# Facies classification in well logs of the Namorado oilfield using Support Vector Machine algorithm.

Alexsandro G. C. *, CPGG/UFBA, Carlos A. C. da P., UFBA, & Geraldo G. N., Hydrolog Serviços de Perfilagens Ltda



This paper was prepared for presentation at the 15th International Congress of the Brazilian Geophysical Society, held in Rio de Janeiro, Brazil, 31 July to 3 August 2017, 2017.



## Abstract

**In this work we propose an alternative form to interpret facies in well logs using the suport vector machine algorithm (SVM), especifically, the suport vector classifier (SVC) present in the scikit-learn package of Python programming language. The dataset used was the well log data of two different wells, from Namorado field in Campos basin, Brazil, labeled with our interpretation and some informations of the core Data.**

**Two faciological interpretations were made in wells NA02 and NA04, that contain a turbidity sandstone as reservoir. These wells were used in the training of the machine learning algorithm to predict the five main types of facies of the well log data.**

**The goal is to build a model of the distribution of lithological labels in terms of four predictors features: Gamma-ray, resistivity, density and neutron porosity logs, using the classifier algorithm in well NA07.**

## Introduction

Well logs present a detailed plot of formation parameters versus depth. From the plots, interpreters can identify lithologies, differentiate between porous and nonporous rock and quickly recognize pay zones in subsurface formations. The ability to interpret a log lies in recognizing the significance of each measurement (Varhaug, 2016). In many practical cases, particularly in offshore oil exploration, the absence of outcrops combined with limited number of core well collaborate with the insufficient geological support to identify the facies of interest to all boreholes in the oilfield to supply the geologic information needed for formation evaluation (Lopes & Andrade, 2013).

A variety of multivariate statistical approaches, supervised or unsupervised, to the problem of facies classification have been applied using wire-line log curves measurements, such as: K nearest neighbor (KNN), fuzzy logic, neural networks and others (Dubois et al., 2007). Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and other learning tasks. This algorithm is a generalization of a simple and intuitive classifier called the maximal margin classifier (James et al., 2015). There are applications of this technique to recognize seismic data patterns for exploration and reservoir characterization applications. The SVM is recommended to classify data with nonlinear characteristics.

A different approach were applied to interpret facies from the classification algorithm (SVC), and were confronted with the conventional interpretation, to evaluate the performance of the algorithm in this problem .

## Theory

A support vector machine constructs one or a set of hyperplane in a high or infinite dimensional space, which can be used for classification, regression among other tasks. In general, if our data can be separated by hyperplanes, there will be an infinity number of modes to do it. A natural choice is the maximal margin hyperplane (also known as the optimal separating hyperplane), which is the separating hyperplane that is farthest from the training observations (James et al., 2015).

First, we are going to introduce the concept of maximal margin classifier.

*Maximal Margin Classifier*

The mathematical definition of a hyperplane is given by the Eq.(1):

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0, \quad (1)$$

and suppose that we have a $n \times p$ data matrix **X** that consists of $n$ training observations in $p$-dimensional space,

$$x_1 = \begin{pmatrix} x_{11} \\ ... \\ x_{1p} \end{pmatrix}, ..., x_n = \begin{pmatrix} x_{n1} \\ ... \\ x_{np} \end{pmatrix}, \quad (2)$$

These observations are going to fall into two classes: $y_1, ..., y_n \in \{-1, 1\}$, where $-1$ represents one class and $1$ represent the other class. Assume that it is possible to build a hyperplane that separates the training observations perfectly according to their class labels. The natural choice to separate the training observation is the *maximal margin hyperplane*, which is the separating hyperplane that is farthest from the training observations, shown in Figure 1.

Analysing the Figure 1, we percept that three training observations are equidistant from the maximal margin hyperplane and lie along the dashed lines indicating the width of the margin. These observations are known as support vector, since they are vectors in p-dimensional space (in Figure 1, $p = 2$). The maximal margin hyperplane depends directly on the support vectors, but not of the other observations.

In this work there will be more than two classes.

