Brazilian Journal of Geophysics, 42 (2024), 2, 1-20
Brazilian Geophysical Society
ISSN 2764-8044
DOI: 10.22564/brjg.v42i2.2315

P-WAVE VELOCITY LOG SIMULATION FOR PETROELASTIC INVERSION USING GARDNER'S EQUATION, NEURAL NETWORKS AND ENSEMBLE METHODS BASED ON TREES: A CASE STUDY OF THE SANTOS BASIN, BRAZIL

Caique Pinheiro de Carvalho[1], Jéssica Lia Santos da Costa[2], Tobias Fonte-Boa[1], Tiago Amâncio Novo[1], Maria José Campos de Oliveira[1], and Fernanda Moura Costa[1]

[1]Universidade Federal de Minas Gerais - UFMG, Programa de Pós-graduação em Geologia, Instituto de Geociências, Departamento de Geologia, CPMTC-IGC, MG, Brazil
[2]Baker Hughes, Macaé, RJ, Brazil

*Corresponding author email: caique.pc94@gmail.com

ABSTRACT. One of the main tools for reservoir characterization is analyzing well-log data. The importance of such methods stems from petrophysical properties estimation, such as porosity, which is very important to the oil and gas industry. In scenarios where data is hard to collect, data loss and technical failures during the acquisition impose an extra challenge. Thus mathematical and petrophysical models are good candidates to fill information gaps in the well-log dataset. In such a way, the rock's petroelastic and petrophysical properties can be successfully estimated. Several studies correlate the velocity of compressional waves (V_P) to other basic well data. In this study, we used the Gardner equation and Machine Learning methods such as Neural Networks, Random Forest and Gradient Boosting regressions to generate V_P logs. We used real-world data acquired from twenty wells of the pre-salt formation from Santos Basin in Brazil to train and test the Machine Learning methods and evaluated the data estimated by those models using statistical metrics. We calculated the acoustic impedance from the estimated logs and used it to create a prior model for a petroelastic inversion, which allowed us to estimate the natural logarithm of the acoustic impedance for a seismic volume. The Machine Learning methods presented lesser errors between estimated and measured velocities when compared to Gardner's equation.

Keywords: well logging; artificial neural networks; machine learning; linear regression

INTRODUCTION

In oil field exploration, the sonic or acoustic log (DT) analysis is fundamental and one of the most powerful interpretation tools for petrophysical study. This method investigates the travel time of an elastic wave through the rock formation, which, among other applications (e.g., calibration of seismic data, identification of lithologies, stratigraphic correlation), is applied in petroelastic inversion and porosity calculation. In this situation, the P-wave speed can be correlated with some physical characteristics of the rock formation, such as porosity, pressure, type of rock matrix and fluid, and pore shape, crucial for determining potential production in hydrocarbon reservoirs.

Nevertheless, in deep-water fields where depths reach thousands of meters, technical failure during the well-data acquisition results in information loss, imposing an extra challenge to well-data interpretation. Yet, due

to specific relationships among rock physical properties (e.g. Gardner et al., 1974), missing acoustic data can be inferred through other well-log information, such as gamma-ray, density, neutron, and resistivity.

The speed of compressional waves (VP) can be correlated to other basic well data (e.g. Gardner et al., 1974; Oloruntobi and Butt, 2019; Carrasquilla et al., 2022; Carvalho et al., 2022). To understand how the P-wave is affected by these petrophysical properties, we applied three different estimation methods, the Gardner equation (Gardner et al., 1974), which relates the velocity of the compressional wave to the rock density, Neural networks (NN), one of the Machine Learning (ML) methods that have already been applied in geophysics (Lim and Kim, 2004; Rolon et al., 2009; Aleardi, 2015), and also ensemble methods based on trees, such as the Gradient Boosting and Random Forest regressors which are also applied in geoscience (Sahin, 2020).

In our work, we bring new data from one of the largest oil reservoirs in the world. We take advantage of the extensive well-log dataset from Santos Basin, focusing on Búzios pre-salt oil field (BOF), an ultra-deep-water reservoir along the Brazilian coast (Figure 1). Filling the gaps of P-wave velocity in our data allowed us to estimate the acoustic impedance for a seismic volume afterwards through a petroelastic inversion for a broad area, the acoustic impedance volume might then be applied for a porosity estimation using the same ML techniques. Another important application of our work is in the well to seismic tie process, the algorithm might fill the gaps where the P-wave velocity information is missing and the calculated acoustic impedance can then be used to calculate the reflectivity and the synthetic seismogram (de Macedo et al., 2020).

The data from the well-log curves was loaded and then split into training and test data to apply the ML methods and also on Gardner's equation, we considered the P-wave velocity as the labeled data for our regression, finally, we applied the statistical metrics to compare the results obtained with the different ML methods and with Gardner's equation. This work distinguishes itself from previous researches by the amount of data used for training the ML models and the set of combinations of input data for the P-wave velocity estimation for the Búzios oil field. To demonstrate the usefulness of our methodology, we show an acoustic impedance section at the end which was calculated after the P-wave velocity gaps were filled in the well log.

GEOLOGICAL CONTEXT

The Búzios oil field (BOF) is a highly productive and promising deep-water target found in the Santos Basin, one of the Brazilian continental margin's most extensive offshore petroleum reservoirs. Santos Basin is located within the Cabo Frio high and the Florianópolis platform along the coast of the Brazilian states of São Paulo and Rio de Janeiro (Figure 1). The BOF is situated seaward from the continental shelf slope where the current water depth reaches up to 3000 m, and its sedimentary column thickness can overtake more than 4000 m down from the ocean floor.

The Santos Basin developed during the evolution of the South Atlantic continental margin resulting from the Gondwana breakup event in the Late Jurassic-Early Cretaceous (Brune, 2016). The tectonic evolution of this basin can be divided into three main phases: (i) rift, (ii) post-rift (i.e. tectonic sag-phase), and (iii) drift (de Mio and Chang, 2005; Moreira et al., 2007).

In this study, we focus on the pre-salt section that is associated with the rift and post-rift stages. The rift stage is represented by lacustrine sediments (i.e. continental siliciclastics, talc-stevensite ooids with interbedded lacustrine coquinas and organic-rich shales) of the Camboriú, Piçarras, and Itapema formations. The post-rift

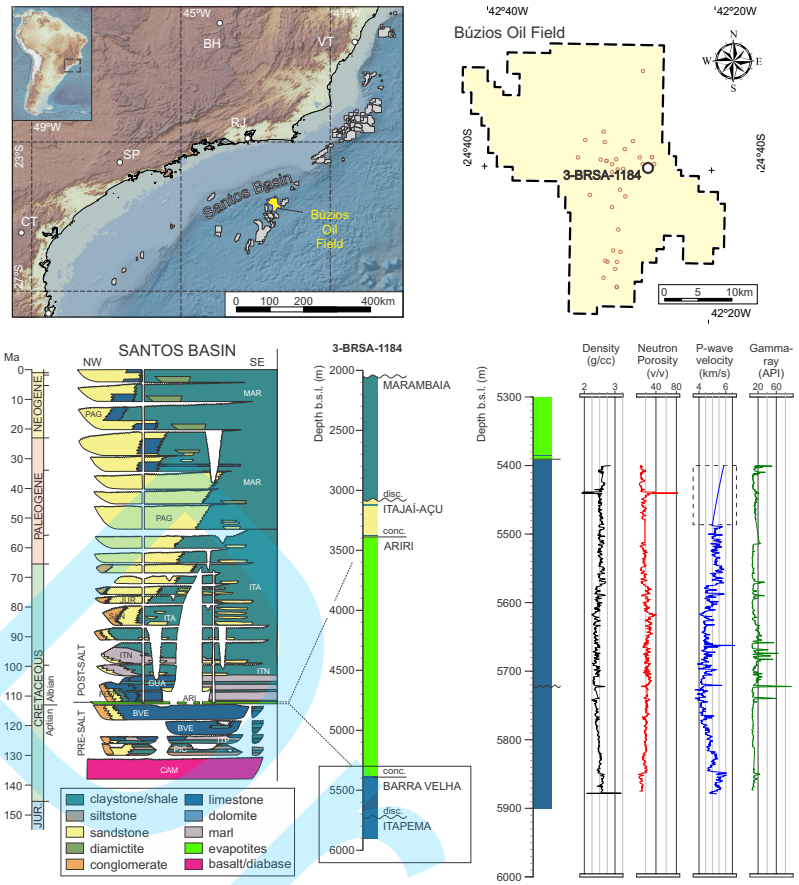

Figure 1: Map of the Santos Basin, focus on the Búzios field. On the lower left section, there is an example of the litostratigraphy sequence (Moreira et al., 2007). On the lower right, we present just an example of a well log, that was not included in our dataset, which lacks the sonic log measurements for a specific range of depth.

sequence is represented by the Aptian non-marine and shallow-water sequence (i.e. lacustrine carbonates and shales) of the Barra Velha Formation (Gomes et al., 2020), followed by evaporitic deposit (i.e. anhydride and halite) of the Ariri Formation (Moreira et al., 2007), featuring the typical sequence of a continental-to-marine transitional environment also registered in other adjacent and correlated basins along Brazilian southeast coast (e.g.,Campos Basin;(Winter et al., 2007)).

