PREDICTION OF DEPOSITIONAL FACIES USING DATA MINING ON WELL LOGS FROM NAMORADO OIL FIELD, OFFSHORE BRAZIL

Lucas Lima de Carvalho

Laura Lima Angelo dos Santos

Luana Karina Câmara de Medeiros

Edmilson Helton Rios

Paulo Couto

lucascarvalho@poli.ufrj.br

lauralimaang@gmail.com

luana.medeiros@outlook.com

edmilson@petroleo.ufrj.br

pcouto@petroleo.ufrj.br

Programa de Engenharia Civil, COPPE/Universidade Federal do Rio de Janeiro, Cidade Universitária, Rio de Janeiro, Brazil.

Nadege Bize-Forest

NBize-forest@slb.com

Schlumberger Brazil Research & Geoengeneering Center, Parque Tecnológico, Cidade Universitária, Rio de Janeiro, Brazil.

Abstract. *Geophysical well logging is performed in almost all wells drilled for petroleum exploration. They are important for formation evaluation, well integrity, reservoir surveillance and reserve calculations. Accurate and high resolution identification of geological facies crossed by a well is only possible with the description of rock samples coming from drill cuttings, whole cores or lateral plugs. However, these technics are very expensive, time consuming and limited to specific intervals. This paper describes then how to predict geological facies using data mining techniques applied to conventional wireline logs, such as gamma ray, resistivity, density*

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

and neutron. Supported Vector Machine and Random Forest classification algorithms were trained using the raw and preprocessed dataset of the cored interval of the Namorado Oil Field, one of the most productive turbidite reservoirs in Campos Basin, Southeast Brazil.

Key words: *Depositional Facies, Namorado Sandstone, Machine Learning, Supported Vector Machine, Random Forest*

1. INTRODUCTION

Well log responses are indirect measures that are commonly applied on the interpretation of multiple properties of the rock and the fluid contained on its pore space. For this study, the logs were used on the prediction of different rock types, known as facies. The models were trained and validated on intervals where the facies were already defined, through the direct description of the whole cores.

The paper proposes a workflow focusing on data management that intends to prepare the log data sets and the facies to predict. To do so, it was applied several methods that overcome the conventional statistical analysis and suit better to our dataset type.

Within this purpose, the chosen facies classification found in the literature for the Brazilian turbidites of Namorado Field was the one suggested by Barboza (2005) who simplified 21 depositional lithofacies described by Zarpelon et al. (1997 apud Barboza, 2005) into 7 facies only. Other papers have already attempted to apply machine learning methods on Namorado dataset to predict facies. Lammoglia et al., 2014 used Fuzzy Logic and Neural Networks with the objective to recognize only three classes of rocks (shale, sandstone and limestone) and Alexsandro et al. (2017) applied Support Vector Machine algorithm to identify 5 rock classes (shale, marl, sandstone with Hydrocarbon, sandstone without HC and carbonates). We propose here a new machine learning workflow to identify the 7 facies described by Barboza in 2005.

The dataset contains 4 wireline logs (gamma ray, resistivity, density and neutron) and whole core descriptions from 17 wells located in Namorado Field (Campos Basin, offshore Brazil).

Data preparation and data analysis were fundamental and main part of the workflow has involved data cleaning, transformation, integration, normalization, facies distribution analysis and Principal Component Analysis (PCA). These steps were carefully executed to reduce noise from the original dataset and optimize the information taken from it. Lastly, two machine learning algorithms, Support Vector Machine and Random Forest were applied into the original and improved datasets to generate the facies model.

2. GEOLOGICAL AND DATA SETTINGS

Namorado Oil Field, discovered in 1975, is part of the Campos Basin, located in the northern coast of Rio de Janeiro State. The post-salt reservoirs are marine turbidites related to the first important transgression over the Albian limestone shelf.

Studied intervals range from 2900 m to 3400 m, and include the passive margin deposits from the Albian carbonate platforms overlapped by Albian and Cenomanian gravity flows, such as marine turbidites and debris flows (Zarpelon et al., 1997 apud Barboza, 2005).