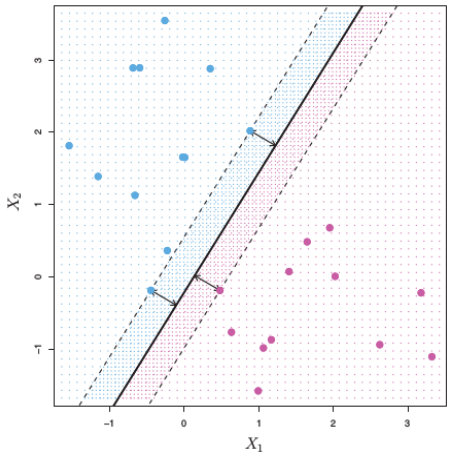


Figure 1: The are two classes of observations, shown in blue and in purple. The maximal margin hyperplane is shown as a solid line. The margin is the distance between one of the dashed lines and the solid line. An Introduction to Statistical Learning with Application in R (James et al. 2015).

*Support Vector Classifier (SVC)*

The SVC is a powerful classification technique proposed by Cortes & Vapnik (1995). The optimization criterion here is the width of the margin between the classes, and the empty area around the decision boundary defined by the distance to the nearest pattern (Bagheri & Riahi, 2013).

Considering the task of constructing the maximal margin hyperplane based on a set of $n$ observations $X_1,...,X_n \in \Re^p$ associated to the class labels $y_1,...,y_n \in \{-1,1\}$. A linear function can be defined as follows:

$$f(x) = W^t X_n + b, \tag{3}$$

where,

$$W^t X_n + b \geq 1 \qquad if \quad y_n = +1 \quad for\ all\ n, \tag{4}$$

and

$$W^t X_n + b \leq 1 \qquad if \quad y_n = -1 \quad for\ all\ n. \tag{5}$$

$W$ is the gradient vector and is perpendicular to the plane in the $\chi$ space. The Eqs. (4) and (5) can be rewritten as Eq. (6):

$$y_n(W^t X_n + b) \geq 1. \tag{6}$$

The square of the margin is inversely proportional to:

$$\|W\|^2 = W^t W, \tag{7}$$

and to maximize the margin, we have to minimize $\|W\|^2$.

To do this we use a standard optimization technique. We construct a Lagrangian:

$$L(W,b,\alpha) = \frac{1}{2}\|W\|^2 + \sum_{n=1}^{N_s} \alpha_n (y_n(W^t X_n + b) - 1), \quad \alpha_n \geq 0, \tag{8}$$

and $L$ must be minimized with respect $W$ and $b$, and maximize for each $\alpha_n$. $\alpha_n$ are the Lagrange multipliers.

Derivating $L$ with respect to $W$ and $b$, we obtain the Eqs. (9) and (10):

$$\frac{\partial L(W,b,\alpha)}{\partial W} = \left(W - \sum_{n=1}^{N_s} \alpha_n y_n X_n\right) = 0, \tag{9}$$

$$\frac{\partial L(W,b,\alpha)}{\partial b} = \sum_{n=1}^{N_s} y_n \alpha_n = 0, \tag{10}$$

and from Eq. (9) we can written:

$$W = \sum_{n=1}^{N_s} \alpha_n y_n X_n. \tag{11}$$

Now, we can get the following equation:

$$L(\alpha) = \sum_{n=1}^{N_s} \alpha_n - \frac{1}{2} \sum_{n=1}^{N_s} \sum_{m=1}^{N_s} y_n y_m \alpha_n \alpha_m X_n^t X_m, \tag{12}$$

or

$$L(\alpha) = \sum_{n=1}^{N_s} \alpha_n - \frac{1}{2} \sum_{n=1}^{N_s} \sum_{m=1}^{N_s} \alpha_n Q \alpha_m, \tag{13}$$

and

$$Q = \begin{bmatrix} y_1 y_1 X_1^t X_1 & y_1 y_2 X_1^t X_2 & \dots & y_1 y_{N_s} X_1^t X_{N_s} \\ y_2 y_1 X_2^t X_1 & y_2 y_2 X_2^t X_2 & \dots & y_2 y_{N_s} X_2^t X_{N_s} \\ \dots & \dots & \dots & \dots \\ y_{N_s} y_1 X_{N_s}^t X_1 & y_{N_s} y_2 X_{N_s}^t X_2 & \dots & y_{N_s} y_{N_s} X_{N_s}^t X_{N_s} \end{bmatrix}. \tag{14}$$

$L$ must be maximized with respect to the $\alpha_n$. $Q$ is called quadratic coefficients matrix.