METHODS

All data processing in this study was done using Python and Dug Insight (Dug, 2021), which included loading and graphing logs, selecting wells to utilize, simulating logs, and creating graphs with results. Statistical analysis were performed on the simulation results. The seismic and well log data were provided by ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis).

Data

For this work, we selected 20 well-log data from BOF. The well-log curves used in this work were the density log (ρ_b), neutron porosity (NPHI), gamma-ray (GR),resistivity log, and P-wave velocity (V_p) (Figure 2). For an initial test, the ρ_b, NPHI, resistivity and GR curves were used as input for training the ML algorithms,

while the V_p was used as the label. During the procedure, 18 wells were used for training and cross-validation, while 2 were used for additional tests and generating the results. Figure 3 shows the seismic section with the location of well 9-BRSA-1197-RJS, one of the wells used for additional tests and results. This well lacked the sonic log info between depths of 5400-5440 and 5660-5700 meters and it was selected to apply the ML models and compare them with Gardner's equation in the discussion section.

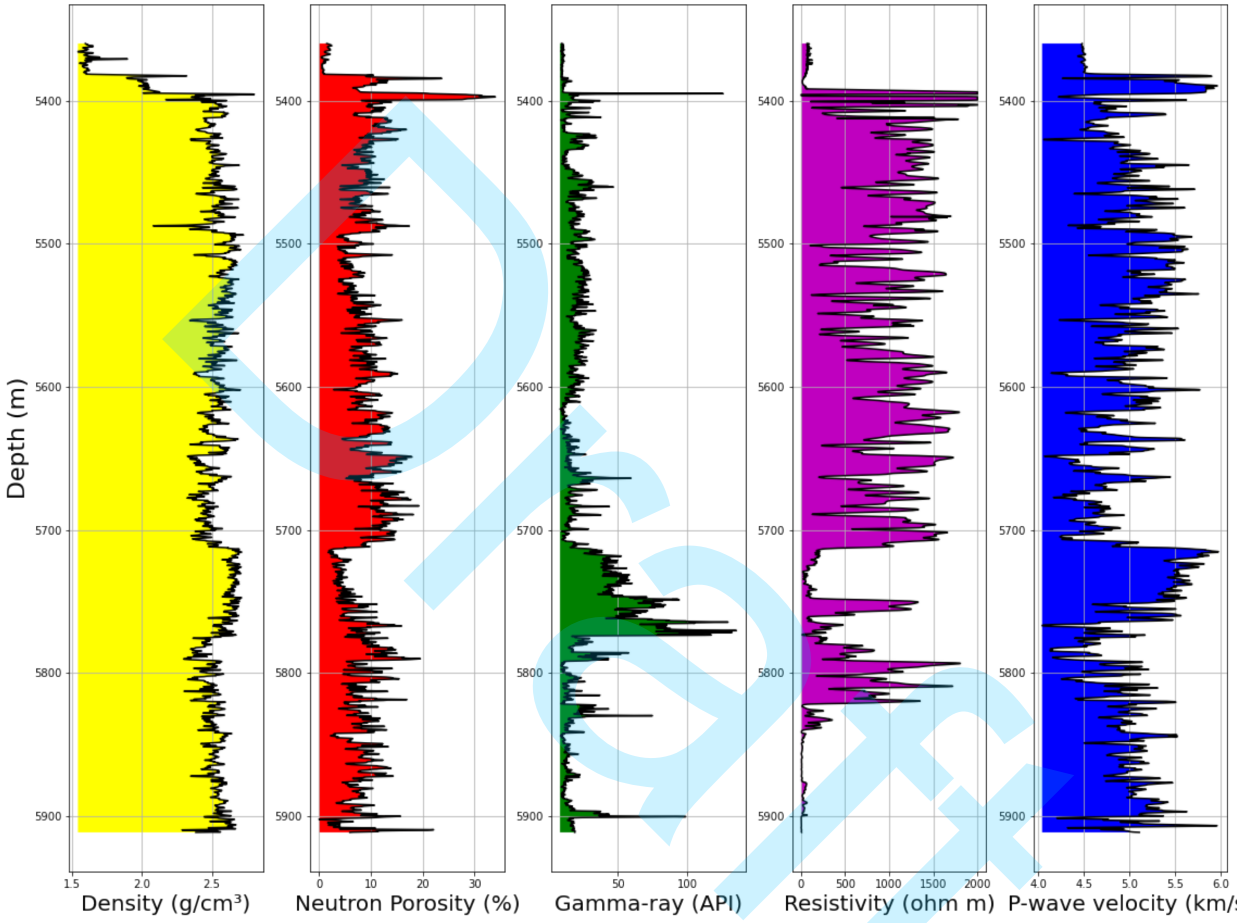

Figure 2: Well log curves used in this work. Density, neutron porosity, gamma-ray and resistivity were used as input, and P-wave velocity was used as the correct output.

The ML algorithms used were the MLP Regressor, the Random Forest Regressor and the Gradient Boosting Regressor, all implemented with the Scikit-Learn library in Python (Pedregosa et al., 2011). The Gardner equation was implemented with the parameter values from Gardner et al. (1974), we isolated the V_p variable on the equation and estimated it by having the density curve as input. Two metrics were selected for the evaluation of the different algorithms. One was the Root mean squared error (RMSE) which penalizes large errors (Chai and Draxler, 2014), and the other was Pearson's correlation coefficient which evaluates a linear relationship between measured and estimated values (Sedgwick, 2012; Thirumalai et al., 2017).

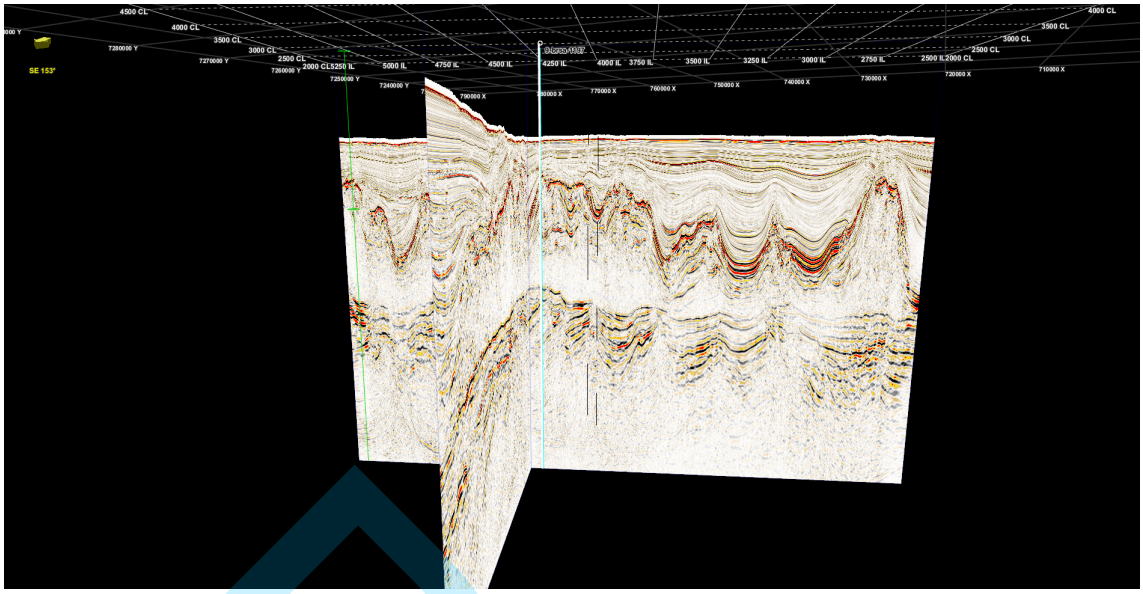

Figure 3: 3D seismic section where the well 9-BRSA-1197-RJS is located. The pre-salt layer for the well location ranges from around 5400 to 5700 meters.

Gardner's Equation

The empirical equation estimated by Gardner et al. (1974) shows that the relationship between density (ρ_{b}) and velocity (V_P) is:

$$\rho_{\mathrm{b}} = k\,[V_P]^B\,, \tag{1}$$

where k and B are the empirical constants, and their numerical values change accordingly to the units of measurement for density (ρ_{b}) and velocity (V_P), the units that we used in this work were g/cm^3 for density and km/s for velocity, thus, the values for the empirical constants were $k = 1.74$ and $B = 0.25$, these are the values found by Gardner et al. (1974).