The whole cores are labeled in 7 different depositional facies (Table 1), Barboza (2005) based on the combination of the 21 facies originally described by Zarpelon (1997). Each facies is related to specific shallow water platforms and deep water slope environments. The facies classification is mainly based on the texture and granulometry of the rock reflecting the energy of the depositional environment. It does not consider the rock composition itself which adds an extra challenge to the facies prediction workflow, as some logs are sensitive to the chemical composition of the formation and not the granulometry. 742 m of whole core description are available from 17 wells located in Namorado Field covering 3870 m of reservoir.

Table 1: Depositional facies of the Namarado reservoir as described by Barboza (2005).

Label	Description
1	Conglomerates
2	Stratified and Massive Sandstones
3	Turbidites (thick layers)
4	Turbidites (thin layers)
5	Matrix-Supported Conglomerates
6	Debris-Flow Deposits
7	Mudrocks

The wireline logs available for the wells are gamma ray, resistivity, density and neutron with a 0.2 m sampling rate.

The density log (RHOB) is the formation bulk density, based on the mass to volume ratio of a measured interval. The measurement is derived from the electron density of the formation, obtained by radiating the formation with gamma rays and measuring the number of collisions between the gamma rays and the formation electrons (Schlumberger, 2016).

The neutron log (NPHI) log measures the hydrogen concentration on the formation. The tool emits neutrons into the formation that interact with the hydrogen atoms and loose energy. The hydrogen concentration (or index) is then converted to neutron porosity (Schlumberger, 2016).

The gamma ray log (GR) responds to naturally occurring radioactive elements (potassium, uranium and thorium). Clay-rich formations and arkose sandstones usually produce more radiations than carbonate and quartz-rich formation (Schlumberger, 2016).

The resistivity log (ILD) is a measure of the degree to which the rock and fluid can imped the flow of an electric current. It provides a strong response for the type of fluid contained in the pore space of the rock. Saline water provides a low values of resistivity and hydrocarbons, hig values (Schlumberger, 2016).

3. METHODOLOGY

3.1. Data analysis

Scatter Plot - Scatter plot (Figure 1) is the way to see all possible plots 2vs2, for each input logs colour coded by the Facies label. One interesting result observed here is that RHOB separates well the two more present facies L2 and L7.

