Enhancing S-Wave Log Construction through Semi-Supervised Regression Clustering

Pamela Carolayne R. Bolsem* (Federal University of Pará, Brazil), João Rafael Barroso S. Silveira (Federal University of Pará, Brazil), José Jadsom S. De Figueiredo (CPGf-UFPA & INCT-GP)& João Lucas M. da Silva (Federal University of Pará, Brazil)

This paper was prepared for presentation during the 18th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, 16-19 October 2023.

Abstract

This work presents a method for estimating shear wave velocity (DTS) log from other well-logs. Using Bayesian Gaussian mixture clustering, significant patterns in DTS data are detected by integrating clustering and regression approaches with machine learning. To reliably estimate DTS values, regression models (Rando Forest-RF, Least-Square Regression-LSR, Multi-Layer Perceptron-MLP) are used. The mean squared error (MSE), mean absolute percent error (MAPE) and R2 evaluation metrics show the RF method's superior effectiveness in explaining data variability. The proposed methodology improves reservoir characterization, and oil exploration, and gives useful information about subsurface rock formations.

Introduction

In recent years, there have been great advances in the research and application of intelligent systems as powerful tools for extracting quantitative formulations between two data sets (inputs/outputs), which have a fundamental dependence on the petroleum industry (Na'imi et al., 2014). Due to their low resolution, seismic data are sometimes the only data used to study reservoir structures. Because of their extensive spatial coverage, researchers have sought to use seismic data and their properties as predictive variables in lithology prediction and reservoir characterization projects. Well logging is essential for the oil and gas industry to understand the petrophysical and geomechanical properties (He et al., 2019).

Accurate knowledge of shear-wave velocity is essential for petrophysical evaluation. Information about the shear-wave velocity along with compressional-wave velocity and formation bulk density can be used to estimate the dynamic rock mechanical properties (Tixier et al., 1975). Direct measurement of DTS by geophysical well logs, on the other hand, might be difficult and costly. As a result, using machine learning techniques to forecast DTS based on other characteristics accessible in geophysical data has proven to be a promising approach.

In this study, we propose an innovative approach that combines clustering and regression techniques using machine learning. The clustering process is conducted using the Bayesian Gaussian Mixture method, enabling the formation of clusters with comparable attributes. This facilitates the identification of distinct patterns and behaviors associated with the shear wave (DTS). The DTS prediction is then performed in each identified cluster using three regression algorithms: MLP (Multi-Layer Perceptron), RF (Random Forest), and LSQ (Least Square Regression). These algorithms can learn the relationship between the input factors and the DTS, yielding in precise and dependable predictions.

We employed well-accepted metrics such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and R2 score to assess the effectiveness of regression models. These measures enable quantification of prediction quality and comparison of algorithm performance. We show that integrating clustering and regression in a synergistic manner with cutting-edge machine-learning methods can significantly enhance forecasts of shear wave velocity. By utilizing an integrated method, it may be possible to characterize oil reserves more precisely and gain a deeper understanding of the characteristics of subsurface rock formations. Consequently, oil exploration and production activities may be improved.

Methodology

Applying clustering techniques to well logs entails multiple phases. Five Norwegian Sea logs were first chosen for training and testing, along with two more logs for algorithm validation. The "Viking Graben" in the south and north is where this choice was made (information from (Bormann et al., 2020) NPD (the Norwegian Petroleum Directorate) provided the information. The region has a variety of geology, including Permian evaporites, sandstones, and shales. Due to the abundance of DTS (Shear Wave Travel Time) records and the accessibility of data on the local lithology, these data were chosen. The blind data, on the other hand, which were excluded from the training set and did not come from the study area, matched the Cambo Oil Field in the North Sea. Off the coast of the United Kingdom (UK), the field is situated about 125 kilometers northwest of the Shetland Islands, and there are around 250 kilometers between the two study fields. Shale was the most prevalent lithology in the dataset used to train the machine learning algorithms in this work, followed by sandstone, marl, and shaly-sandstone. Machine learning libraries including lasio, pandas, numpy, and scikit-learn (Pedregosa et al., 2011) were used in this work, among others. The data were loaded during the research's execution, and then they underwent preprocessing and correlation analysis to get them ready for the regression procedure. Following data training, the characteristics and targets were chosen, and outliers were eliminated.