Gardner's equation is a systematic relationship between the velocity and density of many sedimentary rocks in situ. The empirical relationship allows estimating the reflection coefficients from the velocity information. Gardner et al. (1974) also concluded that Gassmann's theory is valid for sedimentary rocks at interrelated elastic constants, densities, and P-wave velocities for different rock components and for the entire consolidated rock, with the structure or skeleton being an important component. Microcracks can be present in rock and slow down the P-wave velocity. Nevertheless, lithostatic and tectonic stresses can close them and induce the velocity increase. The elastic parameters of rocks without microcracks can be estimated using the theories of Voigt and Reuss and the elastic constants of crystals (Swan and Kosaka, 1997).

Neural Network

Neural Networks (or Artificial Neural Networks) are an important part of Artificial Intelligence and were developed by authors such as McCulloch and Pitts (1943) and Rosenblatt (1958) as a mathematical model inspired by the information processing that occurs in the biological neurons in the brain. The neural network models we selected for this work are supervised learning methods, which means that they require pairs of input and labeled output data values for training, validating, and testing. The mathematical model of a neural network is

based mainly on a matrix multiplication operation.

According to Amr (2020), the Multilayer Perceptron (MLP) is a subset of feedforward neural networks and one of its most commonly used types. In this model, each of the input features is multiplied by a weight and summed to get the output, sometimes an extra bias factor is also added, to get a non-linear output, an activation function is used after the summation. A more detailed mathematical expression for a single layer NN is given by:

$$Y_{\mathrm{k}} = f(W_{\mathrm{km}} X_{\mathrm{m}} + B_{\mathrm{k}}), \tag{2}$$

where W_{km} is the matrix containing the weights of the neural network that will be multiplied with the input vector: X_{m}, and then summed with the bias vector: B_{k}, the result of this operation is the input for the activation function f, and gives the calculated output vector Y_{k}.

A different notation for the neural networks model is the equation 3 which calculates each element y_{i} of vector Y_{k} individually:

$$y_{\mathrm{i}} = f\left(\sum_{j=1}^{m} w_{i,j} x_j + b_i\right), i = 1, 2, 3...k. \tag{3}$$

The same operation expressed in a matrix multiplication will be:

$$Y_{\mathrm{k}} = f\left(\begin{bmatrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k1} & w_{k2} & \cdots & w_{km} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}\right). \tag{4}$$

We used the MLP Regressor function from the Scikit-Learn library. We used the GridSearchCV function for the tuning of hyperparameters, the test was conducted for the selection between the solvers: 'sgd', which is the stochastic gradient descent method (Amari, 1993), and 'adam', which is a modification from the stochastic gradient descent method (Kingma and Ba, 2014). Other hyperparameters considered were constant or adaptive learning rates. The activation functions were also tested among the logistic, identity, hyperbolic tangent and rectified linear. The selected score metric was the negative mean squared error.

Ensemble methods based on trees

The other algorithms applied in this work were the Random Forest and the Gradient Boosting Regressor, which are ensemble methods. A very powerful and applied technique, an ensemble of methods is a learning algorithm that combines the predictions of multiple statistical models to improve the final prediction, an ensemble method can be applied both for classification and regression problems (Dietterich, 2000; Breiman, 2001; Iaccarino et al., 2024).

To understand the concept of the Random Forests algorithm, it is important to first understand the definition of a decision tree classifier. A decision tree is constructed by analysing a set of training samples with known

class labels and make a series of questions about features related to these samples. Each question is contained in a node and every internal node points to one child node for each possible answer to the question forming a hierarchy of questions (Kingsford and Salzberg, 2008).

A Random Forest algorithm combines the predictions of decision trees such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Random Forest can be applied for regression by growing trees depending on a random vector such that the tree predictor takes on numerical values as opposed to class labels (Breiman, 2001).

The Gradient Boosting Regressor or GBR is an algorithm that combines the intuitions from boosting and gradient descent to construct ensembles of decision trees. In this algorithm, the gradient of a cost function is calculated with respect to the predicted values of the ensemble and new decision trees are added iteratively to the structure to shift the algorithm in the negative direction of the gradient, other important parameters are the maximum depth of the trees and the learning rate of the gradient descent (Friedman, 2001; Iaccarino et al., 2024).

The hyperparameters for the Random Forest and the Gradient Boosting Regressors utilized were the default ones from the Scikit-Learn library, with the exception of the maximum depth for the Random Forest Regressor. For the Gradient Boosting Regressor, the loss function was the squared error of the regression, the learning rate of the gradient descent is equal to 0.1 and the maximum depth of the trees is 3. For the Random Forest Regressor, the number of trees is 100, the criterion function is the squared error and the maximum depth of the tree is 2.

Petroelastic inversion

As an example of how our methodology may be applied, a 3D model of the natural logarithm of the acoustic impedance will be created for the ML model that presents the best metrics result. We will utilize this 3D model as a priori guess for a petroelastic inversion of a 3D poststack seismic data.

We will use the PyLops library (Ravasi and Vasconcelos, 2020) to calculate the petroelastic inversion. The importance of this algorithm is to provide a variable which contains a petroelastic property of the medium, which might be applied to calculate petrophysical properties such as porosity, oil and water saturation. This algorithm requires as input the information of the wavelet, the seismic trace and, as optional parameter, the priori model for the natural logarithm of the acoustic impedance, this last optional parameter will significantly improve the inversion result. Figure 4 shows an example of a wavelet signal which is a required parameter for the inversion algorithm.

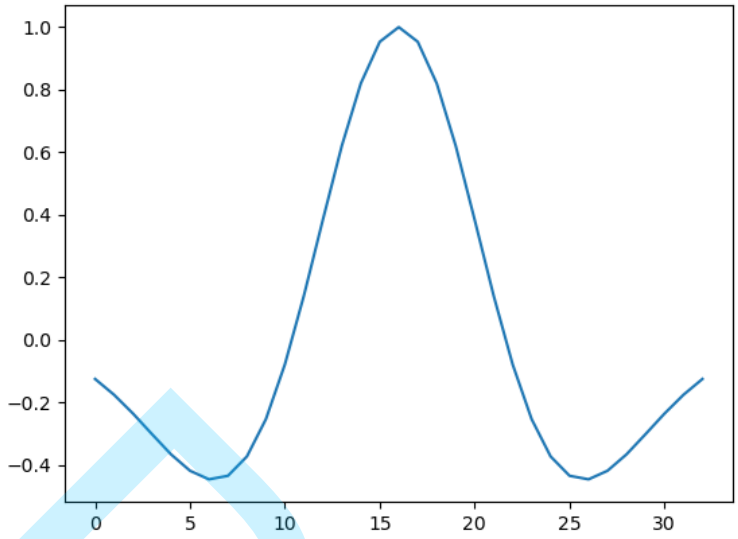

Figure 4: Signal example of a Ricker wavelet. One of the necessary input parameters for the petroelastic inversion.

The first step to create the 3D priori model is to use the estimated P-wave velocity (green curve shown in Figure 12) and the bulk density log (black curve from Figure 12) to calculate the natural logarithm of the acoustic impedance for each well. The acoustic impedance is calculated by the multiplication of the density log with the velocity log, then the natural logarithm is applied.

The second step is to bring the calculated natural logarithm of the acoustic impedance of each well to the seismic scale, since the samples on the well occur about every 15 centimeters and on the seismic data they are sampled about every 5 meters, to adjust the scales, we used the Numpy interp Python function. The last step is to make a linear regression for each well location between the seismic data and the natural logarithm of the acoustic impedance which will allow us to create the 3D priori model using the 3D seismic data as input.

RESULTS

The hyperparameter tuning applied for the MLP Regressor selected 'adam' as the best optimizer and a constant learning rate. The other hyperparameters consisted of one hidden layer containing 100 neurons and a second layer with the number of neurons consistent with the labeled data output size, the activation function hyperbolic tangent (tanh) achieved the best result among the functions which were tested.

The results we obtained for comparing Gardner's equation and model 2, 6 and 10 from Table 1 are shown in Figure 5 and the results for the metrics are shown in Table 2. In Figure 5 it is possible to observe the fit for the P-wave velocity provided by the different models for the pre-salt layer of a well that ranges from 5450 m to 6100 m. Model 10 achieved the lowest RMSE calculated between measured and estimated velocities which was 0.2523 and the highest correlation coefficient which was 0.8573 as shown in Table 2.

The estimation for the Gardner equation is shown in Figure 5. The RMSE calculated among the velocities for Gardner's equation was 0.7061, much higher than the other ML models, while the correlation coefficient was 0.8051 as shown in Table 2.

Table 1: Models used to train the data and the parameters used as input.

