

Computational Intelligence Techniques and Phylogenetic Trees for Identification of Sedimentary Petrofacies

Camila M. Saporetti¹, Leonardo Goliatt¹,
Leonardo C. de Oliveira², Egberto Pereira³

¹Faculdade de Engenharia - Universidade Federal de Juiz de Fora (UFJF),
Campus Universitário, São Pedro, 36036-330, Juiz de Fora, MG

²Sísmica de Reservatórios - Petrobras,
Avenida República do Chile, 65, Centro, 20031-912, Rio de Janeiro, RJ

³Faculdade de Geologia - Universidade do Estado do Rio de Janeiro (UERJ),
Rua São Francisco Xavier 524, 20559-900, Rio de Janeiro, RJ

camilasaporetti@ice.ufjf.br, leonardo.goliatt@ufjf.edu.br,
leogeo.oliveira@petrobras.com.br, egberto@uerj.br

Abstract. *The heterogeneities are characterized by several sedimentary petrofacies. Petrofacies identification involves manual processes and time-consuming analyses. The study of the diagenesis has been encouraged by petroleum companies, in order to understand the distribution of porosity in sandstones. This work aims to analyze the use of clustering approaches to identify petrofacies and assist the analysis of petrographic data. In addition, this study introduced the use of the phylogenetic analysis tools to understand the diagenetic process that occurred during sedimentary rock formation. The proposed methodology reaches similar results to those obtained by the conventional method of individualization while allows for reducing time and cost in the individualization task.*

1. Introduction

Exploration and optimization of production of oil fields are subject to the definition and distribution of heterogeneities of hydrocarbon reservoirs, what makes these tasks are paramount. For a rock to be regarded as an appropriate reservoir for exploration and extraction, it must have an acceptable reach, good porosity, a permeable magnitude and a hydrocarbon recovery factor, among other factors. Such properties, called petrophysics, are linked to the depositional pathway of the sediment accumulation units, especially to the sedimentation conditions and the diagenetic process, being the same of extreme importance for the definition of the quality of the reservoir. The heterogeneities are characterized by means of several sedimentary petrofacies, a set of petrographic features that individualize a group of rocks.

Diagenesis is the set of transformations that the sedimentary deposit suffers after deposition, comprehension in the changes in the conditions of pressure, temperature, Eh, pH and water pressure, resulting in dissolution and precipitation through aqueous solutions in the pores. The process ended in the transformation of inconsiderate deposit into rock, or lithification.

The sedimentary petrofacies can be identified by counting under a petrographic microscope of transmitted light, in order to evaluate the heterogeneities present in the reservoir of interest. This procedure is usually long because it involves the process of sampling, data generation and subsequent interpretations of the same. As previously mentioned, the petrophysical properties are related to the diagenetic history of the formation of rocks, in order to initiate a study of the use of phylogenetic analysis to aid in the understanding of diagenesis and in the characterization and individualization of petrofacies. Due to the large amount of data, not all information is used with the manual method. Consequently, it becomes interesting to change the use of manual methods for automatic analysis through computational tools.

In this context, Computational Intelligence techniques, more precise clustering methods, appear as a useful mechanism to assist in the determination of petrofacies. In recent years several papers have used techniques based on computational intelligence to assist in solving problems of reservoir characterization and the employee methods based on phylogeny in various problems. [Ruzgar and Erciyas 2012] developed a technique called Clustered Neighbor-Joining (CNJ) which is based on the use of clustering methods followed by a technique to construct phylogenetic trees. In this work was used in data Y-DNA haplogroups.

[Kadkhodaie-Ilkhchi et al. 2013] applied a hierarchical clustering technique to determine electrofacies in well profile data (gamma ray, sonic transit time, apparent density and neutron porosity). [Martinelli and Eidsvik 2014] employed grouping strategies to construct sequential designs for Bayesian Networks and Random Markov Fields of pre-oil prospects, helping in the decision of which well to drill first. [El Sharawy and Gaafar 2016] used Cluster Analysis and Principal Component Analysis to identify electrofacies in seismic profile data (gamma ray, sonic, density and neutron).