The objective of this study is to forecast DTS, or the transit time of shear waves. The variable factors include

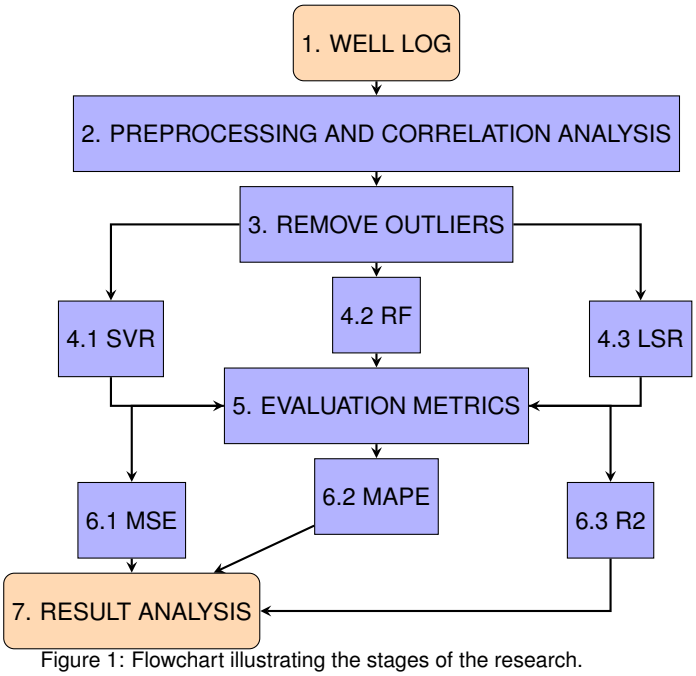

Figure 1: Flowchart illustrating the stages of the research.

DTC (heat wave transit time), RHOB (density), NPHI (neutronic porosity), and DEPTH (depth). Figure 2 illustrates the correlation between the input logs versus the target log (DTS). The second part performed clustering using Bayesian Gaussian Mixture, in addition to building a regression model separately for each cluster. After that, MLP (Multi-Layer Perceptron), RF (Random Forest), and LSQ(least square regression), models were built for each cluster. To create different training and testing sets, the data was divided.

The effectiveness of the various strategies was then evaluated using a variety of measures. Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and R^2 score, the sckit-learn package includes routines for each of these measures. The prediction error of the regressor is calculated using the mean squared error (MSE) function and compared to the actual target value. The better performance of the regression model is indicated by a lower MSE.In the initial step, the analysis of results involved evaluating the Mean Squared Error (MSE) metric, defined as:

$$\text{MSE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2, \qquad (1)$$

where n_{samples} represents the number of samples in the dataset. For each sample, the squared difference between the actual value (y_i) and the corresponding predicted value ($\hat{y}_i$) is calculated. These squared errors are then summed and divided by the total number of samples to obtain the mean. (Wang and Bovik, 2009), During this analysis, the MSE was used to quantify the average of the squared errors between the predicted values ($\hat{y}$) and the actual values (y).

An alternate performance statistic for regression models that has a relatable meaning is the Mean Absolute Percentage Error (MAPE). Hetre, the Mean Absolute Percentage Error (MAPE) is a metric used to evaluate the performance of regression algorithms. It measures the average percentage difference between the predicted values ($\hat{y}$) and the actual values (y). The formula for MAPE is:

$$\text{MAPE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{|y_i - \hat{y}_i|}{\max(\varepsilon, |y_i|)}, \qquad (2)$$

where n_{samples} represents the number of samples in the dataset. For each sample, the absolute difference between the predicted value ($\hat{y}_i$) and the actual value (y_i) is divided by the maximum of a small positive value ε and the absolute value of y_i. This division ensures that the metric does not produce undefined results when y_i approaches zero.

The amount of variance in the dependent variable (y) that can be explained by the independent variables in the model is expressed in terms of the coefficient of determination, also known as R-squared. Evaluating the percentage of variation that can be explained, acts as a gauge of the model's goodness of fit and shows how well the model can predict samples that have not yet been seen. The interpretation of R-squared R^2 may not be directly comparable across various datasets because the variance is affected by the dataset. The score can be positive, which indicates that the model performs worse than a constant model. However, it is vital to remember that the greatest possible value is 1.0. An R^2 score of 0.0 would be produced by a constant model that only predicts the expected (average) value of the dependent variable without taking the input features into account. Defined as,

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad (3)$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \qquad (4)$$