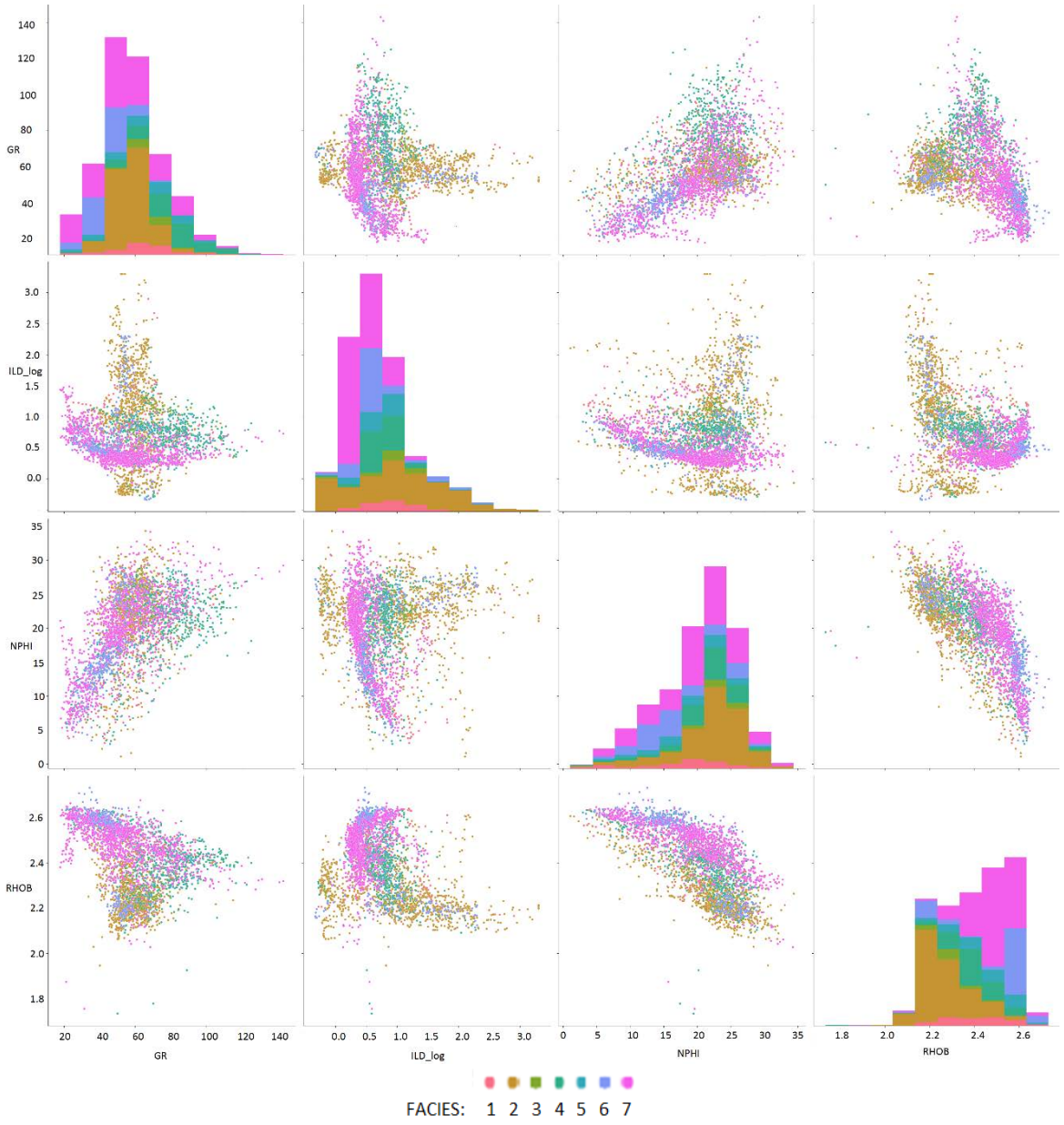

Figure 1: Matrix of crossplots with the distribution of facies for the available data set. The main diagonal shows histograms for each well log: gamma ray, resistivity, neutron and density.

Principal component analysis (PCA) **-** The principal component analysis is a method used for dimensional reduction of a large dataset. This method makes the calculation of a new base for the given vectorial space, making orthogonal transformations. So, given one dataset with dimension n and k samples, the method find a matrix $P_{n,m}$, where Y=XP, Y being a dataset with m dimensions and k samples.

The transformation represents the maximum variance possible in the first component, then the second maximum variance in the second component and subsequently the x maximum variance in the x component.

This method is widely applied to visualize the data given in a bigger dimension in 3 or 2 dimensions, and also to reduce the calculation time in unsupervised methods.

The steps to generate the matrix $P_{n,m}$ are the following:

a) Calculate the covariance matrix of the normalized data:

$$C_{i,j} = \frac{1}{N-1} * \sum_{q=1}^{N} X_{q,i} * X_{q,j}, \quad (1)$$

b) Calculate eigenvalues and eigenvectors;

c) P = [v1, v2], v1 and v2 being the eigenvectors corresponding to the bigger eigenvalues.

In our dataset 83% of the variance is supported by the first two components (Figure 2). The plot shows that different facies are not well clustered, and can not even have a well-defined boundary.