[Oliveira and Pereira 2009] was hypothesized that the diagenetic changes that control the petrophysical characteristics of the reservoirs present a close relationship with the stratigraphic framework and he characterized in stratigraphic and petrographic terms the sandstones interdigitated with the shales of the Ponta Grossa Formation. The database used in this work is the result of the research cited.

This work aims to apply clustering techniques to determine sedimentary petrofacies in a way that can assist the geologist/petrologist in this task and apply a method of Phylogenetic Analysis to analyze the behavior of the samples. Therefore, was opted for the use of unsupervised techniques, since the idea is not to have the samples previously categorized.

2. Materials and Methods

The data analyzed are petrographic thin-sections collected from stratigraphic wells that reached the Tibagi, Devonian member of the Paraná Basin. This basin is located in a tectonically stable region, which has an area of 1400000km^2 with parts situated in Brazil, Argentina, Paraguay and northern Uruguay, as shown in Figure 1.

The manual process by [Oliveira 2009] is presented below: the collected samples were impregnated to obtain thin-sections for the characterization of the diagenetic process of the sandstones. This description originated data to analyze the diagenetic process and to

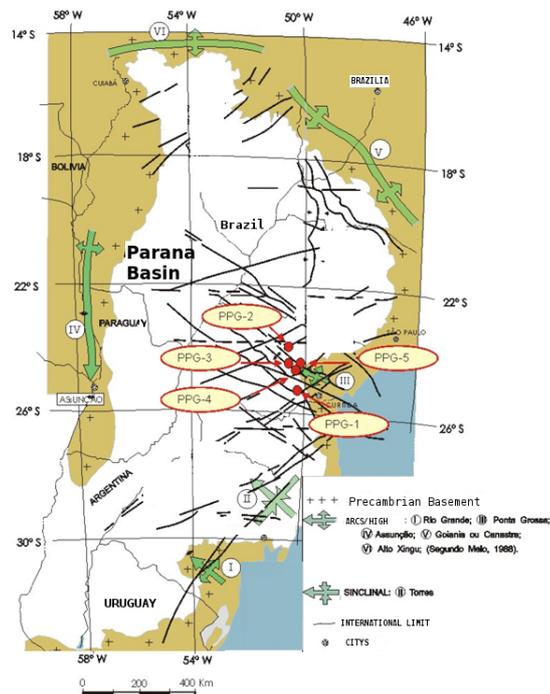


Figure 1. The location of the Paraná Basin and the position of the wells (red dots) (modified from [Oliveira and Pereira 2009]).

establish its influence on the quality of the sandstones as reservoir of hydrocarbons. These thin-sections were analyzed under a Leica DMLP microscope with polarized light and reflected and photographed by Nikon Coolpix 9900 digital camera. The sandstones were characterized petrographically from the [Wentworth 1922] and textural grading of [Folk 1980]. The selection of the grains was based on the work of [Beard and Weyl 1973]. The samples were impregnated with blue epoxy resin to visualize the pores, according to the procedure proposed by [Cesero et al. 1989] and [Lindholm and Finkelman 1972]. The dyeing technique was used to identify carbonates. This methodology is in accordance with the publication of [Evamy 1963], which guides the use of a solution of alizarin (10%) and ferricyanide in dilute HCl (0.15%). All these processes were carried out in LGPA at UERJ. The samples were analyzed quantitative way through the count of 300 points in each thin-section, spaced 0.3 mm, where it was sought to recognize the diagenetic modifications and the chronological relationship between them, based on the textured and faciological relations observed.

In the samples from the Paraná Basin, were counted for each thin-section, the percentages of 22 constituents: quartz, total feldspar, muscovite, opaque minerals, tourmaline, zircon, rutile, glauconite, chlorite, pseudo matrix, lithoclast, bioclast, secondary growth of quartz, kaolinite, illite/smectite, pyrite, siderite, carbonaceous cement, siliceous cement, ferruginous cement, intergranular porosity and intragranular porosity. Each thin-section prepared from samples collected in the drill hole is a sample with 22 features, which correspond to the amount of constituents counted. The database presents information from 5 wells and 44 samples in total.