Cortes and Vapnik (1995) solves the following primal optimization problem (Eq. (15)):

$$\begin{gathered} \min_{W,b,\varepsilon_n} \tfrac{1}{2} W^t W + C \textstyle\sum_{n=1}^{N_s} \varepsilon_n \\ subject\ to\ y_n(W^t \phi(X_n) + b) \geq 1 - \varepsilon_n, \\ \varepsilon_n \leq C \ \ and \ \ \varepsilon_n \geq 0 \end{gathered} \tag{15}$$

where $\phi(X_n)$ maps $X_n$ into a higher-dimensional space and $C > 0$ is the regularization parameter. Due the possible high dimensionality of the vector variable $W$, usualy we solve the following dual problem:

$$\begin{aligned} &\min_{\alpha} \quad \tfrac{1}{2}\alpha^{\mathbf{t}}\mathbf{Q}\alpha - \mathbf{u}^{\mathbf{t}}\alpha \\ &subject\ to\ \mathbf{y}^{\mathbf{t}}\alpha = \mathbf{0} \\ &0 \le \alpha_i \le C, \quad n = 1,2,...,N_s \end{aligned} \tag{16}$$

$\mathbf{u} = [\mathbf{1},...,\mathbf{1}]^{\mathbf{t}}$ is the unit vector, Q will be rewritten as the matrix $Q_{n,m} = y_n y_m K(X_n, X_m)$, where $K(X_n, X_m)$ is the kernel function, defined by Eq. (17):

$$K(X_n, X_m) = \phi(X_n)^t \phi(X_m). \tag{17}$$

There are many examples of kernels function. In this work is used the gaussian RBF kernel (Eq. (18)).

$$K(X_n, X_m) = exp(-\gamma \|X_n - X_m\|^2) \tag{18}$$

The decision function is:

$$sgn(W^t \phi(X) + b) = sgn\left(\sum_{n=1}^{N_s} y_n \alpha_n K(X_n, X) + b\right). \tag{19}$$

## Methodology

The wells NA04 and NA02 were used as training data. This set of data were labeled using the flowchart proposed by Nery (2013), shown in the Figure 2, to define the oil zones and the shales zones, and facies information from the core data.

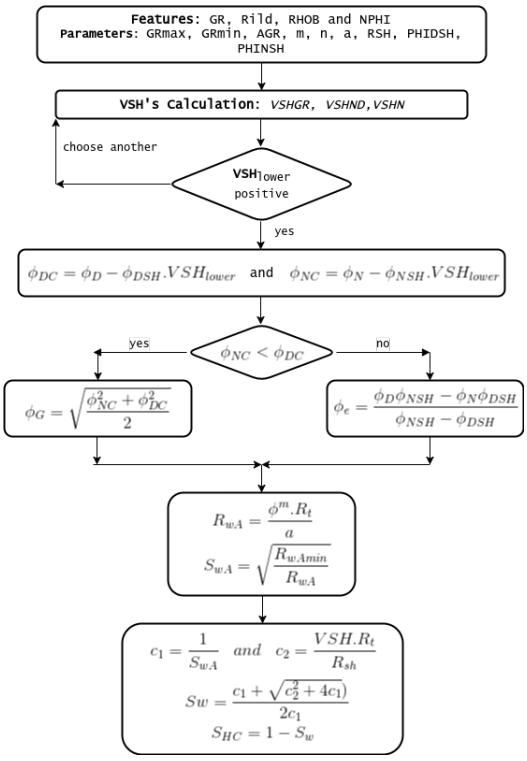


Figure 2: Flowchart used to calculate Saturation in hydrocarbon.

The twenty facies, presented in the well, were grouped into five facies and labelled as shown in Table 1:

| Facies | Label |
|---|---|
| Shale | 1 |
| Marl | 2 |
| Sandstone with HC with $S_{HC} \geq 45\%$ | 3 |
| Sandstone with $S_{HC} < 45\%$ | 4 |
| Carbonates | 5 |

Table 1: Facies used in the labeling of training data of the Namorado oilfield. This table contains a summarize of the main lithologies present at the well.

Figures 3 and 4 display the interpretation of the well NA02 and NA04.

