

ELECTROFACIES CHARACTERIZATION USING SELF-ORGANIZING MAPS

Michelle Chaves Kuroda¹, Alexandre Campana Vidal², Emilson Pereira Leite³
and Rodrigo Duarte Drummond¹

Recebido em 16 agosto, 2010 / Aceito em 27 dezembro, 2011
Received on August 16, 2010 / Accepted on December 27, 2011

ABSTRACT. The characterization of electrofacies is essential for reservoir modeling. However, this is a process that depends on many variables, with errors and associated noise that interfere on visual interpretation. In order to minimize uncertainties, this paper proposes the use of the artificial neural network called Auto-Maps Organizing, which is a computational algorithm inspired on the brain function that maps and groups similar information. The petrophysical data used, refers to Namorado Field on Campos Basin, of which were studied the neutron porosity, gamma ray, density and sonic profiles, to classify the field's reservoirs lithology. From the results it was possible to define the reservoir geometry and to detail its features with more accuracy.

Keywords: neural networks, self-organizing maps, recognition of lithofacies.

RESUMO. A caracterização de eletrofácies é essencial para a modelagem de reservatórios. Todavia, esse é um processo dependente de muitas variáveis, com erros e ruídos associados que dificultam a interpretação visual. Para minimizar incertezas, este trabalho propõe a utilização da rede neural artificial denominada Mapas Auto-Organizáveis, um algoritmo computacional inspirado no funcionamento cerebral que mapeia e agrupa informações semelhantes. Os dados petrofísicos utilizados são referentes ao Campo de Namorado, Bacia de Campos, dos quais foram estudados os perfis porosidade neutrônica, raio gama, densidade e sônico, com o objetivo de classificar a litologia dos reservatórios do campo. Através dos resultados foi possível delimitar a geometria dos reservatórios e detalhar suas características de forma mais precisa.

Palavras-chave: redes neurais, mapas auto-organizáveis, reconhecimento de litofácies.

¹Universidade Estadual de Campinas, Prédio da Reitoria V, Barão Geraldo, 13083-872 Campinas, SP, Brazil. Phone: +55(19) 3521-4659; Fax: +55(19) 3289-1562 – E-mails: mckuroda@ige.unicamp.br; rdrummond@gmail.com

²Universidade Estadual de Campinas, Prédio do DGRN, Rua João Pandiá Calógeras, 51, Barão Geraldo, Caixa Postal 6152, 13983-970 Campinas, SP, Brazil. Phone: +55(19) 3521-5198; Fax: +55(19) 3521-4552 – E-mail: vidal@ige.unicamp.br

³Universidade Estadual de Campinas, Prédio do DGRN, Rua João Pandiá Calógeras, 51, Barão Geraldo, Caixa Postal 6152, 13983-970 Campinas, SP, Brazil. Phone: +55(19) 3521-4697; Fax: +55(19) 3521-4552 – E-mail: emilson@ige.unicamp.br

INTRODUCTION

The geological models of hydrocarbon reservoirs are based on the estimates of reservoir properties, such as: lithology, porosity, permeability and fluid type (e.g. Alhoniemi et al., 1999; Bhatt & Helle, 2002; Matos et al., 2004; Stundner & Oberwinkler, 2004). The use of robust mathematical methods aims at reducing uncertainties when generating such models, the more reliable the results the more they help to minimize exploration risks and costs.

Among these, artificial intelligence technologies are the most used methods in geosciences, with emphasis on Artificial Neural Networks (ANN) (Haykin, 1999). These ANNs have the ability to model non-linear relationships among the variables involved, are highly adaptable to input dataset and do not require prior knowledge of data statistical distribution. These fundamental features rank ANN among the most important clustering and classification methods of the last decade (Haykin, 1999; Stunder & Oberwinkler, 2004), and it is therefore, largely used in reservoir geology, which involves geological heterogeneities on various scales and treatment of data sets from several sources (Kohonen, 2001; Bhatt & Helle, 2002; Stundner & Oberwinkler, 2004; Esposito et al., 2008). As example of such application, Stundner & Oberwinkler (2004) mention three basic ANN uses to treat data from well logs: prediction of permeability and/or porosity; generation of synthetic wells and determination of electrofacies.

This work proposes to use this method to analyze the electrofacies of the geophysical dataset of the Namorado Oil Field, Campos Basin, in order to identify in detail the lithologies related to reservoir rocks. The choice to specifically study reservoir lithologies is due to its importance in building the geological model of the reservoir, whose main information sources are the data from well logs (Matos et al., 2004).

The method proposed in this work is an unsupervised ANN, in which *a priori* information about the lithology is not used to classify the electrofacies. The ANN Self Organizing Maps (SOM) (Kohonen, 2001; Coléou et al., 2003) stands out for this type of sample.

Coléou et al. (2003) advocate that SOM results are more accurate compared to other multivariate statistical techniques such as PCA, ICA and K-Means, using as comparison criteria redundancy, noise and continuity. According to Stundner & Oberwinkler (2004), SOM also allows mixing statistical forecasting methods to obtain information from wells, by combining multivariate data visualization with clustering ability.

In most cases, the traditional methods are efficient to separate clearly distinct facies, such as reservoir and non-reservoir. However, the detailing of single facies presents the limitation of

indirect profile data and lower density of core information, which prevents the accurate use of a supervised method. Additionally, petrophysical measurements may be linked to changes of physical properties and not to lithologic variations. However, SOM's great sensitivity and learning capacity enables group identification, resulting in more accurate characterization of reservoir rocks.

NAMORADO OIL FIELD

Discovered in 1975, the gigantic Namorado Oil Field located in the north-central part of the Campos Basin at 80 km from the shore, has an area of 21 km² (Fig. 1), depth between -2900 and -3400 m and estimated average thickness between 90 m and 180 m (Johann, 1997; Souza Jr., 1997). The area is limited by structural and stratigraphic closures, with top and bottom boundaries clearly recognized in well profiles and seismic lines.

The reservoir is deposited in a complex of turbidite channels, along a system of deltaic fronts (Barboza, 2005), called turbidite sandstones of Namorado.

Among the 21 facies present in the core analysis, described by Petrobras and released by Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), the main reservoir rocks refer to two sandstone facies displaying metric thickness in the cores. The dominant reservoir facies in the cores is massif, medium-graded sandstone, arcosean and well selected. The second in dominance is amalgamated coarse sandstones, graded from conglomeratic-coarse sand fraction at the base to coarse sand fraction at the top. Mean porosity and permeability of these two facies are 26% and 400 mD, respectively (Bacoccoli et al., 1980).