According to the results obtained by the quantitative analysis of petrographic data,

it was possible to identify the sandstones (by manual classification) in 6 different petrofacies: P-1, P-2, P-3, P-4, I-1 and I-2 [Oliveira 2009]. For Well 1 (PPG1) 12 thin-sections classified from PPG1-1 to PPG1-12 were analyzed. According to the manual analysis the samples PPG1-1, PPG1-9 to PPG1-12 were identified as belonging to the petrofacies named I-2; samples PPG1-2 to PPG1-8 to petrofacies I-1. From the Well 2 (PPG2) 11 thin-sections were analyzed, PPG2-1 to PPG2-11, where the PPG2-1 sample was classified in the P-2 petrofacies and the remainder in the P-1 petrofacies. In Well 3 (PPG3) the analysis was performed on 12 samples, PPG3-1 to PPG3-12. The PPG3-11 thin-section was classified as belonging to P-3 and the remaining 11 samples were pooled to P-1. Well 4 (PPG4) had 5 samples analyzed, PPG4-1 to PPG4-5, where all were classified as referring to P-4. Four thin-sections from Well 5 (PPG5), PPG5-1 to PPG5-4, plate PPG5-3 was identified as relative to P-3 and the remainder to P-1.

The methodology used in this work was into three parts: (1) Intra-Well Analysis, with the objective of analyzing the behavior of the methods in front of each well, individually, (2) the development of an Inter-Well Analysis, in order to verify the division of the thin-sections into groups when information from all wells is stored in the same database and (3) Phylogenetic Analysis technique in order to understanding of diagenesis and the characterization and individualization of petrofacies.

Intra-Well Analysis is intended to define the petrofacies and the way the samples were separated in these petrofacies [Cevolani et al. 2011]. This procedure consists of three techniques: clustering methods, silhouette analysis [Rousseeuw 1987] and Principal Component Analysis (PCA). In the Inter-Well Analysis, each cluster generated in the previous step was represented by the mean of its constituents and presented in groups by the methods applied in the Intra-Well Analysis. This procedure consists of two techniques: clustering methods and Principal Component Analysis (PCA). The Inter-Well Analysis is the complementation of Intra-Well, as it provides a deeper understanding of the behavior of nearby drill holes.

For the Phylogenetic Analysis was used the Neighbor-Joining method [Saitou and Nei 1987], silhouette analysis and for visualization was used ETE 3 [Huerta-Cepas et al. 2016]. The clustering algorithms used in the first two parts were: DBSCAN (Density Based Spatial Clustering of Applications with Noise), K-Means and Ward (hierarchical), and later a comparison was made of the groupings found by these algorithms. These methods were chosen because they used different approaches.

DBSCAN [Ester et al. 1996] is a density-based algorithm where it finds the number of clusters by density distribution corresponding points. The method assumes that clusters are regions of high density separated by regions of low density. The technique begins with the random choice of a starting point p , called the central point. If a minimum number of points, $MinPts$, distance ϵ or less, a ϵ -neighborhood is generated, forming an initial grouping. Since the method is iterative, these points can move, belonging or not to a grouping. The points not contained in some grouping generated will be called noise. The procedure is finalized when no new point can be added in some grouping.

K-Means [Xiong et al. 2009] is one of the most widely used and one of the simplest clustering algorithms. Through the selection of initial centroids, randomly, each sample is assigned to the nearest centroid. Later new centroids are created from the mean

value of all samples designated for each previous centroid. The calculation of the difference between the old and the new centroids, repeating the process of the assignment of samples to the centroids, until this value is less than a threshold. It can be said that the repetition occurs until the centroids do not move significantly. The number of groupings to be generated is passed as a function parameter.