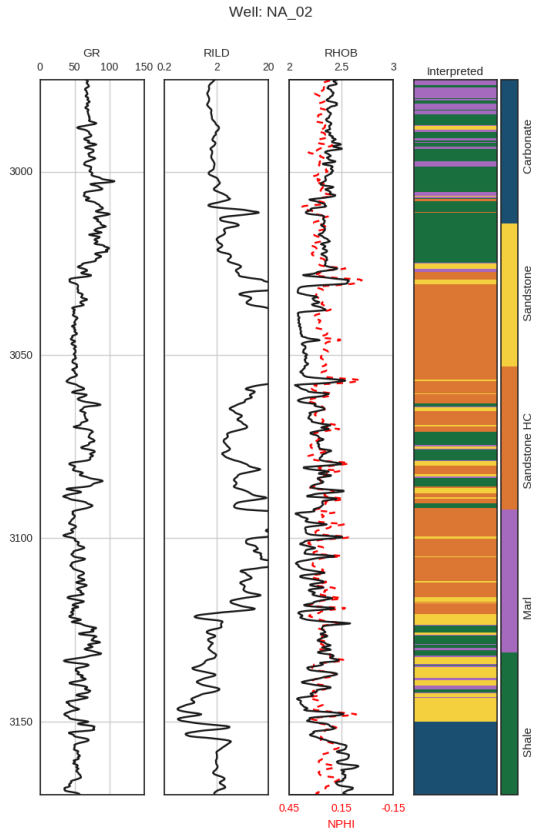


Figure 3: Well logs with the facies interpretation of the well NA02 made using the flowchart and informations of the core.

After labeling the facies in the well in function of the characteristics: gamma-ray, resistivity, density and neutron porosity, the training data was fit to build the SVC predictor to the algorithm interpret the blind test (well NA07).

To evaluate the quality of the prediction we used the $F1_{score}$ equation, defined by:

$$F1_{score} = 2\frac{precision \times recall}{precision + recall}. \tag{20}$$

The Eqs. (21) and (22) show the *precision* and *recall*,

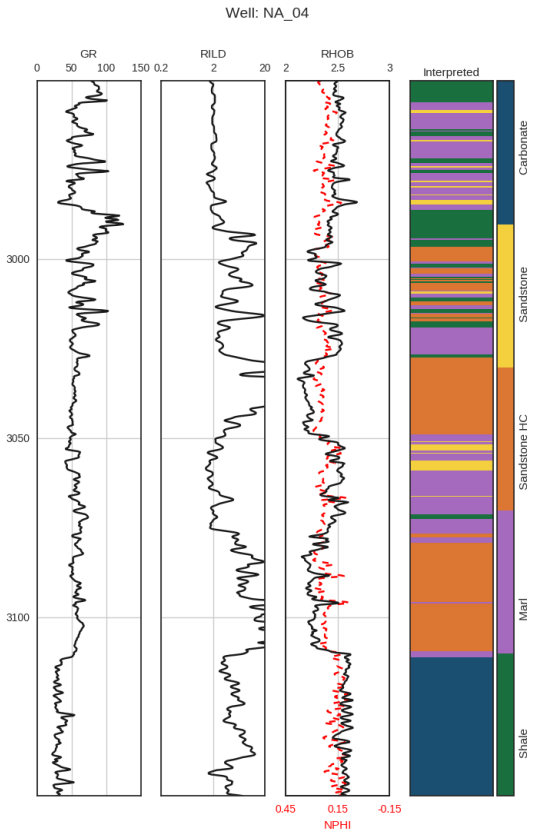


Figure 4: Well logs with the facies interpretation of the well NA04 made using the flowchart and informations of the core.

respectively:

$$precision = \frac{t_p}{t_p + f_p}, \tag{21}$$

$$recall = \frac{t_p}{t_p + f_n}, \tag{22}$$

where $t_p$ is the number of true positive of the class, $f_p$ is the number of false positive and $f_n$ the falses negatives.

**Results**

To evaluate the behavior of the physical properties in relation to the facies, a set of crossplot are made to analysis the distribution of the labels, shown in the Figure 5:

The predictor variables are the four wireline values.

After we trained the classifier using the training set to creat our model, we could predict the facies in a set of data test selected by a module of the Scikit-learn package to evaluate the accuracy of the classifier. Was selected $10\%$ of the training data to be our test data, and the results about the performace of the predictor is shown in Table 2.

The Table 2 shows that the predictor has difficulty in classify sandstone in the test data.

In the SVC algorithm were used the parameters $C = 1000$ and $\gamma = 0.1$ on the fit of the data set.