Results and Discussions

After using regression methods in conjunction with data clustering, comparative graphs were made between the regression values and the actual value of the shear wave velocity (DTS). Each cluster was compared independently using one of three regression methods: RF (Random Forest), LSR (Least Square Regression), and MLP (Multi-Layer Perceptron). The graphics displayed the actual DTS values in green, while the values predicted by LSR, MLP, and RF were depicted in blue, red, and pink, respectively. This visual analysis enables us to assess the efficacy of the regression approaches in calculating the DTS. Figure 3 and 4 shows the comparative graph of the predicted and actual values of the DTS for the well drilled in the Viking Graben and Cambo region, called "blind well", which was not used in the training of the algorithm. Additionally, metrics that took into account the average Mean Squared Error (MSE) (see Table 2, Mean Absolute Percentage Error (MAPE), and R2 score were generated for each cluster (see Table 1). These metrics gave a numerical evaluation of the accuracy of the forecasts in each situation.

By analyzing the metrics, you can gain insight into the performance of the regression methods. In general, all methods showed very close results, indicating a satisfactory ability to predict DTS. When the R2 score,

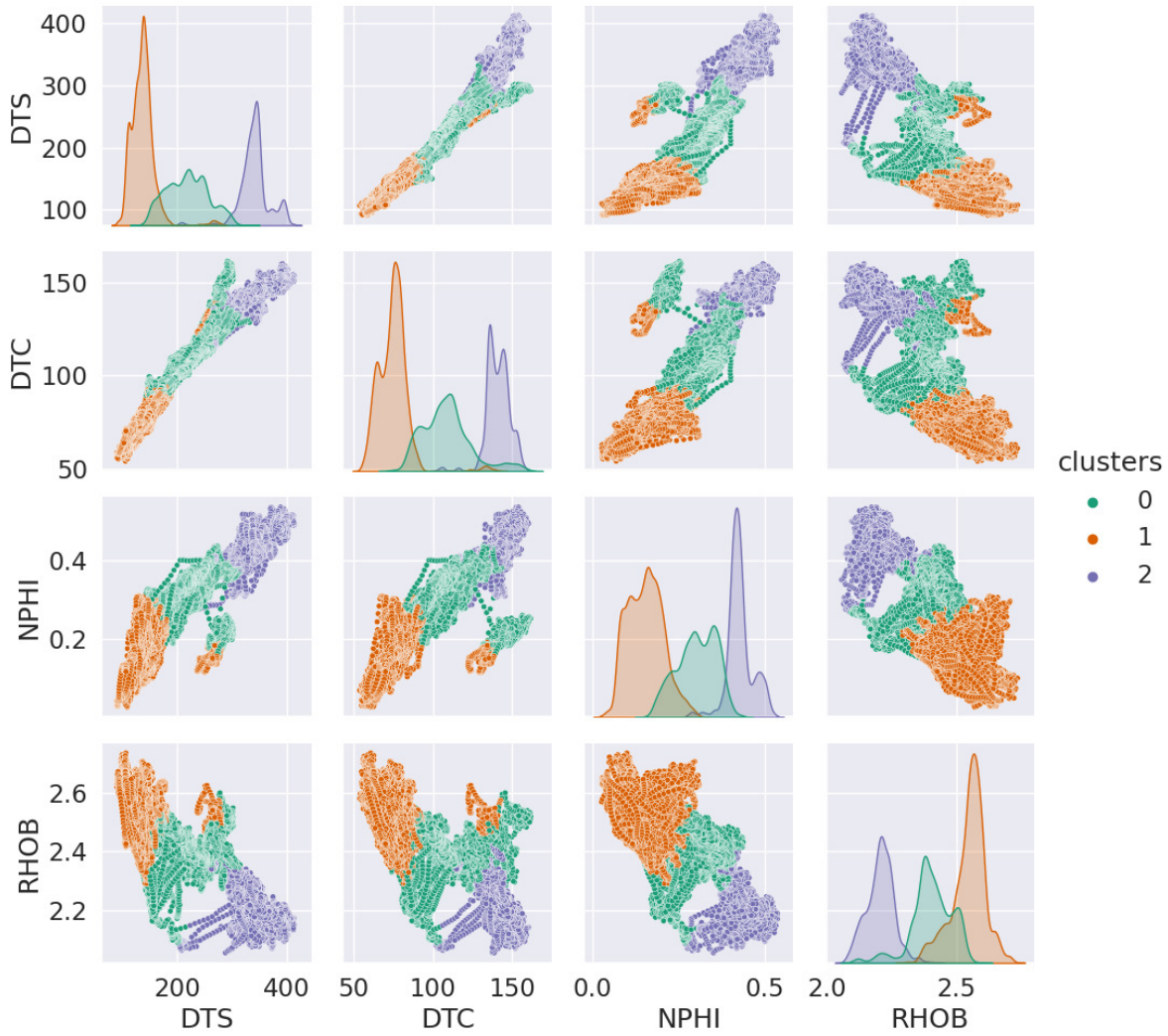

Figure 2: Correlation maps of the well-logs inputs and output (DTS).

Table 1: Results for R^2 metric for the test, validation, and blind logs related to each method of regression.

Cluster	LSR	MLP	RF
TESTE_CLUSTER	0.988597	0.992259	0.998031
BLIND1_CLUSTER	0.973148	0.967607	0.954748
BLIND2_CLUSTER	0.954204	0.947413	0.913577

Table 2: Results for MSE metric for the test, validation, and blind logs related to each method of regression.

Cluster	LSR	MLP	RF
TESTE_CLUSTER	0.011413	0.007747	0.001970
BLIND1_CLUSTER	0.026852	0.032393	0.045252
BLIND2_CLUSTER	0.045796	0.052587	0.08423

which represents the proportion of variance explained by the model, was considered, the RF approach shone out, with values greater than 0.95 in all clusters. This implies that RF was able to explain the majority of the variability in the shear wave velocity data. In terms of error metrics, MSE, and MAPE, the RF technique yielded the lowest values across all clusters. This suggests that, when compared to LSR and MLP, RF predicted DTS values more accurately.

When analyzing 2, we can observe that RF consistently achieved the lowest MSE values across all clusters, indicating a lower mean squared error between the predicted and actual values. On the other hand, LSR and MLP generally had slightly higher MSE values compared to RF.