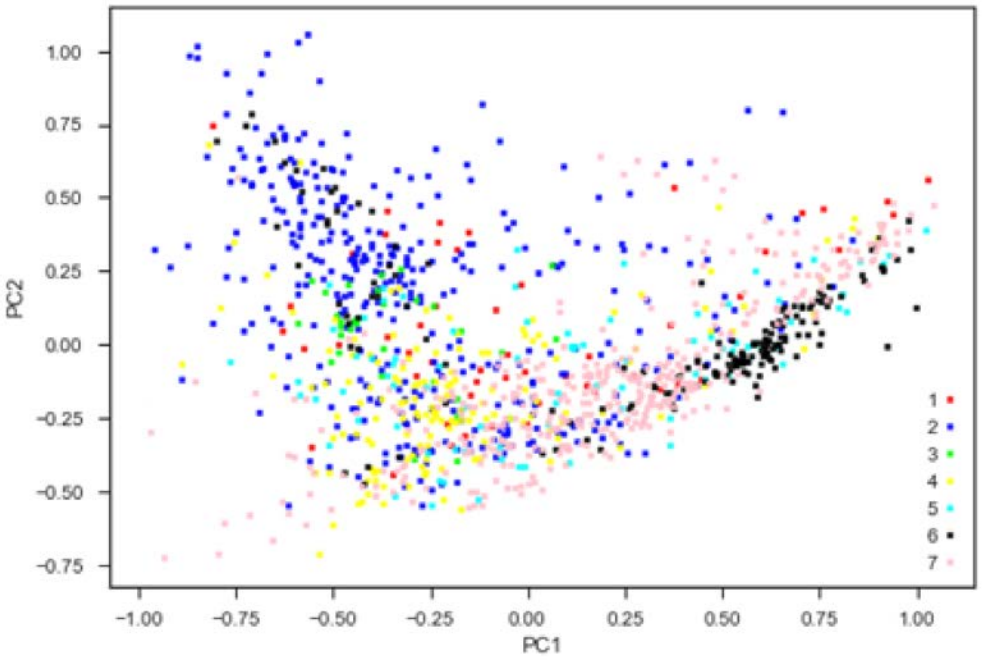

Figure 2: Cross plot of principal components showing the facies label.

3.2. Data preprocessing

Data Cleaning – Boundary Removal - Inherent to large datasets, facies class noise may occur when a facies is inaccurately labeled. It can be caused by the subjectivity of the labeling process done by different geologists describing the cores or by data entry errors. As a solution to avoid class noises and to improve the dataset quality, the label boundaries were removed. Other difficulty faced was the depth shift issues between core description and log data, especially in the transition zone of two different facies. In this case study, 5 cm was removed at the bottom and top of each facies boundary.

Data Integration – It corresponds to the processing and merging of log data, carefully done avoiding redundancies and inconsistencies of the physical measurements (García et al., 2015). For this dataset, the neutron porosity (NPHI) and bulk density (RHOB) were combined using linear transformations (Equation 2) to calculate the "shale volume" (Vsh). Vsh is commonly calculated considering the lower and higher GR values, with a direct relation between regular shales and high gamma ray values. Nevertheless, in this reservoir, this relationship between gammaray and shale content is not direct considering two facts: some sandstones provide high GR response due to its high potassium content; and facies L7 (Mudrocks) are not only shales, but mudstones and marls, clay-size particles with low gamma ray response as they are mainly carbonate lithology.

VSH_ND – A shale volume curve was generated by using the linear response (equations bellow) and using default values for neutron porosity (NPHI) and density (RHOB) values for the matrix, fluid and shale (Schlumberger, 2009).

$$VSH\_ND = \frac{X_1 - X_0}{X_2 - X_0}, \tag{2}$$

Where $NPHI$ is the neutron porosity log response, $NPHI_{MA}$ and $NPHI_{Sh}$ are the standard value for the neutron porosity of respectively the sandstones and the mudrocks. $RHOB$ is the density log response, and $RHOB_{MA}$ and $RHOB_{Sh}$ are the standard values for the densities of the sandstone and the mudrocks. $X_0 = NPHI_{MA}$

$$X_1 = NPHI + M_1 \times (RHOB_{MA} - RHOB)$$

$$X_2 = NPHI_{Sh} + M_1 \times (RHOB_{MA} - RHOB_{Sh})$$

$$M_1 = \frac{NPHI_{FL} - NPHI_{MA}}{RHOB_{FL} - RHOB_{MA}}$$

Data Transformation – The dataset was divided in different zones (Zonation) based on the similarity of the behavior of neutron porosity and bulk density curves among the wells. This correlation was suggested by Faria et al. (2001 apud Barboza, 2005), for the whole reservoir, to identify five different stratigraphic sequences. These stratigraphic formations always appear in the same order, but not all are present in all wells. Figure 3 represents the different distribution of the facies through all the zones.