Figure 5: Crossplots of the distribution of the lithologies for the training data. The main diagonal shows a histogram of each property evaluated.

| Lithology | $precision$ | $recall$ | $F1_{score}$ |
|---|---|---|---|
| Shale | 0.94 | 1.00 | 0.97 |
| Marl | 0.76 | 0.86 | 0.81 |
| Sandstone HC | 0.95 | 0.95 | 0.95 |
| Sandstone | 0.69 | 0.48 | 0.56 |
| Carbonate | 0.88 | 0.85 | 0.87 |
| Average | 0.87 | 0.88 | 0.87 |

Table 2: Parameters related to prediction quality in the test data in each facies. Were selected $10\%$ of the total set of training data.

The well NA07 was chosen to be applied to the blind test. Applying the predictor in the features of the blind data, the algorithm made the interpretation shown in Figure 6, that also contains the interpretation using the flowchart and the core information.

The Table 3 presents the parameters related to prediction quality in the blind test compared with the prediction.

| Lithology | $precision$ | $recall$ | $F1_{score}$ |
|---|---|---|---|
| Shale | 0.99 | 0.88 | 0.93 |
| Marl | 0.17 | 0.38 | 0.24 |
| Sandstone HC | 0.78 | 0.84 | 0.81 |
| Sandstone | 0.76 | 0.59 | 0.66 |
| Carbonate | 0.88 | 0.96 | 0.92 |
| Average | 0.84 | 0.80 | 0.81 |

Table 3: Parameters related to prediction quality in the blind data in each facies compared with the prediction.

The prediction was more accurate to delimitate the carbonate limit, shown in the final depth of the well. Some facies, present in the well, were not informed to the algorithm due the simplified interpretation. In the interval between $3100m$ and $3150m$ some facies are classified as carbonate while in the interpretation made

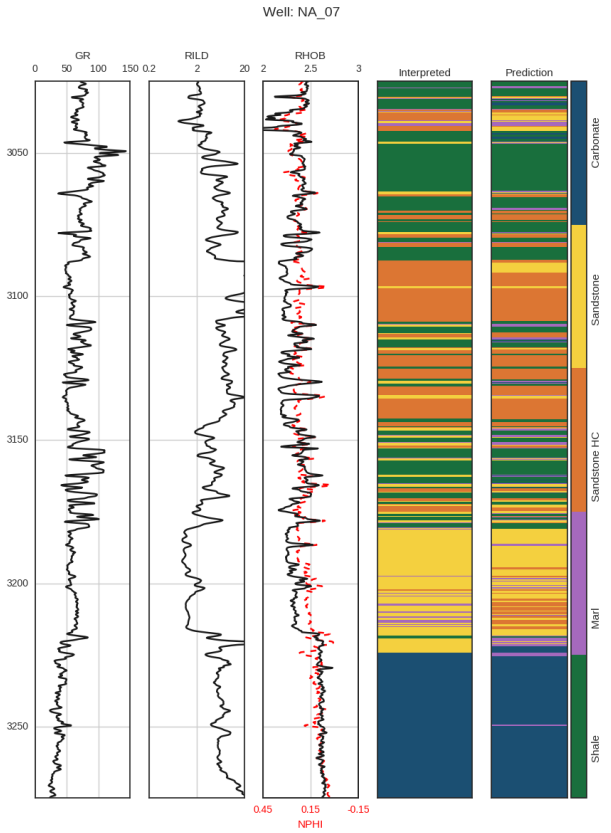


Figure 6: Features with the facies interpretation of the well NA07. The fourth track shows the interpretation made by the interpret and the fifth track shows the interpretation made by the SVC algorithm.

by us indicated as sandstone. In this case, the classifier was not completly wrong because in this interval there is occurrence of sandstone cemented with calcite. Near $3150m$ the algorithm predicted intercalation of sandstone, shale and marl, coherent with the core information.

It is possible to affirm that the algorithm presents a certain difficulty in distinguishing sandstone and marl, possibly due to the accuracy of the training data that does not cover with detail the whole well.

The Figures 7 and 8 show the behavior of the properties in the well facies interpreted by the interpreter and the prediction in the well NA07. It is noticeable to notice that the properties are distributed in a similar way in both figures, showing the efficiency of the SVC algorithm in classifying of the facies in the well NA07.

## Conclusions

Analysing the results obtained by the SVC algorithm was possible to say that the prediction made by algorithm was similar to the interpretation made by the interpreters. In some cases, the algorithm shows more efficiency to descrimanate some typical lithological answers in the well logs, for example marls and carbonates.