In summary, based on the R2 metric, RF outperformed LSR and MLP in the test and blind log clusters. Additionally, in terms of the MSE metric, RF consistently achieved the lowest values, indicating higher precision in its predictions.

When analyzing the comparative graphs in Figures 3 e 4 of the methods' predictions, we were able to confirm the analysis conducted using the metrics, as we observed a good fit of the predicted DTS curves with the original data curve. This observation applies to all prediction methods used, reinforcing their high efficiency.

However, upon closer examination of the graphs, it can be noted that the RF method slightly outperforms the others. Its predicted DTS curves exhibit an even better

fit with the original curve, displaying a greater similarity in terms of shape and trend. Additionally, RF demonstrates fewer outliers compared to the other methods, indicating a greater ability to capture the variations and patterns present in the data.

These additional results further support the efficiency of the RF algorithm in DTS prediction, providing greater confidence in its estimates. In the context of this study, RF showed superiority over LSR and MLP both in terms of performance metrics and the visual analysis of the prediction curves.

Conclusions

Finally, this research demonstrates the utility of combining clustering and regression methodologies with machine learning to predict shear wave velocity (DTS) in the petroleum business. The suggested approach provides a more accurate calculation of DTS by utilizing seismic data and well-logging information, which is critical for petrophysical evaluation and understanding of subsurface rock properties. The Bayesian Gaussian Mixture clustering method makes it easier to identify unique DTS patterns, while the MLP, RF, and LSQ regression methods allow for exact predictions within each cluster. MSE, MAPE, and R2 scores were used to evaluate the performance of the regression models, with RF regularly surpassing LSR and MLP in terms of accuracy. These findings illustrate the use of machine learning approaches to improve reservoir characterization and oil exploration are two examples of activity. This integrated method can lead to a better knowledge of subsurface rock formations, resulting in more precise oil reserve characterization and better business decision-making.

Acknowledgments

Acknowledgments to the UFPA Faculty of Geophysics, CNPq for financing the creation of this study with a Scientific Fund (Initiation Scholarship) and to Dr. José Jadsom for the assistance of his knowledge during the research.

References

Bormann, P., P. Aursand, F. Dilib, P. Dischington, and S. Manral, 2020, Force machine learning competition.

He, J., H. Li, and S. Misra, 2019, Data-driven in-situ sonic-log synthesis in shale reservoirs for geomechanical characterization: SPE Reservoir Evaluation & Engineering, **22**, 1225–1239.