Ward [Seidel et al. 2008] is a hierarchical clustering method, which creates nested clusters, joining them gradually. It is based on Ward's technique, which is a minimization approach. Ward's minimum variance criterion minimizes the total variance within the cluster. At each step, the pair of clusters with the minimum distance is clustered. And in each step finds the pair of clusters that takes the minimum increase in the total variance within the cluster after the join.

An exhaustive search is performed for each parameter in a range according to each method. The parameters chosen are those that maximize the value of the silhouette. The definition of the silhouette analysis is described below.

Silhouette analysis is a technique proposed by [Rousseeuw 1987]. It is a geometric method based on the compaction and separation of clusters with the intention of to analyze the quality of the formed groupings. The optimal number of clusters is defined by the larger silhouette coefficient resulting from the silhouette analysis. For each sample i the value s_i is defined by the following formula:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (1)$$

Considering that sample i belongs to cluster A, a_i is described as the mean dissimilarity of sample i in relation to all other samples of cluster A. Let B be a grouping other than A, b_i is the minimum mean dissimilarity of sample i in relation to all other samples of B. The silhouette coefficient of a set of data is given by the mean of the individual coefficients of the samples

$$\bar{s} = \frac{\sum_{i=1}^N s_i}{N} \quad (2)$$

where N is the number of samples in the data set. The metric used to calculate the dissimilarity is the Euclidean distance. The value of itself varies from -1 to 1. Result close to 1, indicates that the objects are well grouped.

In order to visualize the results of the Intra-Well and Inter-Well Analysis in two dimensions, the PCA was applied. This technique projects the multidimensional data into a low space dimension, in case two. The main components are linear combinations of the original variables, each one conducting different statistical information with importance decreasing, that is, the first component has maximum variability in the data and each subsequent component has the maximum variance. In this way, they can be analyzed, showing the graphics of the first two main components, which statistical information. The PCA contributes to the interpretation of the collected data and to the separation of the important information from the redundant one.

To visualize the results of the Phylogenetic Analysis was used the Environment for Tree Exploration v3 (ETE3). ETE3 is a computational framework that simplifies the reconstruction, analysis, and visualization of phylogenetic trees and multiple sequence

alignments. The new features include (i) building gene-based and supermatrix-based phylogenies using a single command, (ii) testing and visualizing evolutionary models, (iii) calculating distances between trees of different size or including duplications, and (iv) providing seamless integration with the NCBI taxonomy database.

[Saitou and Nei 1987] created a method called Neighbor-Joining to define the closest pairs of elements or neighbors, in order to minimize the total length of the tree generated. A pair of neighbors is identified as being formed by two elements connected by a branch in a tree without bifurcating root, That is, two branches connected by an inner node. According to [Weir and Ott 1997] it is usually possible to define the structure of a tree by frequent pairings of neighbors. This technique starts with a star-shaped topology. Neighbors are the pairs of genera that, when merged, originate in a tree of smaller length. They must be united in order to generate a new combination. [Saitou and Nei 1987] presented that this technique causes a tree considered good, by pure addition of data, whose distance between each pair of genera is the sum of the lengths of the branches that bind them in the tree [Weir and Ott 1997].

3. Results and Discussions

3.1. Intra-Well Analysis

Table 1 shows the number of clusters, the value found from the Validation Criterion (SC-Silhouette). Figure 2 presents the clusters found by the three methods tested of wells PPG1, PPG2, PPG3, PPG4, and PPG5, respectively. For the PPG1 all methods generated two clusters, coinciding with the manual method in the classification of petrofacies I-1 and I-2. They have grouped 100% of the thin-sections correctly. In the PPG2, the DBSCAN, K-Means and Ward methods found two clusters and the results coincided with the manual classification by correctly identifying the associated petrofacies. For the PPG3 the methods obtained two clusters and the results found coincided with the manual classification. In the PPG4, the methods found two clusters. By the conventional method, a single petrofacies, P-4, was identified. The method did not classify the PPG4-1 and PPG4-5 samples as belonging to P-4. The methods were able to distribute 60% of the thin-sections correctly. In the last well, PPG5, the methods found results that coincided with manual classification. By using the silhouette coefficient one can achieve good results. There are other criteria for group validation, but we chose the silhouette since it takes into account the intra-cluster and inter-cluster distances for each sample.