To formulate more precise interpretations, a larger database is needed with accurate information of the positions of the facies. Probably the algorithm will present

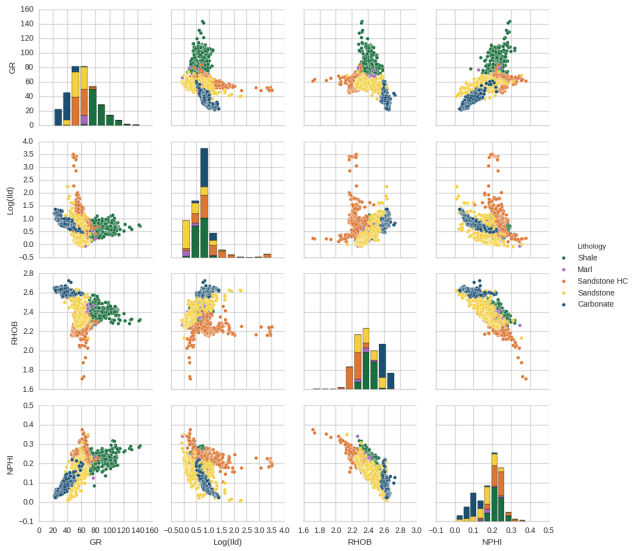


Figure 7: Crossplots of the distribution of the lithologies in relation the well logs of the interpretation made by interpreter in the well NA07. The main diagonal shows a histogram of each property evaluated.

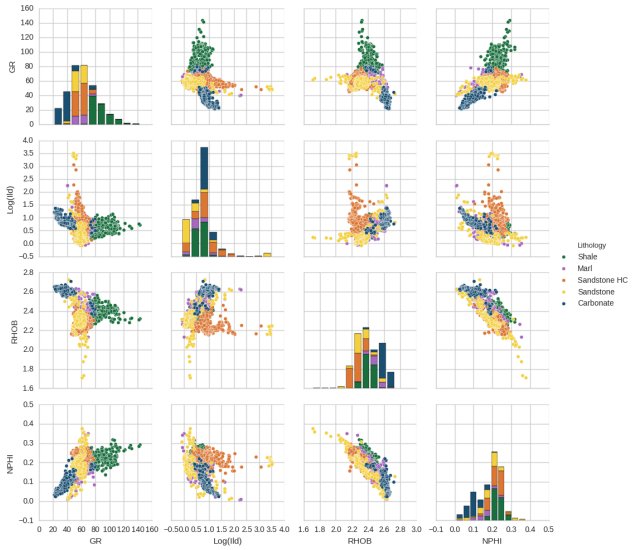


Figure 8: Crossplots of the distribution of the lithologies in relation the well logs of the interpretation made by SVC algorithm in the well NA07. The main diagonal shows a histogram of each property evaluated.

a smaller precision in the faciologial interpretation, but will show a more realistic distribution of the present facies.

In general, the SVC classifier showed efficiency in this study, presenting a relatively high value of $F1_{score}$, which shows that the algorithm learned and applied the interpretations taught to it in an adequate way.

## Acknowledgements

The facility supports CPGG / UFBA and ANP for providing the date is also acknowledged.

**References**

Bagheri, M. & Riahi, M. A. (2013) Support Vector Machine-based Facies Classification Using Seismic Attributes in an Oil Field of Iran. Iranian Journal of Oil & Gas Science and Technology, Vol. 2, 01-10.

Cortes, C. & Vapnik, V. (1995) Support-Vector Networks. Machine Learning, Vol. 20, 273-297.

Dubois, M. K., Bohling, G.C. & Chakrabarti, S. (2007) Comparison of four approaches to a rock facies classification problem. Computers & Geosciences, Vol. 33, 599-617.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2015). An Introduction to Statistical Learning with Application in R. Spriger Science+Business Media New York.

Lopes, D. M. R. & Andrade, A. (2013) Facies identification by Fuzzy Inference. International Congress of the Brazilian Geophysical Society.

Nery, G. G. (2013) Perfilagem Geofísica em Poço Aberto: Fundamentos básicos com ênfase em petróleo. SBGf.

Varhaug, M.(2016) Basic Well Log Interpretation. Oilfield Review, Schlumberger. 52-53.