Na'imi, S., S. Shadizadeh, M. Riahi, and M. Mirzakhanian, 2014, Estimation of reservoir porosity and water saturation based on seismic attributes using support vector regression approach: Journal of Applied Geophysics, **107**, 93–101.

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.

Tixier, M., G. Loveless, and R. Anderson, 1975, Estimation of formation strength from the mechanical-properties log (incudes associated paper 6400): Journal of Petroleum Technology, **27**, 283–293.

Wang, Z., and A. C. Bovik, 2009, Mean squared error: Love it or leave it? a new look at signal fidelity measures: IEEE signal processing magazine, **26**, 98–117.

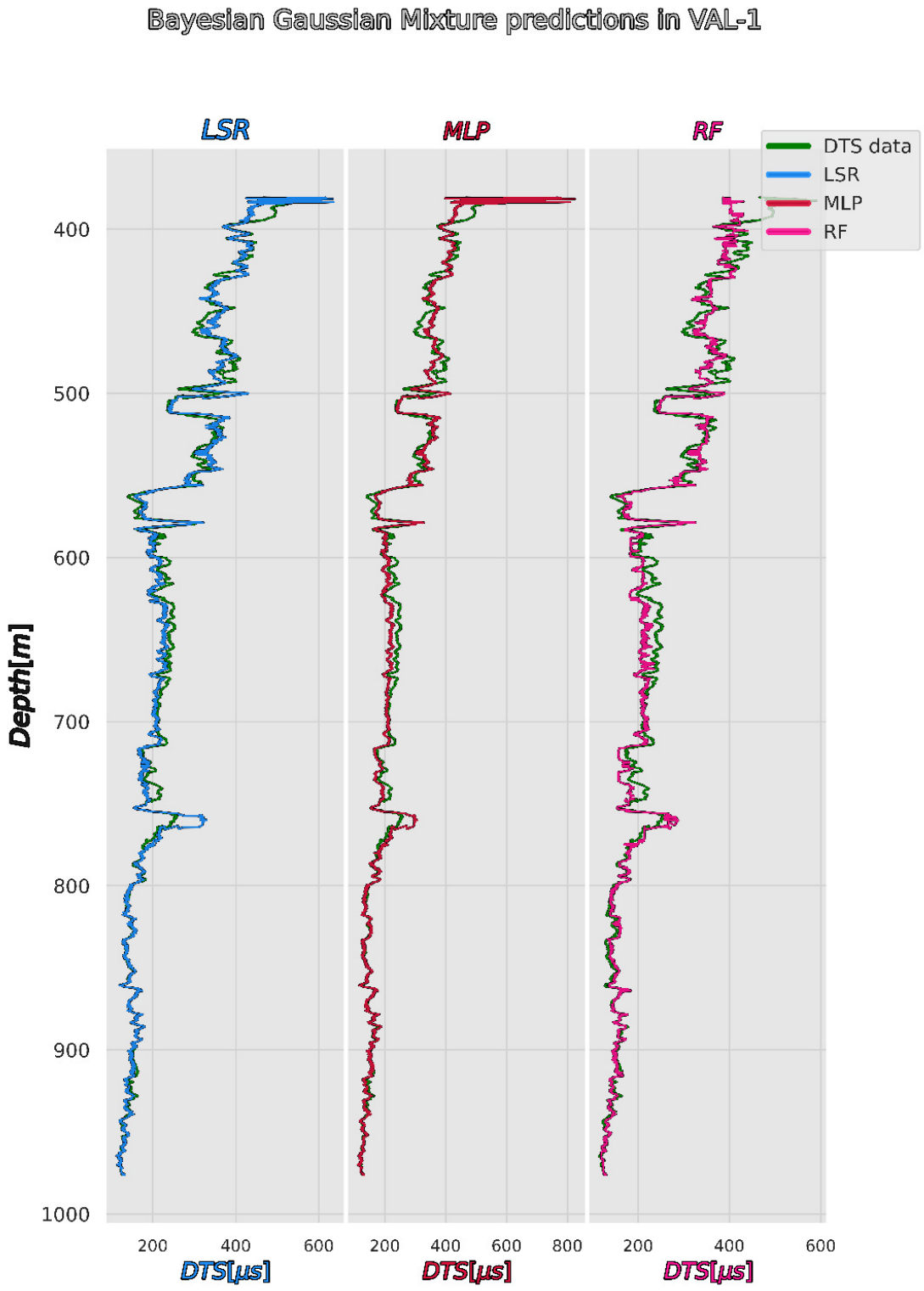

Figure 3: Comparative analysis of predicted and actual values of DTS for the well drilled in the Viking Grab region, known as the 'blind well', which was not used in the algorithm's training.