The facies considered as potential reservoirs may be classified as secondary reservoirs and correspond to: residual conglomerates, interstratified sandstones and shales; interstratified sandstones and finely stratified shales; medium to fine laminated sandstone. The main difficulty for facies analysis is core representativeness. In this case, interstratified sandstones and shale facies are the most sampled.

METHODS

SOM

The method SOM, developed by Kohonen in the late 70s, is an ANN for unsupervised samples, whose objective is to map the input dataset while searching for similarity in a lower dimensional space, delineated by a geometric structure called Self Organizing Maps. SOM units, also called neurons, are connected to their neighboring structures forming maps, typically hexagonal or rectangular, which determine its vicinity, as shown in Figure 2.

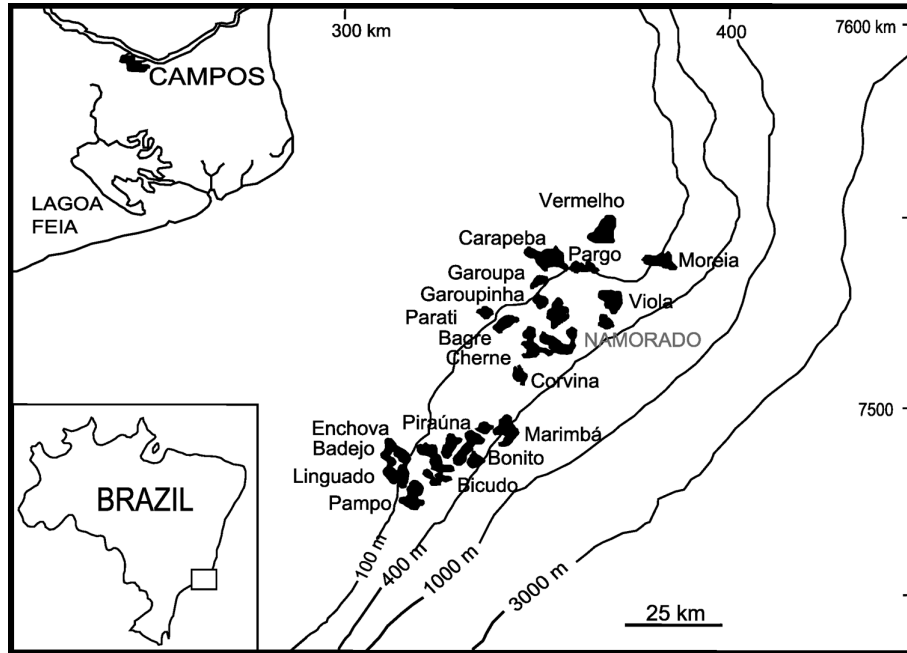


Figure 1 – Location of Campos Basin with major oil fields (Milani et al., 2000).

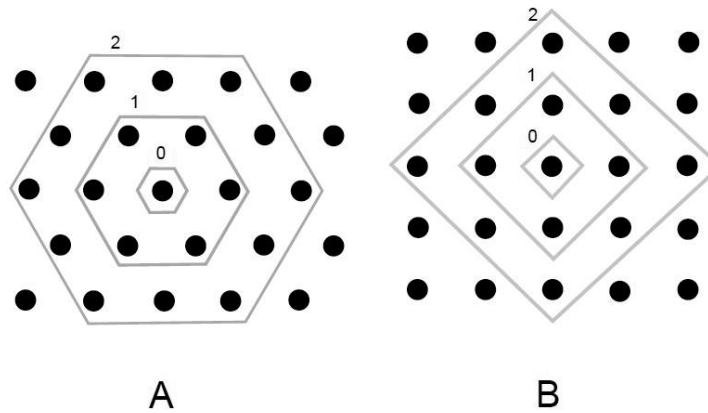


Figure 2 – Two grid configurations and neighborhood levels. (A) represents the hexagonal grid and (B) the rectangular grid (adapted from Costa, 1999).

This map is a flexible grid that molds itself on data similarity. Based on this logic, data close to entry space will have the representation close to the exit space and thus belong either to the same cluster or to nearby clusters. Therefore, SOM preserves the metric and topological relationships of entry space in a two-dimensional network that can be used as a visualization tool displaying different data features, such as their possible organization in clusters (Matos et al., 2004; Leite & Filho, 2010).

The algorithm is based on a competitive learning process, in which only one computational unit output, also called BMU (Best Matching Unit), or local group of computational units provides an active response to input current signal (Gonçalves et al., 1996).

In the input layer, each well sample is a line-vector in the data matrix (where each variable is considered one dimension) represented by the vector: $x = [x_1, x_2, \dots, x_n] \in R^n$. Each unit i of the map has a vector of associated weight:

$$m_i = [m_{i1}, m_{i2}, \dots, m_{in}]^T.$$

The BMUs in this layer are connected to the neighboring layer by the hexagonal grid, which has more neighboring units (6-connected) compared to the rectangular grid (4-connected), as shown in Figure 2.

In this study, the initial random weight values were changed along the procedure to converge to more adequate values inde-

pendent of its initial value (Stundner & Oberwinkler, 2004). The Euclidean distance, the activation function used in the tests, is given by the following equation:

$$d_i = \sum_{j=1}^N (x_j(t) - w_{i,j}(t))^2, \quad (1)$$

where $x_j(t)$ is the j -th entry in a given iteration and $w_{i,j}(t)$ is the weight of BMU j of the input layer connected to neuron i of the output layer.

As each point vector of the input layer has only one associated neuron, the algorithm represents this rule through the following equation, in which the chosen unit (BMU) is represented by the index c :

$$\|x - m_c\| = \min_i \{\|x - m_i\|\}, \quad (2)$$

where $\|\cdot\|$ is the metric Euclidean distance.

After BMU is determined by the shortest distance from the analyzed point, the distances between the unit and its neighbors (determined by the previously chosen grid) are updated, that is, they move closer to input vector in the output space.

Updating i moves the BMU towards vector x , according to the probability density function given by:

$$\begin{cases} m_i = m_i(t) + \alpha(t)h_{bi}(t)[x - m_i(t)], \\ \quad \text{se } m_i \in \text{neighborhood of } c \\ m_i = m_i, \text{ otherwise,} \end{cases} \quad (3)$$

where t is time, $\alpha(t)$ is the learning rate, which varies between 0 and 1 (and decreases along the iterations until its value is no longer significant, that is, they converge) and $h_{bi}(t)$ is the neighboring size, described as:

$$h_{bi}(t) = e^{-\frac{\|r_b - r_i\|^2}{2\sigma^2(t)}}, \quad (4)$$

where r_b and r_i are the positions of neurons b and i in the SOM grid.

Attraction intensity of the input vector is controlled by the average learning, and the number of vectors attracted by the winner BMU is controlled by the neighborhood radius.