3.2. Inter-Well Analysis

Table 2 contains the ratios of the thin-sections and petrofacies to the groups found by the DBSCAN, K-Means and Ward methods, respectively. Where TS are the thin-sections and PF their respective petrofacies for each group. The techniques related, for each group, samples associated to a petrofacies. Figure 3 shows the visualization of the distribution of the groups found in the Inter-Well Analysis and its proximity. The first figure shows how the petrofacies were distributed by the conventional method. The following represent the form in which the petrofacies were separated into groups by the methods DBSCAN, K-Means and Ward, respectively.

In Figure 3 and Table 2 was observed that in the result obtained by DBSCAN method the group GPPG1-0 is isolated, since only it has samples of petrofacies I-1.

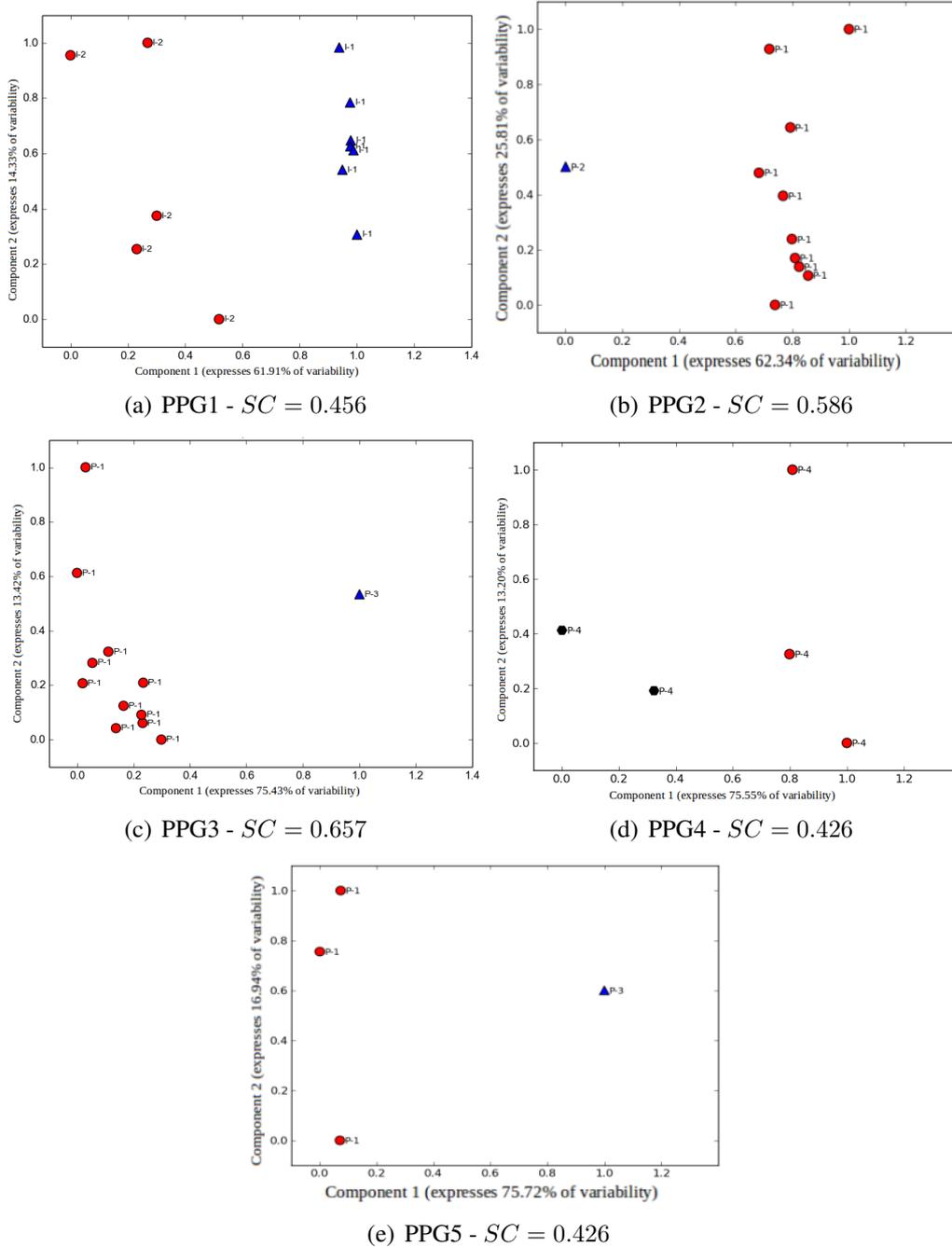


Figure 2. Intra-Well Analysis view of the wells PPG1, PPG2, PPG3, PPG4, and PPG5 respectively.

Table 1. Validation Criterion found by methods in wells.

METHOD	WELL	N° OF CLUSTERS	SC VALUE
DBSCAN	PPG1	2	0.455913
K-Means	PPG1	2	0.455913
Ward	PPG1	2	0.455913
DBSCAN	PPG2	2	0.585717
K-Means	PPG2	2	0.585717
Ward	PPG2	2	0.585717
DBSCAN	PPG3	2	0.656816
K-Means	PPG3	2	0.656816
Ward	PPG3	2	0.656816
DBSCAN	PPG4	2	0.426477