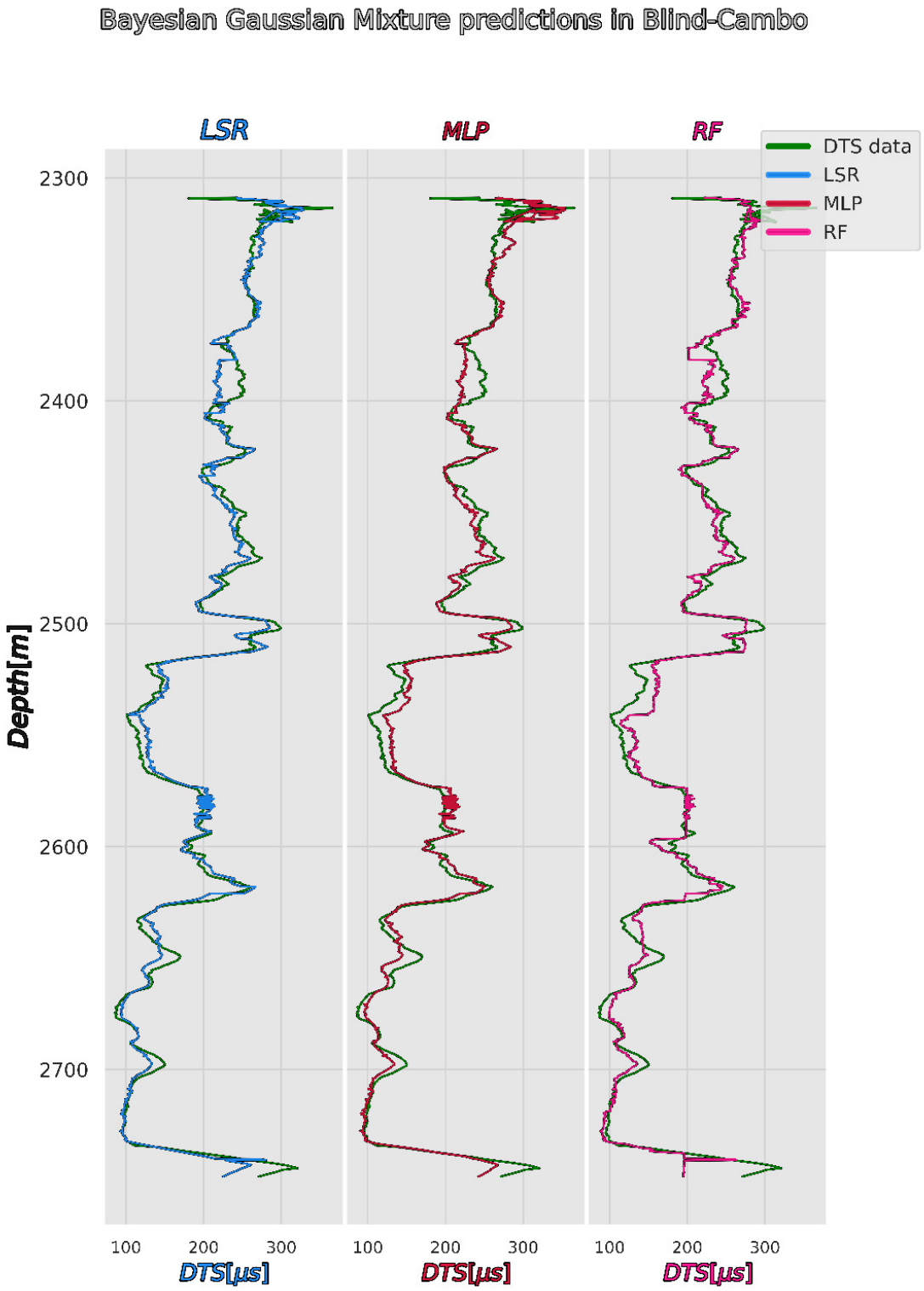

Figure 4: These logs depict data from the Cambo well in Scotland. It's worth emphasizing that this data comes from outside the region, which lends a unique viewpoint to our analysis.