Since the main objective of this work is to identify data clusters, techniques that emphasize correlations among points in the map are going to be used to visualize the SOM.

In order to view map similarities, it is interesting to note the distances between the BMUs, especially in the map of the unified distance matrix, or U Matrix, which is a way of representing data relationships (Ultsch, 1992; 1993).

The U Matrix allows visualizing the distances between the neighboring BMUs and their means. Therefore, the observed image shows possible data clusters formed by the samples, suggesting also the number of existing groups by visual analysis. However, this visual analysis is not always effective or sufficient. Matos et al. (2004) reported on the difficulties of choosing the number of clusters, which is usually performed by empirical systems. Therefore, the semi-automatic grouping of U Matrix data is performed by the traditional method K-Means (Khedairia & Khadir, 2008).

The number of units on the map (number of BMUs) is chosen *a priori* and determines the effectiveness and generalization of the model, since during training the method forms an elastic network that adapts itself to input data.

Vesanto et al. (1999) suggest the following expression to choose the size of the output unit:

$$m = 5 * \sqrt{n}, \quad (5)$$

where m is matrix size and n is the number of trained samples.

However, the number of groups can be pre-determined accordingly, for example, a smaller number will identify predominant patterns, while a larger number will detect all image patterns, including those with low occurrence probability (Matos et al., 2004; Gonçalves et al., 2007).

Subsequently, BMUs can be grouped using the K-Means method, which in addition to providing a better understanding about group formation has the advantage of drastically reducing computational time (Vesanto et al., 1999).

K-Means

K-Means, one of the most popular algorithms to identify clusters (Khedairia & Khadir, 2008), groups together in classes a set of input points without *a priori* information.

In this work, after SOM training, K-Means was used to identify clusters in the U Matrix, an efficient way to characterize the groups of the specified matrix (Rocha & Sousa, 2003).

BMUs are classified as belonging to previously chosen k groups that will split the data according to training that minimizes the following error function:

$$E = \sum_{k=1}^c \sum_{x \in q_k} \|x - c_k\|^2, \quad (6)$$

where k is the number of clusters and $k_n c_k$ is the center of cluster n . It should be noted that SOM and K-Means have the same algorithm when the radius of the SOM neighborhood function is

zero. In this case the adjustment is performed only on the winning unit (BMU).

The advantage of SOM is to enable the visualization of any input dataset onto an output set in two dimensions while preserving the topology. However, there is no need for such sophisticated method for the second group separation, since the data have already been sorted out by the SOM algorithm.

After choosing a less robust algorithm, it is essential to determine the number of clusters of K-Means algorithm. To this end, the Davies-Bouldin index was chosen to automate and assist the decision process, according to Eq. (7):

$$\frac{1}{C} \sum_{k=1}^c l \neq k \max \left\{ \frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_{kl})} \right\}, \quad (7)$$

where C is the number of clusters, S_c is the distance between Q_k and Q_{kl} , and d_{ce} the distance between the centroids of respective clusters (calculated from the Euclidean distances between map units). Therefore, Davies-Bouldin index values lower than 1 indicate separated clusters, while values greater than 1 represent clusters possibly crossed.

It is worth mentioning that this index is a heuristic method that seeks to find a value close to the optimal solutions and therefore, it is not necessarily the best.

ANOVA

Since the dataset used in the study has only 2.5% of core data, a statistical study of the well-log data at each point was necessary to extrapolate the characteristics of the cored interval to the remaining data (87.5%).

Multivariate analysis of variance (MANOVA) was the method chosen to compare mean vectors, whose data are usually derived from statistical parameters. The use of statistical analysis to compare mean vectors depends on the partition of total variance subdivided into variance due to both treatment effect and error. The statistical method used to evaluate the partition of total variance is called ANOVA (Johnson & Wichern, 1999).

In studies of continuous random variables, as in the present case, the assumed multinormality allows a multivariate analysis. The relevance of multivariate analysis is the conjunction use of all involved variables (Regazzi, 2002).

ANOVA works with two hypotheses: H_0 , or null hypothesis, there is no variance of the means of the studied variables; H_1 , or alternative hypothesis, in which there is no evidence of non-variance of the means.

The hypotheses are tested using Levene F, which consists of finding an evidence criterion for rejection of H_0 , given by p -

values as shown in Table 1.

Table 1 – Criteria used to reject the hypothesis H_0 .

$P \geq 0.10$	No evidence against H_0
$P < 0.10$	Weak evidence against H_0
$P < 0.05$	Significant evidence against...
$P < 0.01$	Strong evidence against...
$P < 0.001$	Significantly strong evidence against...

Source: Shimakura (2008).

From Table 1 is possible to infer that for p -values greater than 0.1 there is no statistical evidence of changes in the mean values of the studied variable, whereas as p -value diminishes, there is evidence that the variation of the means increases. It is noteworthy to highlight the careful definition of evidence criteria, which points to the lack of absolute certainty, emphasized mainly by the study of the inconsistencies and not by the direct confirmation of event occurrences.

From this criterion, it is possible to infer whether the values for the cored intervals at each group, obtained using SOM, are representative of the remaining sample along the non-cored intervals.

For the cases where there is variance of the means, that is, when p -values are small, a second statistical analysis is necessary to determine confidence intervals to assess whether this variation is statistically significant.

DATA ANALYSIS

The Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) provided the datasets of 17 wells of the Namorado Oil Field, for which the following profiles were analyzed: sonic (DT), neutron porosity (NPHI), density (RHOB) and gamma rays (GR). Six of the studied wells contained core data, with 21 lithofacies.

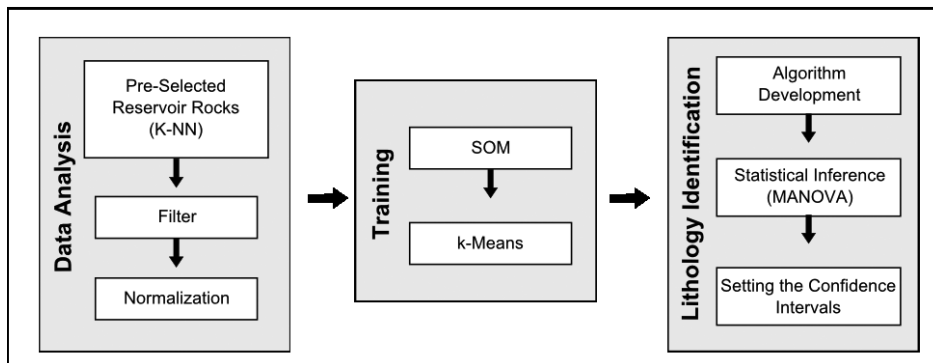
Electrofacies data were classified previously using the supervised k-NN (k-Nearest Neighbor) method (Drummond & Vidal, 2011), which shows 80% accuracy in cross-validation. Three electrofacies groups were identified: reservoir, potential reservoir and non-reservoir. Only reservoir intervals were used in this study. This set consists of 11,600 samples of which only 2.5% correspond to core samples.