K-Means	PPG4	2	0.426477
Ward	PPG4	2	0.426477
DBSCAN	PPG5	2	0.395879
K-Means	PPG5	2	0.395879
Ward	PPG5	2	0.395879

Table 2. Groups obtained by DBSCAN, K-Means and Ward.

GROUPS	DBSCAN		K-MEANS		WARD		[Oliveira 2009]		
	TS	PF	TS	PF	TS	PF	TS	PF	
GPPG1-0	PPG1	2 to 8	I-1	2 to 8	I-1	2 to 8	I-1	2 to 8	I-1
GPPG1-1	PPG1	1, 9 to 12	I-2	1, 9 to 12	I-2	1, 9 to 12	I-2	1, 9 to 12	I-2
GPPG2-0	PPG2	2 to 11	P-1	1	P-2	1	P-2	1	P-2
GPPG2-1	PPG2	1	P-2	2 to 11	P-1	2 to 11	P-1	2 to 11	P-1
GPPG3-0	PPG3	1 to 10, 12	P-1	1 to 10, 12	P-1	11	P-3	11	P-3
GPPG3-1	PPG3	11	P-3	11	P-3	1 to 10, 12	P-1	1 to 10, 12	P-1
GPPG4-0	PPG4	2 to 4	P-4	2 to 4	P-4	1, 5	P-4	1 to 5	P-4
GPPG4-1	PPG4	1, 5	P-4	1, 5	P-4	2 to 4	P-4	–	–
GPPG5-0	PPG5	3	P-3	1, 2 and 4	P-1	3	P-3	3	P-3
GPPG5-1	PPG5	1, 2 and 4	P-1	3	P-3	1, 2 and 4	P-1	1, 2 and 4	P-1

Although the manual classification has designated the P-2 petrofacies for the GPPG2-0 group and P-1 for the GPPG5-0 group, the procedure employed suggests that these two groups have similar attributes. According to the Intra-Well Analysis the petrofacies P-4 was assigned to groups GPPG4-0 and GPPG4-1. The results suggest that the GPPG1-1 group has features associated with the GPPG2-1 and GPPG3-0 groups, as does the GPPG3-1 and GPPG5-1 groups, indicating that in each set the samples share the same features.