Model	Algorithm	Input parameters
1	MLP Regressor	GR, NPHI, Density and Resistivity
2	MLP Regressor	GR, NPHI and Density
3	MLP Regressor	GR and NPHI
4	MLP Regressor	NPHI and Density
5	Random Forest Regressor	GR, NPHI, Density and Resistivity
6	Random Forest Regressor	GR, NPHI and Density
7	Random Forest Regressor	GR and NPHI
8	Random Forest Regressor	NPHI and Density
9	Gradient Boosting Regressor	GR, NPHI, Density and Resistivity
10	Gradient Boosting Regressor	GR, NPHI and Density
11	Gradient Boosting Regressor	GR and NPHI
12	Gradient Boosting Regressor	NPHI and Density

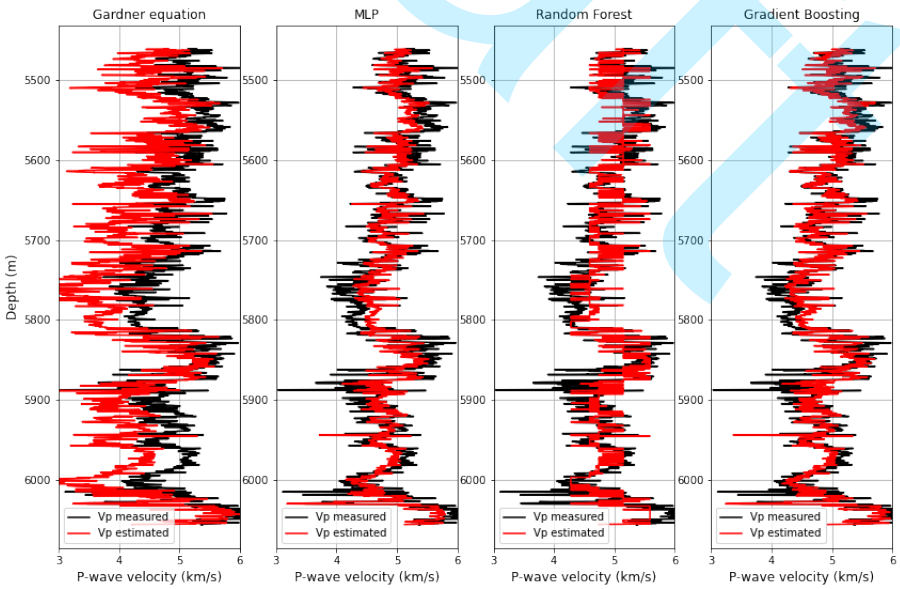

Figure 5: Result for the P-wave velocity estimation using Gardner, MLP, Random Forest and Gradient Boosting regressor, each regression used as input the GR, NPHI and RHOB logs, these estimations correspond to models 2, 6 and 10. The black curve represents the real velocity values, while the red curve represents the calculated values.

Table 2: Results of the metrics used for comparison of the P-wave velocity estimated by Gardner's equation, MLP Regressor, Random Forest Regressor, Gradient Boosting Regressor for each different set of input parameters

Model	RMSE	PEARSON	RMSE (Train)	PEARSON (Train)	RMSE (Test)	PEARSON (Test)
Gardner Equation	0.7061	0.8051	-	-	-	-
1	0.3246	0.7399	0.3922	0.7186	0.3975	0.7155
2	0.2763	0.8445	0.2996	0.8284	0.3091	0.8229
3	0.2984	0.8197	0.3383	0.7706	0.3416	0.7771
4	0.2815	0.8239	0.3244	0.7968	0.3297	0.7938
5	0.3040	0.7783	0.3488	0.7574	0.3626	0.7353
6	0.3050	0.7761	0.3483	0.7591	0.3495	0.7524
7	0.3355	0.7337	0.3754	0.7107	0.3754	0.7161
8	0.3038	0.7784	0.3482	0.7587	0.3597	0.7497
9	0.2682	0.8317	0.2578	0.8756	0.2665	0.8670
10	0.2523	0.8573	0.2736	0.8587	0.2791	0.8570
11	0.2999	0.8089	0.3296	0.7848	0.3366	0.7855
12	0.2728	0.8281	0.2940	0.8347	0.3060	0.8162

We applied the MLP estimation to fill the P-wave velocity for the areas in which this information was missing. Figures 11 and 12 illustrate how the P-wave velocity might be combined with the other logs (GR, NPHI and density (ρ_b)) for interpretation purposes. Figure 6 shows the absolute error for Gardner's equation and models 2, 6 and 10 from Table 1.

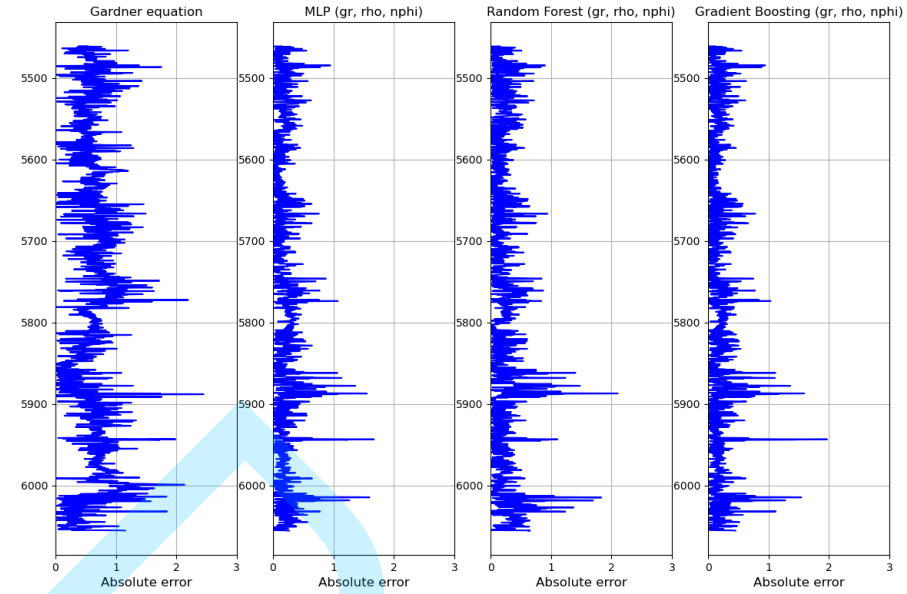

Figure 6: Absolute errors for Gardner, MLP, Random Forest and Gradient Boosting regressor, each regression used as input the GR, NPHI and RHOB logs, these estimations correspond to models 2, 6 and 10.

Figure 7 shows all the results for Gardner and the ML models 1, 5 and 9 which use the GR, NPHI, density and resistivity logs as input. Figure 8 shows the absolute error for each of these methods. The comparison of model 4, 8 and 12 with Gardner is shown in Figures 9 and 10.

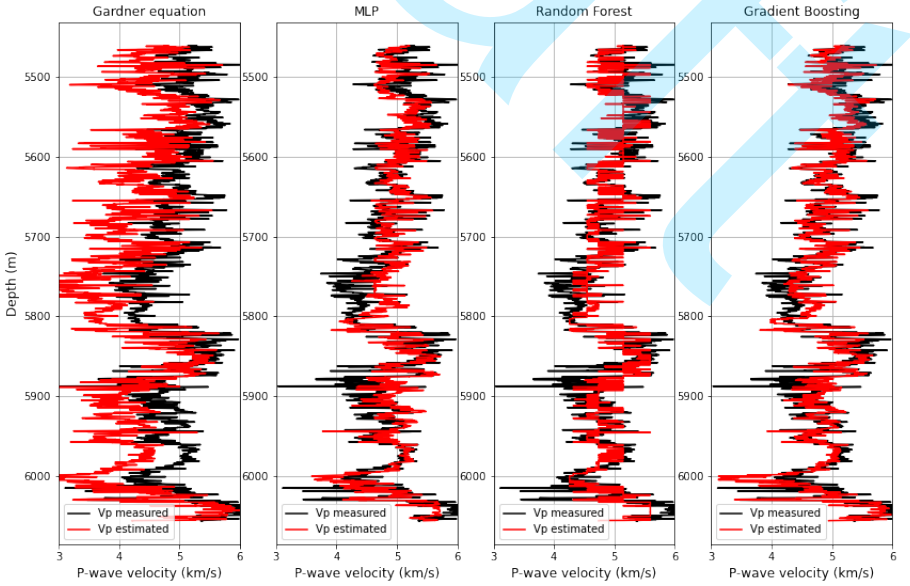

Figure 7: Result for the P-wave velocity estimation using Gardner, MLP, Random Forest and Gradient Boosting regressor, each regression used as input the GR, NPHI, RHOB and resistivity logs, these estimations correspond to models 1, 5 and 9. The black curve represents the real velocity values, while the red curve represents the calculated values.