The database presents six lithotypes, two of which refer to the reservoir rock and four refer to the potential reservoir rock (Table 2).

The analysis of these datasets aims at both finding subdivisions in the reservoir rocks according to geophysical well logs and evaluating the distribution of different types defined in the Namorado Oil Field.

Table 2 – Facies description of Namorado Oil Field.

Facies	Classification	Description	Number of samples
4	Possible reservoir	Residual conglomerates	6
6	Reservoir	Amalgamated coarse sandstone	90
7	Possible reservoir	Fine laminated medium sandstone	4
8	Reservoir	Massif or medium-graded sandstone	115
10	Possible reservoir	Interstratified sandstone/shale	35
11	Potential reservoir	Finely interstratified sandstone/shale	55

**Figure 3** – Schematics of the applied methodology.

In order to use the SOM algorithm to characterize reservoir facies, the work was divided into three stages: Data Analysis, Training and Lithology Identification (Fig. 3).

The performance of neural networks is based on the principle described by Alhoniemi et al. (1999), who emphasized that the results depend mainly on the quality of the used dataset.

The data were processed in five steps, following the method proposed by these authors:

- Data acquisition: the original data are analyzed; their measured variables and the strings representing the nomenclature of lithotypes were transformed into numeric entries;
- Data preprocessing: in this step, the incorrect data are removed or corrected; generally this filtering occurs under special or fixed conditions. In this study all entries, even the ones with data gaps were considered for training, since SOM give good results even when there is information missing;
- Segmentation: the input data are divided into subsets according to criteria established using *a priori* knowledge. The analyzed database considered unsupervised domain and thus there were no subdivisions;

d) Feature extraction: transforms input data vectors in such a way that they describe the problem according to the analysis of the “point of view”. Mathematical modification such as vector decomposition by the Fourier transform, adding or disregarding variables, hinders the training tool. At this stage, the ILD (resistivity) profile was considered redundant for the dataset to be trained by preliminary tests with the chosen method. However, its absence did not interfere with the quality of the results, in addition the ILD profile was disregarded because it did not characterize petrophysical information (as the other variables) but rather, fluid information.

e) Normalization: gives input vectors the correct weight in the SOM training, ensuring that each component will influence equally the training. In this work, data were previously submitted to the equation:

$$\text{Normalized value} = \frac{x_i - \min x}{\max x - \min x} \quad (8)$$

where x_i is the analyzed vector, $\min x$ is the lowest profile value and $\max x$ the highest.

The errors calculated by the SOM method, which define the quality of data mapping, are established by quantization errors

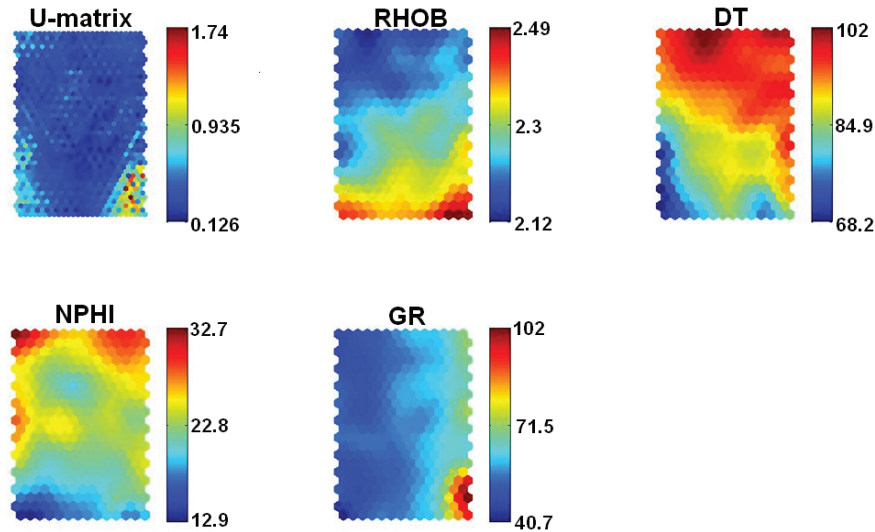


Figure 4 – Component plans of each variable analyzed.

(qe), related to the SOM resolution, and the topographic error (te), related to topology preservation of input data (Kohonen, 2001):

$$qe = \frac{1}{M} \sum_{k=1}^M \|x_k - m_{bk}\| \quad \text{and} \quad te = \frac{1}{M} \sum_{k=1}^M f(x_k) \quad (9)$$

where M is the number of vectors of x data. The function $f(x_k)$ is equal to 1 if the first and the second BMUs of x_k are adjacent and zero otherwise. Subsequent to clustering of the database, algorithms have been developed to assess ANOVA results, as described previously.

TRAINING

The tool Self-Organizing Map in Matlab developed by Vesanto & Alhoniemi (2000), which allows the use of SOM and subsequent regrouping using the K-Means method (Vesanto et al., 1999), was used for several training profiles.

This package offers the possibility of making *a priori* decisions and uses heuristic formulas to choose values, as the Davies-Bouldin index, considered a good indicator of unsupervised analyses (e.g. Matos et al., 2004).

RESULTS

The size of the U Matrix for training the dataset was calculated by Eq. (5) and it was dimensioned in 308 units (14×22).

Figure 4 shows the resulting U Matrix and the contribution of each input variable, represented in the component planes. Smaller values indicate greater similarities between input vectors and nodes, represented by cool colors, while warm colors show

fewer similarities. Thus, groups are characterized by bluish homogeneous regions separated by warm colored regions.

Figure 5A shows the Davies-Bouldin diagram. The function lowest value is at $x = 4$, and so this was defined as the ideal number of groups for the U Matrix, since it minimizes the function described by Eq. (7).

Figure 5B shows the U Matrix regrouped by K-Means. It should be noted that the regions between the clusters have smaller BMU units, which indicates less concentration of samples and, therefore, a good separation among the four groups. The quantization errors for training and topographic were 0.552 and 0.098, respectively.

The profile analysis of each group defined after training shows that the database displays larger variation in the sonic and gamma ray profiles. Density profiles varied very little, since the data set interval is smaller compared to the other variables. When evaluating only the means, it is possible to verify that the groups display different values for most of the profiles; however, this analysis is not enough to determine the lithological classification and therefore, the quality of the reservoirs they represent.