In the K-Means method, the group GPPG1-0 appears isolated, since it is the only one that contains thin-sections of petrofacies I-1. According to the manual classification and Intra-Well Analysis of petrofacies P-2 was attributed to group GPPG2-1 and P-1 to group GPPG5-1, but the methodology employed suggests that these two groups have similar attributes. As occurred in the Intra-Well Analysis the petrofacies P-4 was assigned to the groups GPPG4-0 and GPPG4-1. The results indicate that the GPPG1-1 group shares

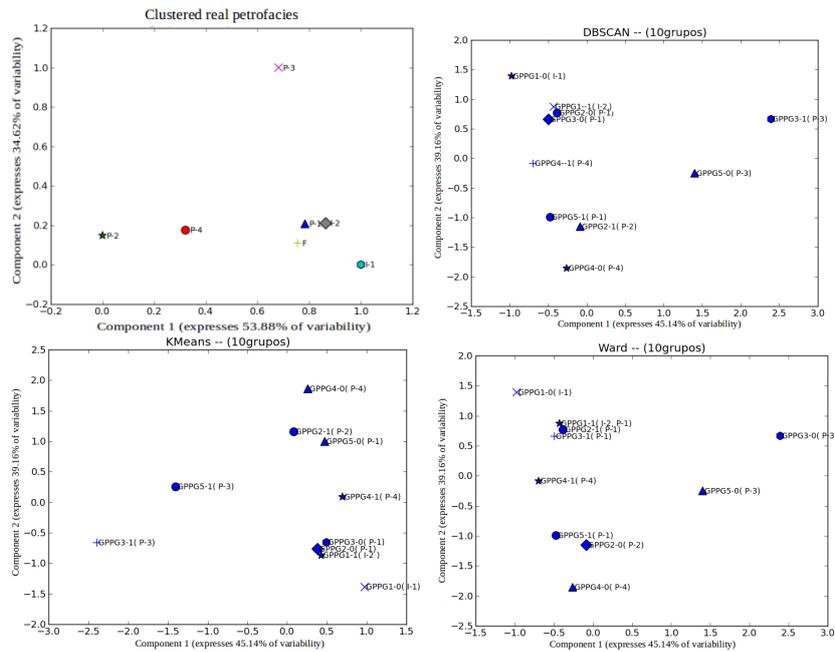


Figure 3. Visualization of the distribution of the groups found in the Inter-Well Analysis and its proximity.

features with the GPPG2-0 and GPPG3-0 groups, as well as the GPPG3-1 and GPPG5-0 groups, indicating that in each set the samples share the similar attributes.

When applying the Ward method, it was observed that the group GPPG1-0 appears separated, since it is the only one that contains thin-sections of petrofacies I-1, not having attributes in common with the other groups. Although the manual classification has designated the P-2 petrofacies for the GPPG2-1 group and P-1 for the GPPG5-1 group, the methodology used indicates that these two groups have similar attributes. In the same way as in the Intra-Well Analysis the P-4 petrofacies were assigned to the GPPG4-0 and GPPG4-1 groups. The results suggest that the GPPG1-1 group shares features with the GPPG2-1 and GPPG3-1 groups, as well as the GPPG3-0 and GPPG5-0 groups, indicating that in each set the samples share the same attributes.

3.3. Phylogenetic Analysis

Figure 4 shows the result of the Phylogenetic Analysis, where it can be observed that the samples belonging to petrofacies I-1 are separated into a subgroup. Samples belonging to P-3 were put together in a subgroup. Two of the five samples of the P-4 petrofacies were between P-1 samples, in the Intra-Well Analysis two samples were separated, which may be an indication that these samples may belong to P-1 and that the diagenetic process undergone by these samples were similar. The samples belonging to I-2 are among the samples of petrofacies P-1, in the Inter-Well Analysis it can be noted that the sample group(s) of the petrofacies I-2 is/are close to the group(s) of P-1, the same assumption made previously may be employed. A nan value appears, this is due to the fact that Neighbor Joining generates star tree, that is, without root. As we wanted to adopt this format to better illustrate the result, the root was nan.

tifying and classifying petrofacies, obtaining high indexes of accuracy compared to the manual method.