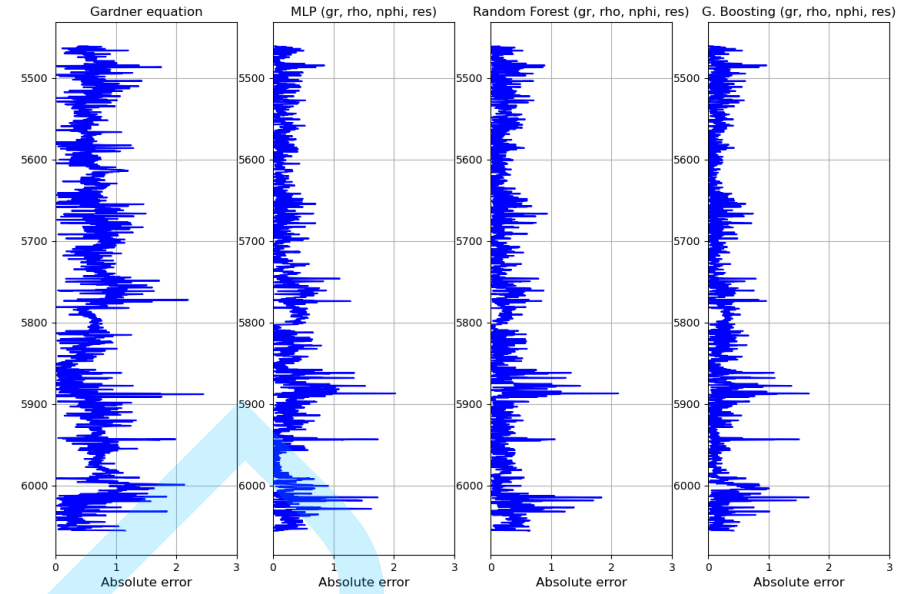

Figure 8: Absolute errors for Gardner, MLP, Random Forest and Gradient Boosting regressor, each regression used as input the GR, NPHI, RHOB and resistivity logs, these estimations correspond to models 1, 5 and 9.

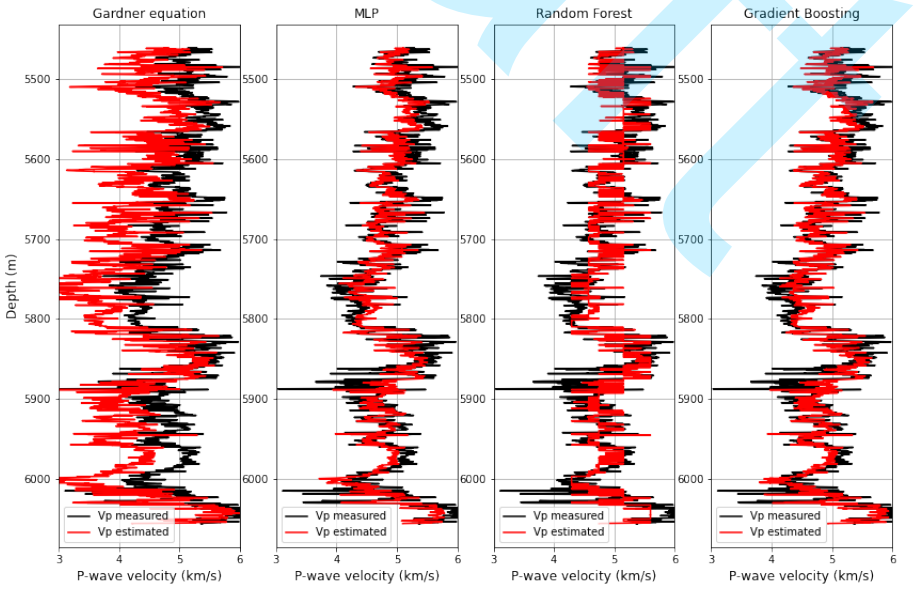

Figure 9: Result for the P-wave velocity estimation using Gardner, MLP, Random Forest and Gradient Boosting regressor, each regression used as input the NPHI and RHOB logs, these estimations correspond to models 4, 8 and 12. The black curve represents the real velocity values, while the red curve represents the calculated values.

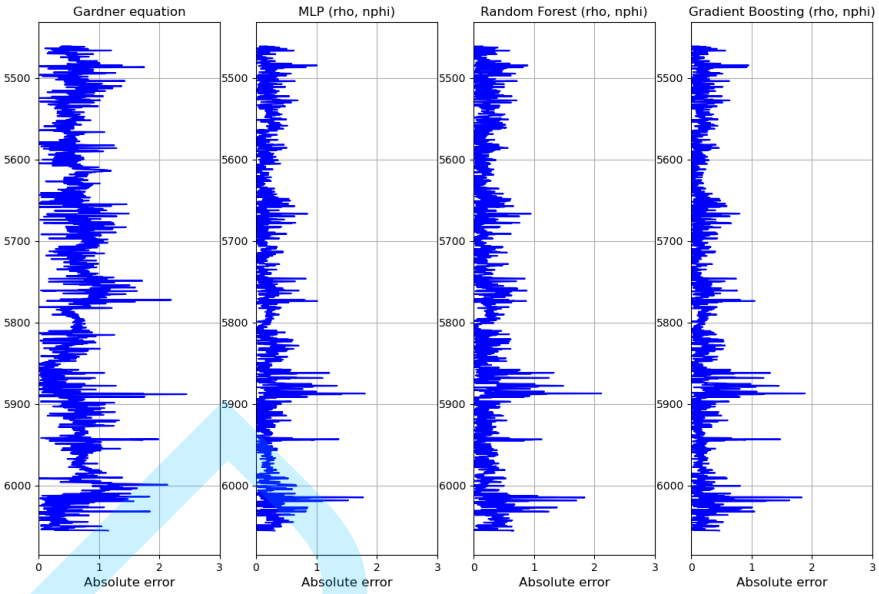

Figure 10: Absolute errors for Gardner, MLP, Random Forest and Gradient Boosting regressor, each regression used as input the NPHI and RHOB logs, these estimations correspond to models 4, 8 and 12.

DISCUSSION

Well logs from a carbonate reservoir in Brazilian pre-salt were used in this investigation. The focus was creating a model to estimate the P-wave velocity log for regions of the well where the sonic log (DT) information was missing. Different combinations of logs that present a good relation with the P-wave velocity log such as Gamma-ray, neutron porosity, density and resistivity were used as input. By testing these different combinations, the objective was to verify the influence of each log on the velocity prediction and if any one of them could be suppressed from the input data set.

Analysing Figures 5 and 6 we can see that Gardner's equation did not provide a good fit and presented the highest absolute error except for depths between 5800 m and 5900 m, Gardner's equation also presented the highest root mean squared error by the metrics results from Table 2. Gardner's equation presented a better correlation coefficient when compared to models 1, 5, 6, 7 and 8; however, a qualitative analysis of the plotted curves from Figure 5 and the errors results demonstrate that these ML models are still a better choice compared to Gardner's equation.

The analysis of the metrics results from Table 2 also show that the lowest RMSE and highest correlation coefficient was from model 10, Figures 5 and 6 illustrate how this model presents a low absolute error except for a region around the depth of 5800 m, this model is also very effective in removing some outliers that are present close to the depth of 5900 m. The metrics results from models 1, 5, and 9 in Table 2 and the results from Figures 7 and 8 show how the addition of the resistivity log does not provide a better estimation for the P-wave velocity as it increases the RMSE and decreases the correlation coefficient when compared to the models where the GR, neutron porosity and density logs are used as input.

The metrics results for models 3, 4, 7, 8, 11 and 12 in Table 2 show how the suppression of the Gamma

ray (GR) information as an input variable does not cause significant increase in the RMSE, furthermore, by comparing models 6 and 8, the results show that the Random Forest Regressor achieved a better performance when the GR variable was suppressed. The suppression of the density log as an input variable causes more error in the P-wave velocity estimation, this can be observed by comparing the metrics results for models 3, 7 and 11 with the metrics for models 4, 8 and 12 in Table 2.

Figures 11 and 12 illustrate how the NN and the Gradient Boosting Regressor methods, specifically, models 2 and 10 were efficient in filling the gaps of sonic log information for well 9-BRSA-1197-RJS shown in Figure 3 which lacked this information for specific depths. The ML methods were also very accurate on estimating the curve in regions where the P-wave velocity information was present (blue curve in Figures 11 and 12), however, they did not provide a good fit between the depths of 5550 m and 5600 m, this can be explained by the presence of a carbonate dark grey shale in this specific depth as shown in the lithological profile in Figures 11 and 12.