Standard deviation values of the profiles in the four groups are very small, which shows the relative group concentration and very few scattered points. Thus, it was concluded that the groups are well separated.

Table 3 shows that the variables with greater variability were separated according to well-defined intervals, demonstrating that the method is sensitive to small variations, as in the case of the density profile data, where even the smaller values were separated satisfactorily.

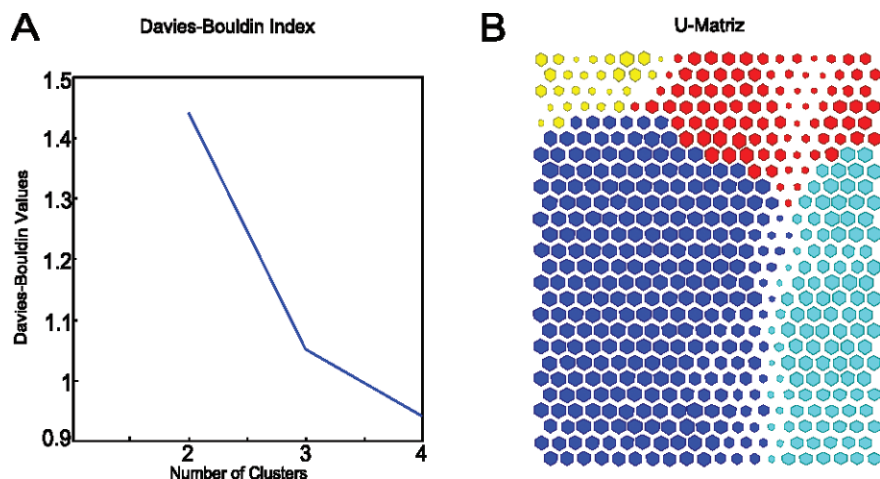


Figure 5 – Image generated applying the tool of Vesanto et al. (1999). (A) Choice of cluster numbers using the K-Means clustering method, (B) U-Matrix regrouped by the method.

Table 3 – Statistical analysis results for: density (RHOB), transit time (DT), neutron porosity (NPHI) and gamma rays (GR) of the 4 groups.

	Group 1				Group 2			
	RHOB	DT	NPHI	GR	RHOB	DT	NPHI	GR
Medians	2.26	88.76	22.97	56.61	2.34	80.18	18.67	53.45
Standard deviation	0.05	6.65	2.33	6.35	0.059	7.08	3.27	7.93
Variance	0	44.34	5.43	40.35	0	50.21	10.75	62.99
Maximum	2.52	106.45	35.02	84.06	2.62	94.64	27.47	75.92
Minimum	2.05	53.47	17.82	33.40	2.17	55.88	0	18.24
Means	2.27	88.24	23.28	56.77	2.35	78.78	18.26	52.44
	Group 3				Group 4			
	RHOB	DT	NPHI	GR	RHOB	DT	NPHI	GR
Medians	2.42	87.33	21.06	90.94	2.19	96.03	25.70	55.38
Standard deviation	0.08	7.69	3.68	14	0.05	4.33	3.48	7.22
Variance	0	59.15	13.55	196.22	0	18.82	12.17	52.24
Maximum	2.61	104.62	29.15	122.94	2.39	110.15	37.23	87.69
Minimum	2.11	64.31	11.69	65.26	2.00	78.28	11.47	37.16
Means	2.42	86.51	20.52	91.67	2.19	96.12	25.88	56.26

Groups were compared to core information associated to them and results are shown in Table 4.

Table 4 shows low core representativeness for each group. Compared to the core, the data showed three groups with high percentage of reservoir rocks, with values 78%, 90% and 83%, and, a group with high lithological concentration of potential reservoirs, 85%.

Three groups of reservoir rocks were determined, with emphasis to the second group that clusters more of these facies. It is worth noting that core description classified two facies for these intervals, while the SOM technique showed a third subdivision as well.

Group three consists of potential reservoir rocks, with very distinct values compared to the other groups.

Table 4 – Supervised data classification, as percentage.

Groups	Core data (%)	Reservoir (%)	Potential reservoirs (%)
1	15.75	78.57	21.43
2	19.48	90.32	9.68
3	30.94	14.29	85.71
4	17.37	83.56	16.44

Table 5 – Lithological classification of supervised data.

Group	Amalgamated coarse sandstone	Massif or medium-graded sandstone	Residual conglomerates	Fine laminated medium sandstone	Sandstone/interstratified shale	Sandstone/finely interstratified shale
1	0.2262	0.5704	0.0271	0.0090	0.0768	0.0905
2	0.2985	0.6119	0.0224	0.0075	0.0224	0.0373
3	0.05615	0.10655	0	0	0.0233	0.8140
4	0.0808	0.9565	0	0.0658	0.1041	0.0658

Table 5 shows the comparison between core data and each group defined by SOM. It is seen that group three, potential reservoir rocks, consists of 81% finely interstratified sandstone and shale, while group four is approximately 93% massif or medium-graded sandstone.

Groups one and two exhibit greater lithological variation of reservoir rocks, group one consists of about 23% amalgamated, coarse-grained sandstone and 52% massif or medium-graded sandstone, while group two consists of 30% and 54% of the same lithologies, respectively.

Due to data complexity and availability of few core data, it is important to analyze statistically the p-value to determine the behavior of the studied profiles. From the Levene F test, according to the criteria used in Table 2, it can be concluded that p-values for the density and sonic profiles are very low, leading to the conclusion that there is significant evidence of the mean variation for the data sets 1, 2 and 4 (Table 6). Therefore, sonic and density profiles of these groups are very likely to have points with very different values that compromise the unity of each group means and thus require further statistical analysis.

Table 6 – P-values.

Variable	Group 1	Group 2	Group 3	Group 4
RHOB	0.0004	0.0000	0.0171	0.0000
DT	0.0095	0.0103	0.2068	0.0147
NPHI	0.3094	0.2583	0.4041	0.6124
GR	0.4650	0.4196	0.3595	0.7736

On the other hand, porosity and gamma rays profile of the same data sets show no variation evidence of sample means and therefore, the null hypothesis cannot be rejected. It is concluded that these two variables do not present statistically significant variations of the sample means and thus it is possible to infer that the data set for these two variables may be represented by the core features, that is, the results show strong evidence

that the data set display behavior similar to the core for these two profiles.

The analysis of *p*-values of group three shows that the density profile has small variation of the sample mean and for the remaining variables there is no evidence against the hypothesis that mean variation is zero. Thus it can be inferred that the core sample is highly correlated to the remaining data, and it is, therefore, representative of the entire set.

However, the variability found for the first two profiles indicates the need to consider the confidence intervals, calculated according to inference statistical analysis to obtain values that identify and confirm its geological features.