Table 2 shows the results of Inter-Well Analysis. This procedure allows visualizing and identifying similarities between thin-sections of different groups and so on as the same petrofacies is distributed along different wells. The results were consistent with those found in Intra-Well Analysis.

The results from the Phylogenetic Analysis (Figure 4) made it possible to visualize the distribution of the samples and the respective petrofacies indicating the samples that similar diagenetic processes. Some results agreed with Intra-Well and Inter-Well Analysis.

The computational method associated with DBSCAN, K-Means and Ward methods can help the geologist/petrologist in the task of identifying the petrofacies, reducing the analysis time in comparison to the manual classification. Since K-Means is a simple method of being optimized, the results are generated faster, the method being more indicated. Phylogenetic Analysis can help in the understanding of the diagenetic process as well as in the classification of sedimentary petrofacies.

References

- Beard, D. and Weyl, P. (1973). Influence of texture on porosity and permeability of unconsolidated sand. *AAPG bulletin*, 57(2):349–369.
- Cesero, P., Mauro, L., and De Ros, L. (1989). Técnicas de preparação de lâminas petrográficas e de moldes de poros na petrobrás. *Boletim de Geociências da PETROBRAS*, 3(1/2):105–116.
- Cevolani, J. T., de Oliveira, L. C., Goliatt, L., and Pereira, E. (2011). 6º congresso brasileiro de pesquisa e desenvolvimento em petróleo e gás.
- El Sharawy, M. S. and Gaafar, G. R. (2016). Reservoir zonation based on statistical analyses: A case study of the nubian sandstone, gulf of suez, egypt. *Journal of African Earth Sciences*, 124:199–210.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.
- Evamy, B. (1963). The application of a chemical staining technique to a study of dedolomitisation. *Sedimentology*, 2(2):164–170.
- Folk, R. L. (1980). *Petrology of sedimentary rocks*. Hemphill Publishing Company.
- Huerta-Cepas, J., Serra, F., and Bork, P. (2016). Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638.
- Kadkhodaie-Ilkhchi, R., Rezaee, R., Moussavi-Harami, R., and Kadkhodaie-Ilkhchi, A. (2013). Analysis of the reservoir electrofacies in the framework of hydraulic flow units in the whicher range field, perth basin, western australia. *Journal of Petroleum Science and Engineering*, 111:106–120.

- Lindholm, R. and Finkelman, R. B. (1972). Calcite staining: semiquantitative determination of ferrous iron. *Journal of Sedimentary Research*, 42(1).
- Martinelli, G. and Eidsvik, J. (2014). Dynamic exploration designs for graphical models using clustering with applications to petroleum exploration. *Knowledge-Based Systems*, 58:113–126.
- Oliveira, L. (2009). *Estudo das relações entre o arcabouço estratigráfico e as alterações diagenéticas observadas na seção Devoniana da Bacia do Paraná*. PhD thesis, Dissertação de Mestrado, Faculdade de Geologia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro.
- Oliveira, L. C. and Pereira, E. (2009). Aplicação de parâmetros diagenéticos para a caracterização do arcabouço estratigráfico do devoniano da bacia do paran. In *5o Congresso Brasileiro de Pesquisa e Desenvolvimento em Petrleo e Gs*.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Ruzgar, E. and Erciyas, K. (2012). Clustering based distributed phylogenetic tree construction. *Expert Systems with Applications*, 39(1):89–98.
- Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.
- Seidel, E. J., Jnior, F. d. J. M., Ansu, A. P., and Noal, M. R. C. N. C. (2008). Comparo entre o mtodo ward e o mtodo k-mdias no agrupamento de produtores de leite. *Cincia e Natura*, 30(1):7.
- Weir, B. S. and Ott, J. (1997). Genetic data analysis ii. *Trends in genetics*, 13(9):379.
- Wentworth, C. K. (1922). A scale of grade and class terms for clastic sediments. *The Journal of Geology*, 30(5):377–392.
- Xiong, H., Wu, J., and Chen, J. (2009). K-means clustering versus validation measures: a data-distribution perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):318–331.