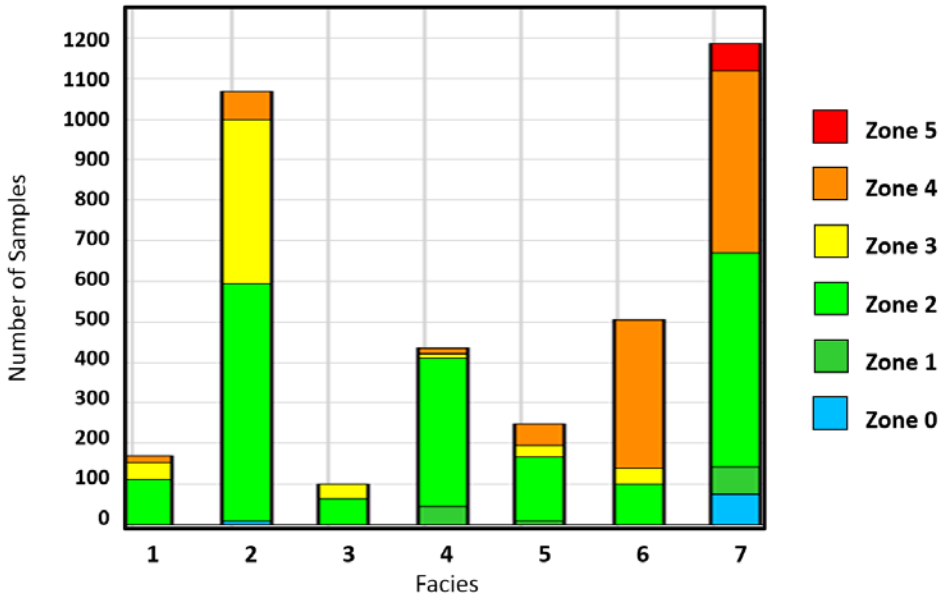

Figure 3: Histogram of facies distribution by stratigraphic zonation.

Data Normalization - The data are normalized using centering and standard deviation, meaning that we subtract each value by the mean of the data set and divide it by the standard deviation.

$$\mu = \frac{1}{N}\sum_{q=1}^{N} X_n \tag{3}$$

$$\epsilon = \sqrt{\frac{\sum_{q=1}^{N}(X_q - \mu)^2}{N-1}}$$

N being the number of samples, and Xq one sample:

$$Y_q = \frac{(X_q - \mu)}{\epsilon}$$

For the resistivity log (ILD) the normalization was applied after logarithmic transformation (base 10).

3.3. Supervised Learning Methods

For this study, we worked with several machine learning methods, most of them using scikit-learning library. This section will resume the theory of the algorithms used.

Support Vector Machine (SVM) - The Support Vector Classifier is a method to find a hyperplane between two perfectly separated classes, maximizing the margin between the planes

containing the samples, minimizing the number of misclassifications. Non-linear limits between classes can be obtained with the addition of different kernels, as showed in Hastie et al. (2009).

Random Forest (RF) - A random forest is a classifier that consists of a collection of tree-structured classifiers {h(**x,θk**), k=1,…} where the { **θk** } are independent and identically distributed random vectors an each tree cast a unit vote for the most popular class at input **x**.

For the construction of each tree we make a bootstrap sampling of size n in a m size dataset with our without replacement (in our case we work with m=n with replacement).

After we grow the decision tree (in our case with max depth=4), and at each node, we randomly select d features without replacement. We then split the node maximizing the information gain with the given samples and features.

4. RESULTS AND DISCUSSION

The results were organized into two groups: 'Original' and 'Modified' datasets. The Original dataset comprehends the provided logs (GR, ILD, NPHI and RHOB) with no pre-processing. The Modified dataset comprehends the new data management workflow proposed in this work (GR, ILD_log, NPHI, RHOB, VSH_ND, Zonation). The models were validated according to the following scheme: the model is trained on 16 wells and tested on 1 well that is left out and this process is iteratively repeated so that all wells, one by one, are taken out as a test data set. The performances of the models are calculated through the mean of all validation tests. The percentage of each predicted facies (labels) as a function of the true facies is presented by confusion matrices in Figure 4 for both models and datasets. The accuracy, which is the ratio of true positive (corrected predicted) by total number of samples, is indicated in Table 2 for the 17 wells. SVM (with C = 1) and Random Forest (with depth = 4) algorithms provided average testing scores of 65 % and 63 %, respectively, with the original dataset and 68 % with the Modified dataset.

The increase of accuracy test between a model trained with the original dataset and a model using the modified dataset can be explained by several reasons. The new generated shale volume curve allows to better differentiate facies 7 (mudrocks) from the other facies, then the original GR indicator. The data zonation may reduce the confusion among facies 1, 5, 6 and 7, considering that facies 6 and 7 usually occur in different zones than facies 1 and 5. The boundary removal step as cleaned depth match issues and delete some outliers.