From Table 7, it can be inferred that any randomly chosen point from the data set has 95% probability of belonging to the intervals described above, which is considered highly reliable.

Gamma ray profiles of groups one, two and four lie between 51 and 57, characterizing good reservoirs, constituted of low shaliness sandstone as described in the core samples, in addition to low density values compared to group three.

On the other hand, group three has high gamma ray values, above 89, representative of potential reservoirs and high density, compatible with shales and fine grained sandstones, the lithology described for this group.

In order to compare the interval values of the profiles within the groups, boxplots associated to core lithologies in the data set were constructed, as shown in Figure 6.

Figure 6 shows how little the profiles vary and how similar the values are for the studied lithologies. Among the analyzed variables, density profile is the most similar among the six groups, very little change is observed for the massif or medium-graded sandstone, interstratified sandstone/shale, residual conglomerate and finely laminated medium sandstone.

Among the analyzed lithologies, the only variable that stands out is the finely interstratified shale/sandstone, with a large range of values; however, the medium-graded or massif sandstone of

Table 7 – Confidence intervals.

	Group 1		Group 2		Group 3		Group 4	
	Valor min.	Valor max.	Valor min.	Valor max.	Valor min.	Valor max.	Valor min.	Valor max.
RHOB	2.26	2.27	2.34	2.35	2.41	2.44	2.18	2.19
DT	87.90	88.59	78.25	79.31	85.22	87.79	95.89	96.35
NPHI	23.16	23.40	18.02	18.51	19.91	21.14	25.69	26.0
GR	56.44	57.11	51.84	53.03	89.33	94.01	55.87	56.64

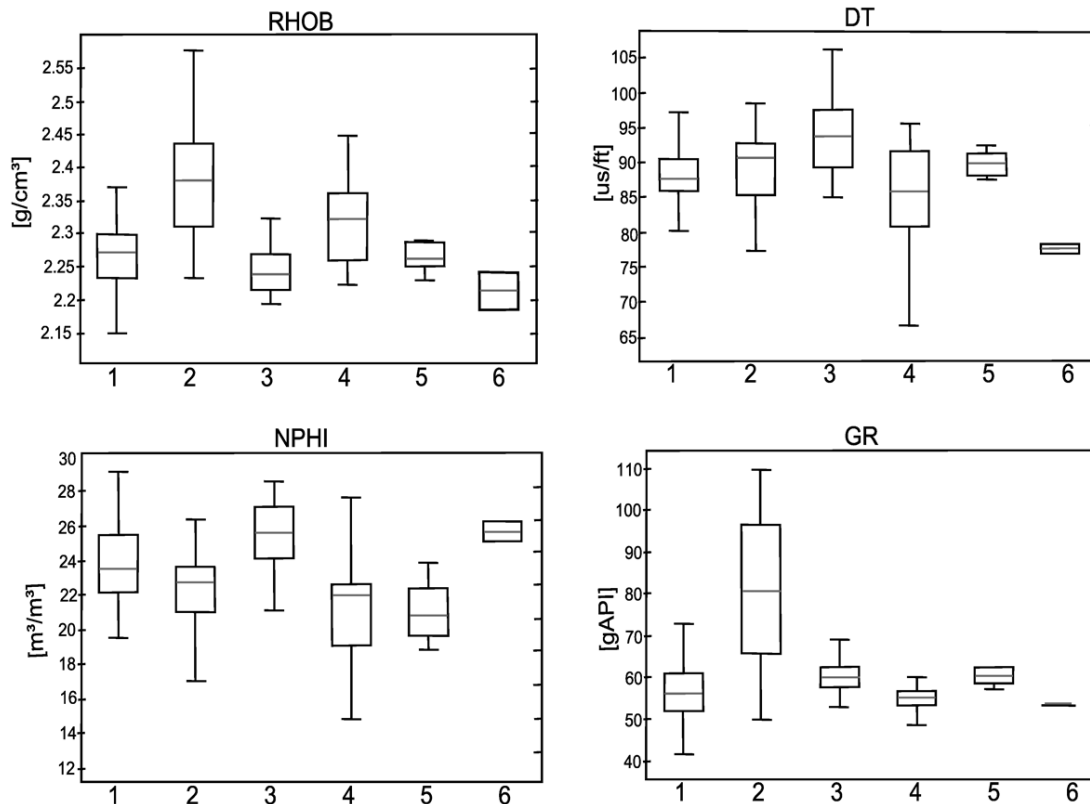


Figure 6 – Boxplots of analyzed lithologies: 1 – Massif or medium-graded sandstone; 2 – Finely interstratified sandstone/shale; 3 – interstratified sandstone/shale; 4 – Amalgamated coarse-graded sandstone; 5 – Residual conglomerates; 6 – Fine to medium laminated sandstone.

the reservoir group displayed density interval very similar to the lithology of the possible reservoir, interstratified shale/sandstone, residual conglomerate and finely laminated medium sandstone.

Within the reservoir group, the coarse-graded amalgamated sandstone lithotype closely resembles the finely interstratified sandstone/shale lithotype. For these two classes, the values of sonic and gamma ray profiles with contained intervals are very similar, however differentiated by the porosity profile, where the possible reservoir facies shows higher values.

The fine laminated medium-graded sandstone values, stand out in the sonic profile, whose interval is below the other classes,

but this variable has only four samples and therefore, a small variation of the values is observed for all profiles. Except for this lithotype, all others display very close average density thus making classification difficult.

The interstratified sandstone/shale class has a porosity profile interval that groups the finely laminated medium-graded sandstone values, and represents the two lithologies with higher porosity values.

In gamma ray profile, the interval of massif or medium-graded sandstone comprises the values of possible reservoir groups; however, the finely interstratified sandstone/shale has