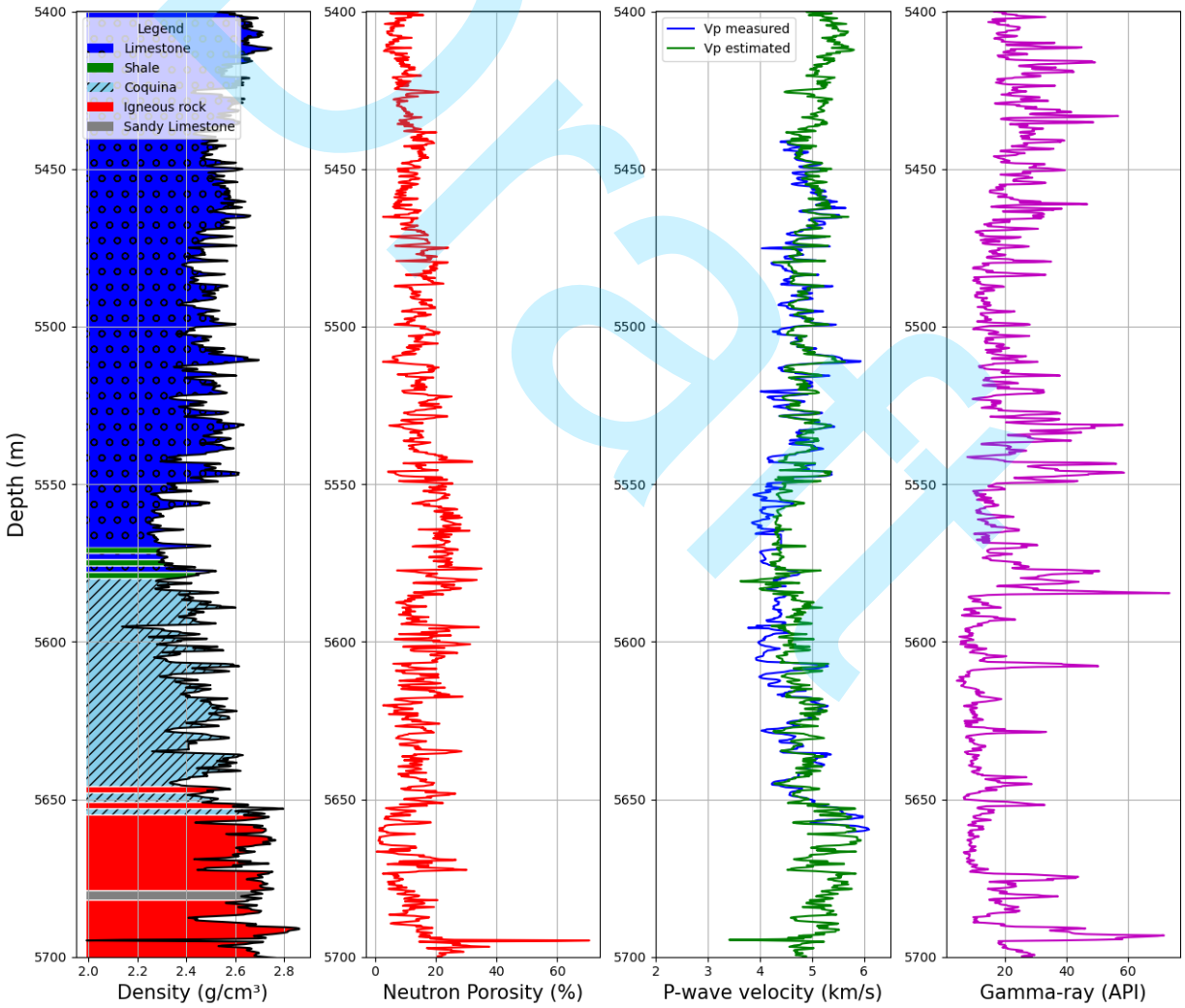

Figure 11: Lithology and well log curves of density, neutron porosity, gamma ray and the result for the P-wave velocity estimation using the MLP Regressor (model 2). The blue curve represents the real velocity values while the green curve represents the calculated values. It was possible to fill the P-wave velocity information where it was absent.

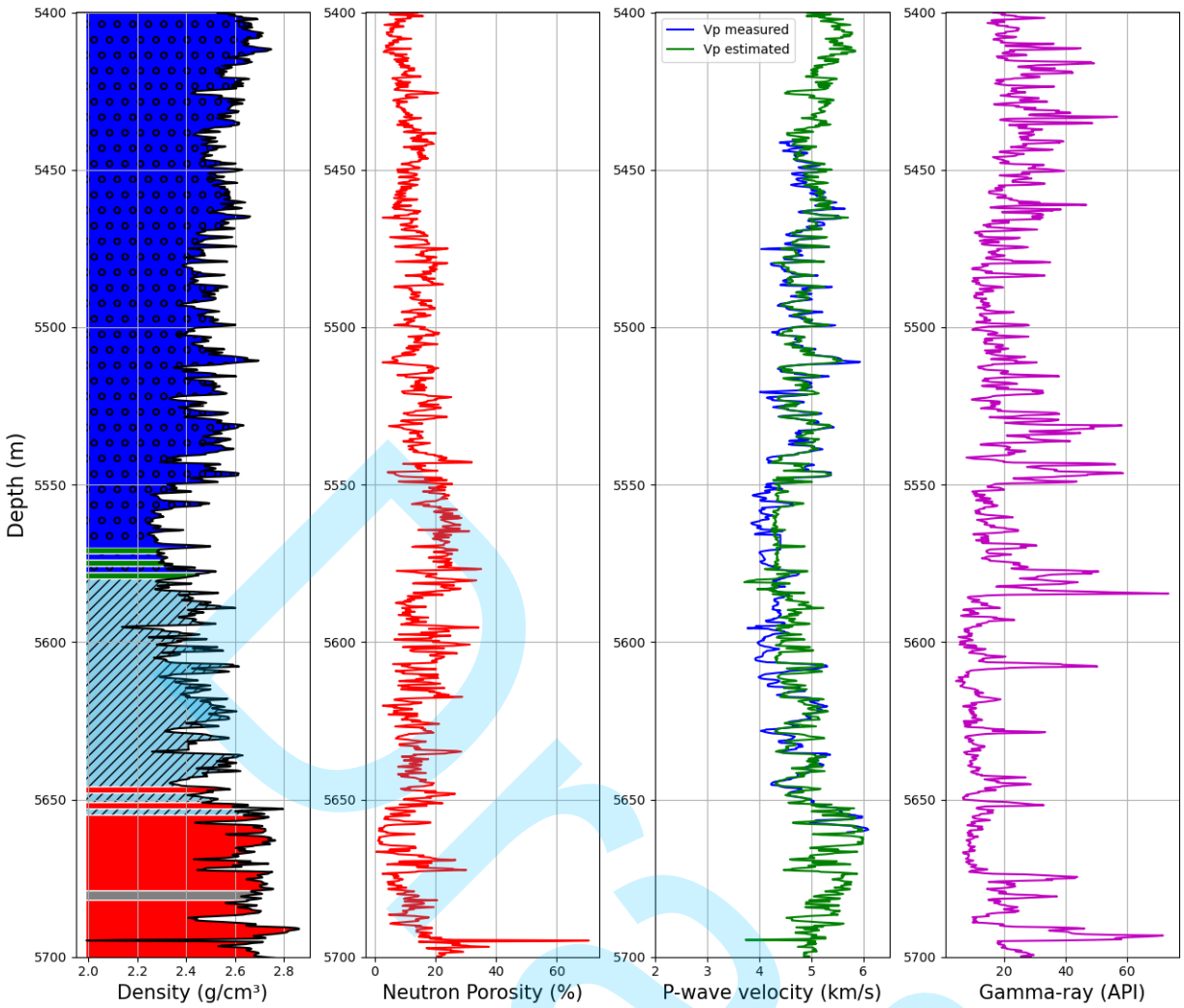

Figure 12: Lithology and well log curves of density, neutron porosity, gamma ray and the result for the P-wave velocity estimation using the Gradient Boosting Regressor (model 10). The blue curve represents the real velocity values while the green curve represents the calculated values. It was possible to fill in the P-wave velocity information where it was absent.

By comparing the results from Figures 11, 12 and 13 we can observe how Gardner's equation fails to reproduce the measured values of P-wave velocity. Among the ML methods chosen for the estimation of the P-wave velocity, the Gradient Boosting Regressor proved to be the most accurate due to its lowest mean squared error and highest correlation with the real values as shown in Table 2 but also proved to be more accurate for depths between 5600 and 5700 m in the 'blind test' well log shown in Figure 12 when compared to the MLP Regressor shown in Figure 11.

Gardner's equation was applied to estimate the P-wave velocity on the same well where the MLP and Gradient Boosting Regressor were applied (Figures 11, 12 and 13). We can observe that the ML methods estimate the measured velocity values more accurately. By comparing the blue and green curves between 5450 and 5650 meters in Figure 13, it is possible to observe that the estimated values of Gardner's equation do not reproduce the measured ones.