Random Forest presented better data distribution, because the method could predict 4 of the 7 facies; while SVM was able to predict only 3 of the 7 (Figure 4 and 5). As expected, the less frequent facies in the training dataset were inexistent or less predicted by both algorithms. This is well observed in the confusion matrix for facies 1, 3 and 5 either in the model using SVM or with Random Forest. Facies 2 and 7 more frequent were very well predicted. It seems that the frequency of the different facies in the training impacts the final results. Indeed, we note that all facies have been predicted by class 2 and class 7 because they represent together 70% of the full data set.

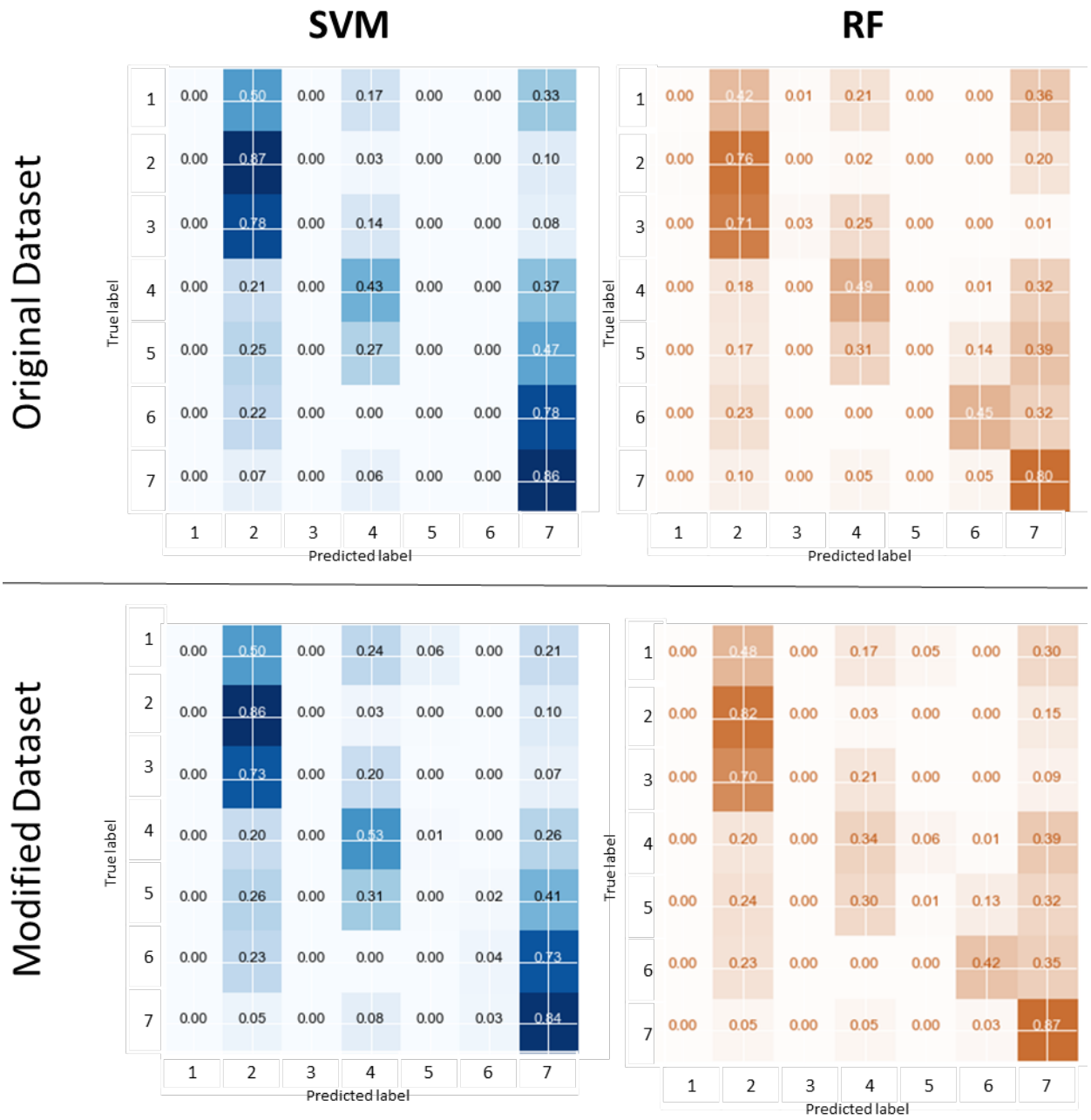

Figure 4*:* Confusion matrices with the prediction results of the depositional facies using SVM (blue) and Random Forest (orange) algorithms on the original and modified datasets.