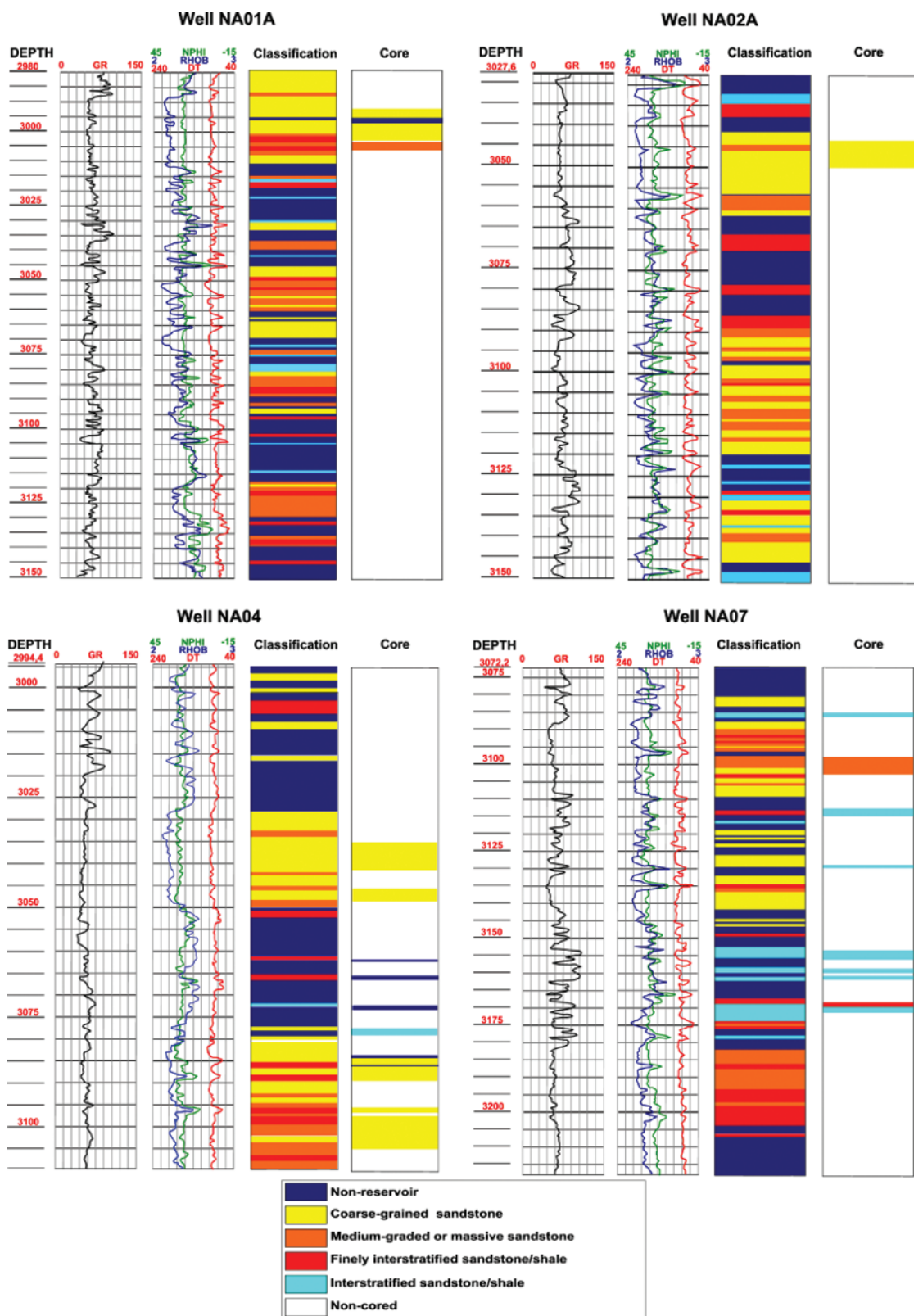


Figure 7 – Lithological classification results for the cored intervals. The white color represents lithologies without core samples. The blue color represents non-reservoir lithologies.

much higher gamma-ray values than the others. This difference found for gamma ray values in this class facilitates pattern recognition for this lithotype.

Since the lithologies differ little, the separation of the lithological groups becomes difficult, especially between reservoir and potential reservoir groups, which have no distinct signature between the profiles.

Massif or medium-graded sandstone shows no significant differences that might distinguish it from the potential reservoir lithotype. For all analyzed profiles, its values are very close to the residual conglomerate values, except for porosity of the potential reservoir rock whose average is relatively lower.

Gamma-Ray profiles with cutoff value of 60° API (indicated by the green color) were used to compare the results with data from well logs. The other three profiles (DT, RHOB and NPHI) are also represented. It should be highlighted the intersection of profiles NPHI and RHOB that indicates reservoir rocks (in yellow).

The four wells displayed in Figure 7 show that the best intervals for reservoir rocks according to the profiles obtained using the SOM method were coarse and medium-graded sandstone facies. The sandstones interbedded with shales were defined correctly in most cases. The comparison with core data shows a very good correlation, it is observed that the results fit core values and classify satisfactorily the wells according to different lithologies.

CONCLUSION

A detailed analysis of the electrofacies of the Namorado Oil Field using the SOM method is shown in this work. The used databases refer to reservoir facies classified after analysis of RHOB, DT, GR and NPHI well logs, by the kNN supervised clustering method. The purpose of the reclassification was to delineate the lithotypes present in these facies while qualifying them either as reservoirs or potential reservoirs.

The U Matrix presented an array of neurons that could be visually grouped; however, the use of K-Means algorithm enabled the separation of other groups and set such limits, which corroborate the subsequent interpretation of data profiles. The two groups determined by the lowest Davies-Bouldin index resulted in the choice of four different groups, two of which could be clustered together since they have very similar characteristics.

The basic statistical analysis of the data showed that group profiles were constant and similar in the four scenarios, thus implying that group behavior remained almost unchanged within different environments. Therefore, a more detailed analysis of the groups was essential.

The diagnosis based on core data led to the conclusion that three of the four groups displayed similar behavior in association with the composition of reservoir rocks. While only one group stands out from the rest, consisting of potential reservoir rocks. The effectiveness of this zoning was also confirmed by comparing with additional information about the Namorado Oil Field and its extrapolation through statistical inference.

ACKNOWLEDGMENTS

The authors thank Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP and PETROBRAS for releasing the data of Namorado Oil Field.

REFERENCES

- ALHONIEMI E, HOLLMÉN J, SIMULA O & VESANTO J. 1999. Process Monitoring and Modeling using the Self-Organizing Map. Integrated Computer Aided Engineering. Available on: <<http://citeseerx.ist.psu.edu>>. Access on: Jan. 17, 2010.
- BACOCOLI G, MORALES RG & CAMPOS OJ. 1980. The Namorado Oil Field: A Major Discovery in the Campos Basin, Brazil. AAPG Bulletin, 63(3): 329–338.
- BARBOZA EG. 2005. Análise Estratigráfica do Campo de Namorado (Bacia de Campos) com base na Interpretação Sísmica Tridimensional. Ph.D. Thesis. Universidade Federal do Rio Grande do Sul, 235 pp.
- BHATT A & HELLE HB. 2002. Determination of facies from well logs using modular neural networks. Petroleum Geoscience, 8: 217–228.
- COLÉOU T, POUPON M & AZBEL K. 2003. Unsupervised seismic facies classification: A review and comparison of techniques and implementation. The Leading Edge, 22: 942–953.
- DRUMMOND RD & VIDAL AC. 2011. Comparação entre as técnicas multivariadas MAF e PCA aplicadas na classificação de eletrofácies. Revista Brasileira de Geofísica, 29(3): 497–509.
- ESPOSITO AM, GIUDICEPIETRO F, D'AURIA L, SCARPETTA S, MARTINI MG, COLTELLI M & MARINARO M. 2008. Unsupervised Neural Analysis of Very-Long-Period Events at Stromboli Volcano Using the Self-Organizing Maps. Bulletin of the Seismological Society of America, 98: 2449–2459.
- GONÇALVES ML, NETTO ML de A & ZULLO JJ. 1996. Um Sistema Neural Modular para Classificação de Imagens Utilizando Mapas de Kohonen. Anais VIII Simp. Bras. de Sensoriamento Remoto 6° Encontro Nacional de Inteligência Artificial, 1: 845–849, Salvador, BA.
- GONÇALVES ML, NETTO MLA & COSTA JAF. 2007. Explorando as Propriedades do mapa Auto-Organizável de Kohonen na Classificação de Imagens de Satélite. Anais XXVII Congr. SBC, 1: 1132–1141.