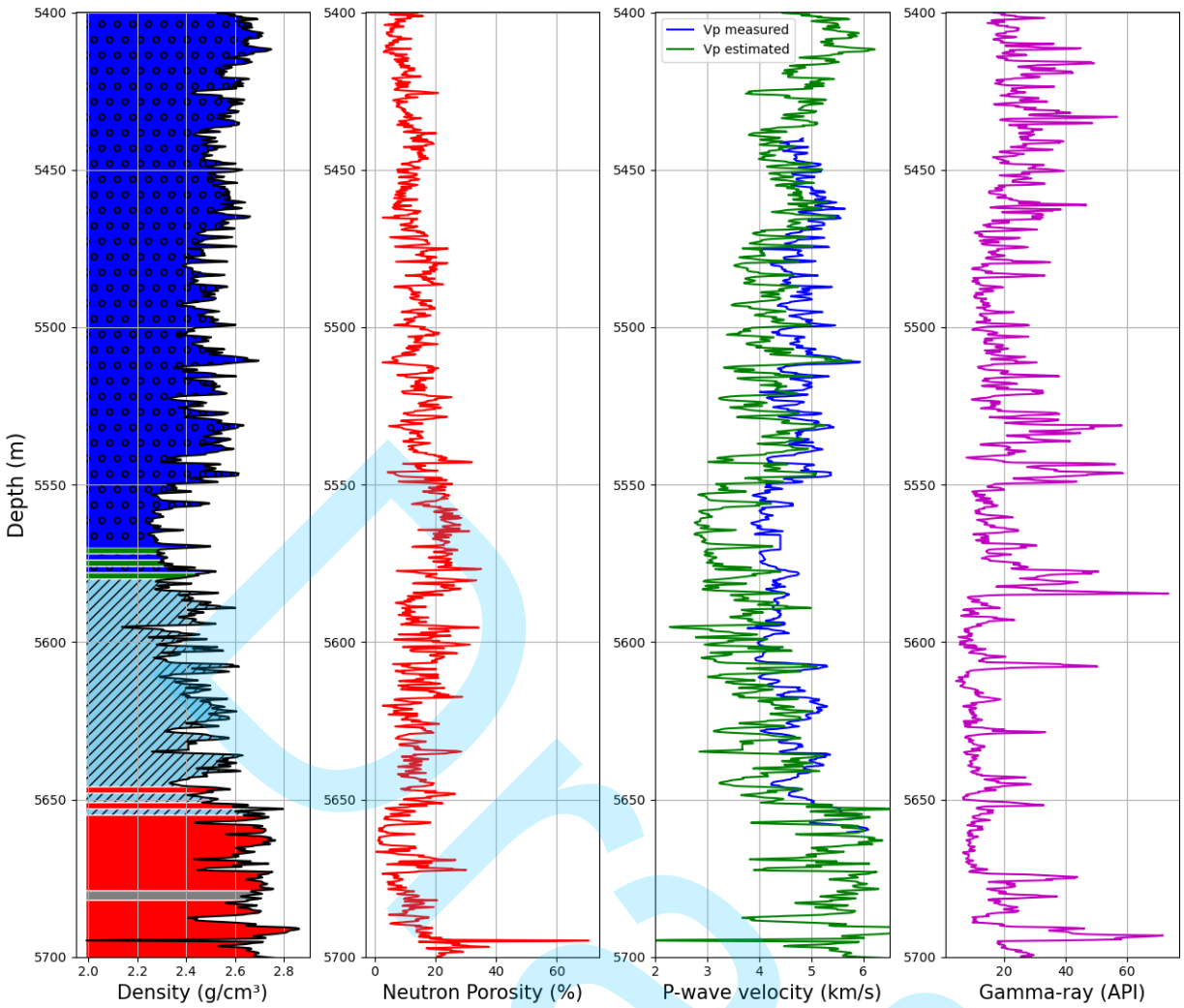

Figure 13: Lithology and well log curves of density, neutron porosity, gamma ray and the result for the P-wave velocity estimation using Gardner's equation. The blue curve represents the real velocity values while the green curve represents the calculated values.

Hyperparameter tuning

The hyperparameter tuning was a crucial step for the selection of the hyperparameters that best suited the data, due to its high computational cost, the tested hyperparameters were reduced to two options between solvers and learning rates and four options of activation functions. Nevertheless, the Gradient Boosting Regressor presented the lowest mean squared error and the highest correlation coefficient with its default hyperparameter values.

Petroelastic inversion

Considering that model 10 presented the best metrics result, we used the calculated P-wave velocity (green curve shown in Figure 12) and the bulk density log (black curve from Figure 12) to calculate the natural logarithm of the acoustic impedance for each well.

After filtering the well log data to the seismic scale, we did the linear regression to create the 3D priori model using the 3D seismic data as input. We used the PyLops library (Ravasi and Vasconcelos, 2020) and

our priori model to calculate the inversion for a poststack seismic data, Figure 14 shows the result for a seismic data cube from the BOF in Santos Basin.

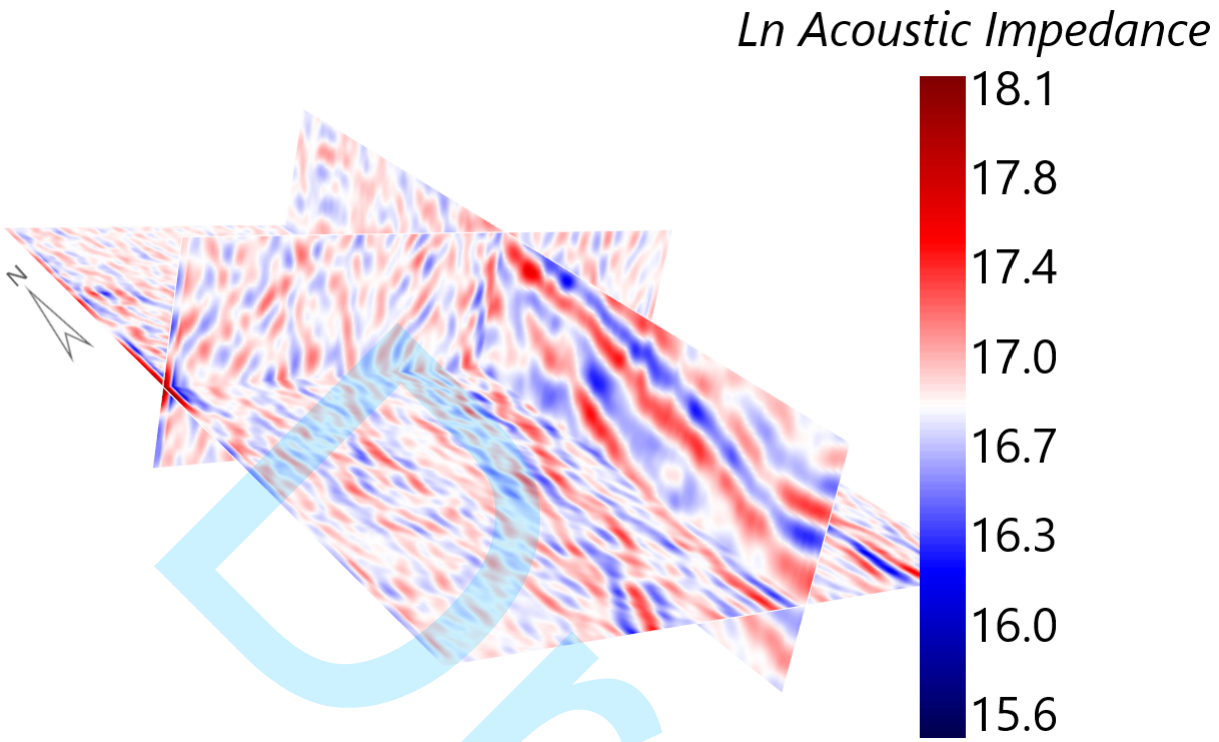

Figure 14: Natural logarithm of the acoustic impedance calculated through a petroelastic inversion from a seismic data. The P-wave velocity data estimated using the ML methods from this work was used to create the priori model for the inversion.

CONCLUSION

The measurement and the estimation of petrophysical and petroelastic parameters such as porosity, permeability and P-wave velocity in carbonate deposits is a complex process, which corroborates the application of the methodology of this work. From the metric relations, we understand that the ML models were more successful in estimating the P-wave velocity when compared to Gardner's equation. This leads us to conclude that ML is a good option for V_p simulation in complex models.

By the analysis of the metrics shown in Table 2 and the curves of Figures 5 and 13 we can conclude that the Gardner's equation would not be as precise as the ML methods for regions of the well where the P-wave velocity is missing, especially for areas where carbonate shales are present.

The case study discussed in this research produced remarkable results with very low errors for the ML methods, especially considering that it is a carbonate reservoir with complex geology. In order to replicate the ML models created in this work, it is only necessary to have access to the same well log data from BOF and to the Python libraries. The number of well logs used in this work were enough to produce accurate results. In future works, it is intended to add other logs to the data set that may be relevant in influencing the P-wave velocity (V_p) estimate and also check how the natural logarithm of the acoustic impedance calculated by the petroelastic inversion can be precise to estimate petrophysical properties such as the porosity of a carbonate

reservoir.