The low score for predicting certain facies can be also explained by the geological and petrophysical definition of the facies themselves. For example, Facies 1 can be predicted by class 2 for 48 % and 30 % by class 7. Beside the low frequency of this facies to train the model, conglomerates are also a general term corresponding to a coarse-grained clastic sediment composed of gravel-size clasts and a matrix. Their petrophysical properties are then close to the Clastic sandstones of Facies 2 and the mudrock of facies 7. Facies 4 is well predicted for 34 % by class 4 bt in 39 % and 20 % on the time it is predicted by class 7 and class 2 respectively.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

Table 2: Accuracy results for the original and modified datasets using SVM and RF algorithms.

Well	Original dataset		Modified dataset	
	SVM	RF	SVM	RF
NA01A	0.65	0.63	0.70	0.69
NA02	0.66	0.67	0.68	0.67
NA04	0.78	0.78	0.80	0.80
NA05	0.34	0.36	0.34	0.35
NA07	0.47	0.52	0.54	0.45
NA11A	0.54	0.64	0.55	0.68
NA12	0.62	0.61	0.62	0.63
NA22	0.35	0.56	0.35	0.53
NA37D	0.93	0.90	0.93	0.93
NA40D	0.48	0.26	0.33	0.22
NA44D	1.00	1.00	0.47	1.00
NA47D	0.31	0.35	0.33	0.33
NA48D	0.87	0.82	0.87	0.87
NA53D	0.80	0.73	0.79	0.79
RJS019	0.69	0.65	0.75	0.68
RJS042	0.67	0.55	0.69	0.69
RJS234	0.61	0.58	0.69	0.73
AVERAGE	0.65	0.63	0.68	0.68

5. CONCLUSION

The non-predicted facies or not well predicted facies by the models is due to the heterogeneity in frequency of the different facies. Facies 2 and 7, the most frequency facies, over represented the models. The low frequency facies (1, 3, 5) are then often classified as 2 and 7. Although both algorithms gave good scoring, with accuracy above 65 %, Random Forest predicted more facies than SVM.

The workflow used for data management for prediction of facies increased by 5% the accuracy tests for the Random Forest or SVM models. Indeed, data cleaning, transformation, integration and normalization processes on the original curves reduced noise and enhance some of the petrophysical properties to react better to the different facies classes. A good understanding of the input data and the geological classification is then necessary to enhance all prediction models.