- HAYKIN S. 1999. *Neural Networks(2)*. Prentice Hall, Virginia, USA, 842 pp.
- JOHANN PRS. 1997. *Inversion Sismostratigraphique et Simulations Stochastiques en 3D: Réservoir Turbiditique Offshore du Brésil, Intégration Géologique, Géophysique et Géostatistique*, Ph.D. Thesis, Académie de Paris, Université Marie Curie. 352 pp.
- JOHNSON RA & WICHERN DW. 1999. *Applied multivariate statistical analysis(4)*. Upper Saddle River, New Jersey: Prentice-Hall, 815 pp.
- KHEDAIRIA S & KHADIR MT. 2008. Self-Organizing Map and K-Means for Meteorological Day Type Identification for the Region of Annaba-Algeria. In: 7th Computer Information Systems and Industrial Management Applications, p. 91–96.
- KOHONEN T. 2001. *Self-Organizing Maps*. Springer-Verlag, New York, 501 pp.
- LEITE EP & FILHO CRS. 2010. Mapas Auto-Organizáveis aplicados ao mapeamento do potencial mineral na região de Serra Leste, Província Mineral de Carajás, Pará. *Revista Brasileira de Geofísica*, 28(3): 397–409.
- MATOS MC, OSÓRIO PLM & JOHANN PRS. 2004. Agrupamento de Mapas Auto Organizáveis de Kohonen Aplicado a Análise de Fácies Sísmicas. In: XV Brazilian Automation Congress, 1: 1–6.
- MILANI EJ, BRANDÃO JASL, ZALÁN PV & GAMBOA LAP. 2000. Petróleo na Margem Continental Brasileira: Geologia, Exploração, Resultados e Perspectivas. *Revista Brasileira de Geofísica*, 18(3): 352–396.
- REGAZZI AJ. 2002. Análise multivariada. Notas de aula INF 766. Departamento de Informática da Universidade Federal de Viçosa.
- ROCHA ACB & SOUZA FAM. 2003. Aplicação do K-Means Cluster a dados de perfis de poços petrolíferos. In: 2^o Congresso Brasileiro de P&D em Petróleo e Gás, Rio de Janeiro. 6 p.
- SHIMAKURA SE. 2008. *Bioestatística A*. Departamento de Estatística, UFPR. Available on: <<http://leg.ufpr.br/~shimakur/CE055/>>. Access on: Feb. 2010
- STUNDNER M & OBERWINKLER C. 2004. Self-Organizing Maps for Lithofacies Identification and Permeability Prediction. In: SPE Annual Technical Conference and Exhibition held in Houston, p. 8.
- SOUZA Jr OG. 1997. *Stratigraphie Séquentielle et Modélisation Probabiliste des Réservoirs d'un Cône Sous-marin Profond (Champ de Namorado, Brésil)*. Intégration des Données Géologiques. Ph.D. Thesis, Université Paris. 128 pp.
- ULTSCH A. 1992. Knowledge acquisition with self-organizing neural networks. In: ALEKSANDER I & TAYLOR J. (Eds.). *Artificial Neural Networks*, Elsevier Science Publishers B.V., 735–740.
- ULTSCH A. 1993. U*-Matrix: a tool to visualize clusters in high dimensional data. University of Marburg, Department of Computer Science, Technical Report, 36, 12 pp.
- VESANTO J & ALHONIEMI E. 2000. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11: 586–600.
- VESANTO J, HIMBERG J, ALHONIEMI E & PARHANKANGAS J. 1999. Self-Organizing Map in Matlab: The SOM Toolbox. *Proceedings of the Matlab DSP Conference*, Espoo, Finland, 59 pp.

NOTES ABOUT THE AUTHORS

Michelle Chaves Kuroda Master's degree in Geology from the Instituto de Geociências, Universidade Estadual de Campinas, in 2012. Bachelor's degree in Mathematics (2007) from the same university. Currently, works as researcher at Centro de Estudos do Petróleo, Universidade Estadual de Campinas. Interested in data computer analysis, with emphasis on geophysical data.

Alexandre Campana Vidal. Bachelor in Geology from the Universidade de São Paulo in 1993, Master's degree in Geoengineering Reservoir from Universidade Estadual de Campinas, in 1997, and Ph.D. in Regional Geology from UNESP, in 2003. Post-doctoral studies in the Departamento de Geologia Aplicada of Universidade Estadual Paulista, from 2002 to 2003. Currently, works as Assistant Professor in the Departamento de Geologia e Recursos Naturais of the Instituto de Geociências da Universidade Estadual de Campinas. Has experience in Geology, with emphasis on Reservoir Geology.

Emilson Pereira Leite. Professor MS-3 in the Departamento de Geologia e Recursos Naturais of the Instituto de Geociências, UNICAMP. Graduated in Geophysics from Universidade de São Paulo (1997), Master's (2000) and Ph.D. (2005) degrees in Geophysics from Universidade de São Paulo, has also worked as a trainee in the Departamento de Ciências Espaciais e da Terra e Engenharia of York University, Canada. Has experience in Geophysics and works mainly with: inversion of geophysical data, integration of geophysical data from geological and remote sensors applied to mineral and hydrocarbon exploration.

Rodrigo Duarte Drummond. Graduated in Computational and Applied Mathematics from Universidade Estadual de Campinas, in 1995. Obtained the Ph.D. degree in Genetics and Molecular Biology, with emphasis in Bioinformática, from the same university, in 2007. Worked at Scylla Bioinformática and currently, works as a researcher in the Centro de Estudos do Petróleo of Universidade Estadual de Campinas. Interested in computational data analysis, with emphasis on gene expression and geophysical data.