ACKNOWLEDGMENTS

The authors acknowledge Universidade Federal de Minas Gerais - UFMG for the computational support, Dug Insight for the software licences, Petrobras for financing the RD&I project, Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) for the dataset and Elita Selmara de Abreu, Humberto Luis Siqueira Reis, Heitor Soares Ramos Filho, Erick Talarico and Matheus Sobreira for the text suggestions.

DATA AND MATERIALS AVAILABILITY

The data used in the manuscript is open and publicly available.

AUTHOR CONTRIBUTIONS

Caique Pinheiro de Carvalho: Idea conceptualization, text writing and editing, data organization, methodological development, neural network model training and testing. **Jéssica Lia Santos da Costa**: Idea conceptualization, text writing and editing, data organization, methodological development. **Tobias Fonte-Boa**: Text writing and editing, data organization, geological context writing, figure editing. **Tiago Amâncio Novo**: Idea conceptualization, geological context suggestions, data request. **Maria José Campos de Oliveira**: Geological context writing, figure editing, text writing and editing. **Fernanda Moura Costa**: Data request and gathering, text writing and editing, geological context writing.

REFERENCES

Aleardi, M., 2015, Seismic velocity estimation from well log data with genetic algorithms in comparison to neural networks and multilinear approaches: Journal of Applied Geophysics, **117**, 13–22, doi: https://doi.org/10.1016/j.jappgeo.2015.03.021.

Amari, S.-i., 1993, Backpropagation and stochastic gradient descent method: Neurocomputing, **5**, 185–196, doi: https://doi.org/10.1016/0925-2312(93)90006-O.

Amr, T., 2020, Hands-on machine learning with scikit-learn and scientific Python toolkits: A practical guide to implementing supervised and unsupervised machine learning algorithms in Python: Packt Publishing Ltd. 384 pp.

Breiman, L., 2001, Random forests: Machine Learning, **45**, 5–32, doi: https://doi.org/10.1023/A:1010933404324.

Brune, S., 2016, Rifts and rifted margins: A review of geodynamic processes and natural hazards: Plate Boundaries and Natural Hazards, **219**, 13, doi: https://doi.org/10.1002/9781119054146.ch2.

Carrasquilla, M. D., C. P. Carvalho, M. D. Costa, I. J. Souza, J. J. de Figueiredo, C. B. da Silva, C. R. Lima, R. S. Silveira, C. E. Manajás, and L. Rautino, 2022, Geological, geophysical and mathematical analysis of synthetic bulk density logs around the world - Part I - the use of linear regression on empirical parameters estimation: Journal of Applied Geophysics, **204**, 104733, doi: https://doi.org/10.1016/j.jappgeo.2022.104733.

Carvalho, C. P., C. B. da Silva, J. J. S. de Figueiredo, and M. D. Carrasquilla, 2022, On the application of neural network regression for density log construction: comparisons with traditional empirical models: Brazilian Journal of Geophysics, **40**, 323–335, doi: http://dx.doi.org/10.22564/brjg.v40i2.2166.

Chai, T., and R. R. Draxler, 2014, Root mean square error (RMSE) or mean absolute error (MAE) – arguments against avoiding rmse in the literature: Geoscientific Model Development, **7, 3**, 1247–1250, doi: 10.5194/gmd-7-1247-2014.

de Macedo, I. A., J. J. S. de Figueiredo, and M. C. De Sousa, 2020, Density log correction for borehole effects and its impact on well-to-seismic tie: Application on a North Sea data set: Interpretation, **8**, T43–T53, doi: https://doi.org/10.1190/INT-2019-0004.1.

de Mio, E., and H. Chang, 2005, Integração de métodos geofísicos na modelagem crustal da Bacia de Santos: Revista Brasileira de Geofísica, **23**, 275–284, doi: 10.1590/S0102-261X2005000300006.

Dietterich, T. G., 2000, Ensemble methods in machine learning: Presented at the International workshop on Multiple Classifier Systems, MCS 2000. Lecture Notes in Computer Science, vol 1857, Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/3-540-45014-9_1.

Dug, 2021, Dug insight: Dug ltd. available on: <https://dug.com/insight>.

Friedman, J. H., 2001, Greedy function approximation: a gradient boosting machine: Annals of Statistics, 1189–1232, doi: https://www.jstor.org/stable/2699986.

Gardner, G., L. Gardner, and A. Gregory, 1974, Formation velocity and density—the diagnostic basics for stratigraphic traps: Geophysics, **39**, 770–780, doi: https://doi.org/10.1190/1.1440465.

Gomes, J., R. Bunevich, L. Tedeschi, M. Tucker, and F. Whitaker, 2020, Facies classification and patterns of lacustrine carbonate deposition of the Barra Velha Formation, Santos Basin, Brazilian Pre-salt: Marine and Petroleum Geology, **113**, 104176, doi: https://doi.org/10.1016/j.marpetgeo.2019.104176.

Iaccarino, A. G., A. Cristofaro, M. Picozzi, D. Spallarossa, and D. Scafidi, 2024, Real-time prediction of distance and PGA from P-wave features using gradient boosting regressor for on-site earthquake early warning applications: Geophysical Journal International, **236**, 675–687, doi: https://doi.org/10.1093/gji/ggad443.

Kingma, D. P., and J. Ba, 2014, Adam: A method for stochastic optimization: arXiv preprint arXiv:1412.6980, doi: https://doi.org/10.48550/arXiv.1412.6980.

Kingsford, C., and S. L. Salzberg, 2008, What are decision trees?: Nature Biotechnology, **26**, 1011–1013, doi: https://doi.org/10.1038/nbt0908-1011.

Lim, J.-S., and J. Kim, 2004, Reservoir porosity and permeability estimation from well logs using fuzzy logic and neural networks: Presented at the SPE Asia Pacific Oil and Gas Conference and Exhibition, OnePetro. Perth, Australia. doi: https://doi.org/10.2118/88476-MS.

McCulloch, W. S., and W. Pitts, 1943, A logical calculus of the ideas immanent in nervous activity: The Bulletin of Mathematical Biophysics, **5**, 115–133, doi: https://doi.org/10.1007/BF02478259.

Moreira, J. L. P., C. V. Madeira, J. A. Gil, and M. A. P. Machado, 2007, Bacia de Santos: Boletim de Geociências da PETROBRAS, **15**, 531–549.

Oloruntobi, O., and S. Butt, 2019, The new formation bulk density predictions for siliciclastic rocks: Journal of Petroleum Science and Engineering, **180**, 526–537, doi: https://doi.org/10.1016/j.petrol.2019.05.017.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R.

Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 2011, Scikit-learn: Machine learning in Python: Journal of Machine Learning Research, **12**, 2825–2830.

Ravasi, M., and I. Vasconcelos, 2020, PyLops—a linear-operator Python library for scalable algebra and optimization: SoftwareX, **11**, 100361, doi: 10.1016/j.softx.2019.100361.

Rolon, L., S. D. Mohaghegh, S. Ameri, R. Gaskari, and B. McDaniel, 2009, Using artificial neural networks to generate synthetic well logs: Journal of Natural Gas Science and Engineering, **1**, 118–133, doi: https://doi.org/10.1016/j.jngse.2009.08.003.

Rosenblatt, F., 1958, The perceptron: a probabilistic model for information storage and organization in the brain.: Psychological Review, **65**, 386, doi: https://doi.org/10.1037/h0042519.

Sahin, E. K., 2020, Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest: SN Applied Sciences, **2**, 1308, doi: https://doi.org/10.1007/s42452-020-3060-1.

Sedgwick, P., 2012, Pearson's correlation coefficient: BMJ, **345**, doi: https://doi.org/10.1136/bmj.e4483.

Swan, C. C., and I. Kosaka, 1997, Voigt-Reuss topology optimization for structures with linear elastic material behaviours: International Journal for Numerical Methods in Engineering, **40**, 3033–3057, doi: https://doi.org/10.1002/(SICI)1097-0207(19970830)40:16<3033::AID-NME196>3.0.CO;2-Z.

Thirumalai, C., S. A. Chandhini, and M. Vaishnavi, 2017, Analysing the concrete compressive strength using Pearson and Spearman: 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), IEEE, 215–218. doi: 10.1109/ICECA.2017.8212799.

Winter, W., R. Jahnert, and A. França, 2007, Bacia de Campos: B. Geoci. Petrobras, **2**, 511–529.

Caique Pinheiro de Carvalho: Idea conceptualization, text writing and editing, data organization, methodological development, neural network model training and testing.
Jéssica Lia Santos da Costa: Idea conceptualization, text writing and editing, data organization, methodological development.
Tobias Fonte-Boa: Text writing and editing, data organization, geological context writing, figure editing.
Tiago Amâncio Novo: Idea conceptualization, geological context suggestions, data request.
Maria José Campos de Oliveira: Geological context writing, figure editing, text writing and editing.
Fernanda Moura Costa: Data request and gathering, text writing and editing, geological context writing.