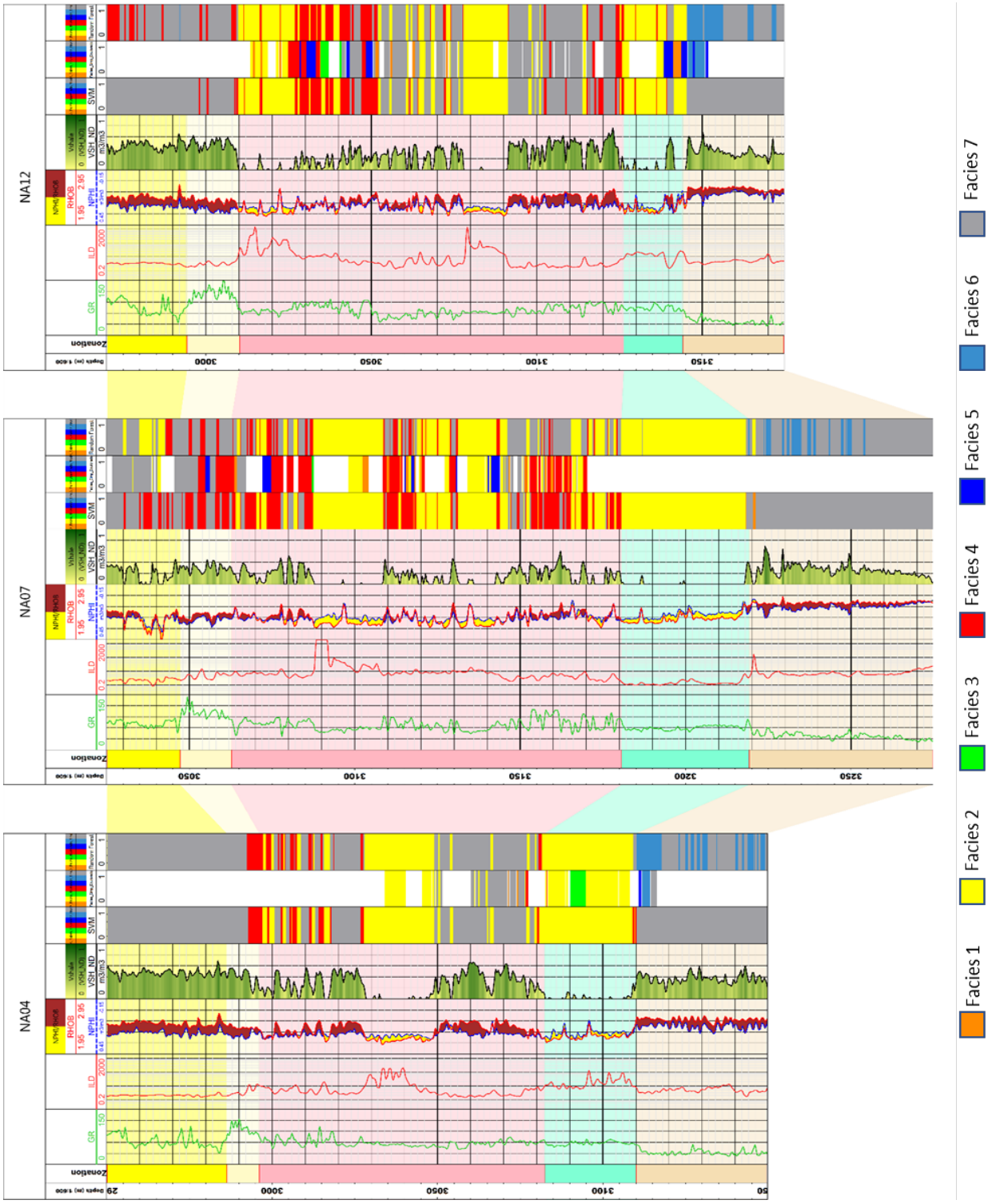

Figure 5: Well correlation for NA04, NA07 and NA12. From track 1 to 9: depth, zonation, GR, ILD, RHOB & NPHI, VSH_ND, SVM-based facies, facies from core descriptions, RF-based facies, respectively.

Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editores), ABMEC, Florianópolis, SC, Brazil, November 5-8, 2017.

Acknowledgements

We would like to thank ANP and Petrobras for funding (project number: 0050.0094368.14.9) and providing the data set for researches at LRAP/COPPE/UFRJ. We extend thanks to BRGC/Schlumberger for providing Techlog software and for research collaborations.

REFERENCES

Alexsandro G. C., Carlos A. C. da P., Geraldo G. N (2017). Facies classification in well logs of the Namorado oilfield using Support Vector Machine algorithm. *Fifteenth International Congress of The Brazilian Geophysical Society*.

Barboza, E. G. (2005). Análise estratigráfica do Campo de Namorado (Bacia de Campos) com base na interpretação sísmica tridimensional. *Phd teses, Universidade Federal do Rio Grande do Sul*.

García, S., Luengo, J., Herrera, F. (2015). Data Preparation Basic Models. *Data preprocessing in data mining*. pp. 39-55. Springer.

Hastie, T., Tibshirani, R., Friedman, J. (2009). Overview of Supervised Learning. *The Elements of Statistical Learning*. pp. 9–41. Springer.

Lammoglia, T., De Oliveira, J. K., & Souza Filho, C. R. (2014). Lithofacies recognition based on fuzzy logic and neural networks: a methodological comparison. *Revista Brasileira de Geofísica*, *32*(1), 85-95.

Schlumberger (2009). Log Interpretation Charts. Texas: Schlumberger.

Schlumberger (2016). The Defining Series: Basic Well Log Interpretation. *Oilfield Review 2016*