AEGC2019 Data to Discovery
Australasian Exploration Geoscience Conference
2-5 September 2019 • Perth, Western Australia
Incorporating the AIG, ASEG, PESA, and WABS

Using machine learning to predict total organic content – case study: Canning Basin, Western Australia

Russell Menezes*
RadixGeo
3/334 Belmont Avenue,
Kewdale, 6105
russell@radixgeo.com

SUMMARY

Total Organic Content (TOC), one of the most important parameters for sweet spot mapping of unconventional plays is estimated using a RockEval Pyrolysis technique. This process being expensive and time-consuming, the sampling rate per well is very low. In this paper, we use Machine Learning to develop a model that can use conventional wireline logs to estimate TOC data. This model is based on wells in the Canning Basin, Western Australia.

Key words: Machine Learning, TOC, Wireline logs, Canning Basin, Sweet spots.

INTRODUCTION

The main objective of this paper is to develop a technique to quantitatively characterize source rocks around the world using conventional wireline logs instead of laboratory based chemical techniques.

TOC is a very important parameter in assessing the potential of unconventional reservoirs. TOC is normally directly measured from core samples in the laboratory using a technique called RockEval Pyrolysis. As this is a laborious process, the measurements usually are limited in number over a small section of the core. Moreover, these measurements are not a true representation of the formation TOC as the sample selection is affected by sampling bias. Very rarely is a well sampled at regular intervals independent of the sampler's decision to selectively sample at the darkest (black shale) locations.

Wireline logs are measurements made, usually, along the entire length of the wellbore with a high sampling rate. Source rocks have special responses to these wireline logs, which make them distinguishable from the surrounding rocks. These logs, therefore, can be used to detect the presence of source rocks. Additionally, a wireline log-based source rock predictor can provide a continuous measurement of TOC throughout the formation interval.

Traditional TOC estimation techniques are based on measuring differences in density and resistivity of the formation, which helps differentiate organic matter from its inorganic counterparts (Schmoker, 1979). As the density of the formation can be affected by the presence of heavier minerals and resistivity by the tightness of the formation, they, by themselves are not good indicators of organic matter. Some other traditional techniques use the Sonic Log (interval transit time), either by itself (Dellenbach et al., 1983; Zhu et al., 2010) or along with resistivity and sonic (Passey et al., 2010) to estimate TOC. The transit time parameter is known to be strongly affected by interbedded carbonate layers and other mineralogical differences of the matrix which is why it is not a good indicator all the time.

Due to the complex nature of unconventional reservoirs, traditional TOC estimation techniques might not always be successful. Moreover, most of the techniques were developed for shales in the US, which are different to the ones in Australia (Tan et al., 2015). This calls for the need of a model-based, data-driven approach which can be used globally to predict TOC values, provided the model is accurately trained.

Previous work on using advanced machine learning techniques like Artificial Neural Networks (ANN) and Support Vector Machine (SVM) algorithms to predict TOC data, has shown that the relationship between TOC data and log values is non-linear (Emelyanova et al., 2016). These types of algorithms can account for non-linear relationships by generating non-linear input-output mapping functions throughout the learning process.

In this paper, we focus on the Goldwyer and Laurel Formations in the onshore Canning Basin of Western Australia. We use wells containing both core TOC data and the basic conventional wireline logs. A model is then trained by calibrating it against core TOC values and this model is used to predict TOC values where core measurements do not exist.

METHOD AND RESULTS

Exploratory Analysis and Data Cleaning

One of the biggest tasks in any machine learning problem is building a clean and accurate database of input data. "Garbage in, Garbage out" is a very famous saying in the data science world. In this step, I filter through the wells and select those wells that have common conventional wireline logs and core TOC data. I then visualize the data using various statistical plots to remove erroneous data. Figure 1A shows the distribution of core TOC data using a violin plot. It is evident that there are few outliers in the dataset (the long, tapered end), which after cleaning, shows a better, balanced distribution (Figure 1B). Figure 2 shows that the individual logs have the right correlation with each other. For example, the interval transit time (DT) decreases with depth, whereas an increase in porosity (NPHI) corresponds to increased interval transit time (DT).

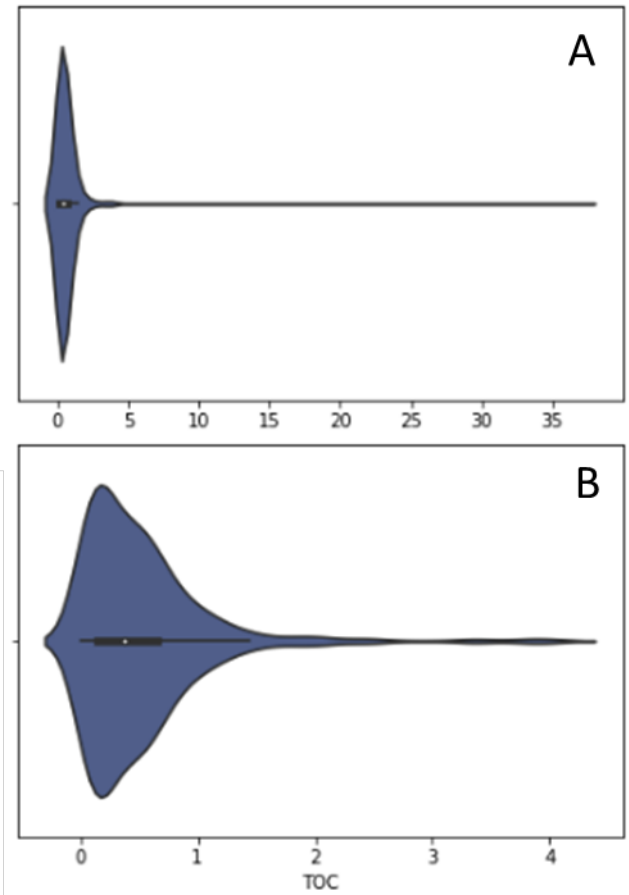

Figure 1. Violin plot showing the distribution of core TOC data before (A) and after (B) data cleaning.

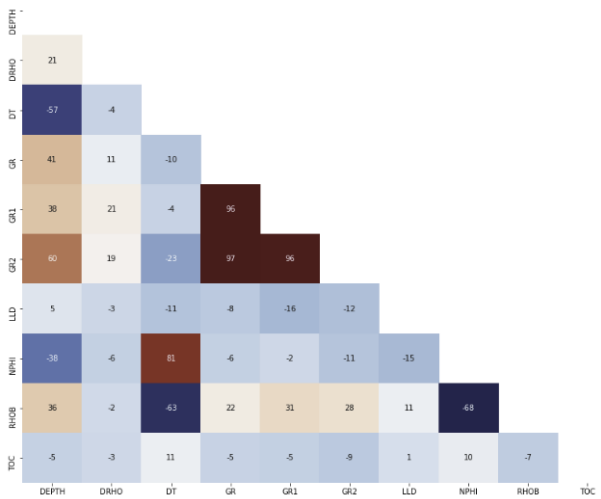

Figure 2. Heatmap showing the correlation between individual wireline logs. Text within each box corresponds to R^2 percentage correlation value.

Feature Engineering

Feature engineering is the process of using one's domain knowledge to help machine learning algorithms work more effectively by generating new input features. In this paper we use our geology knowledge along with information regarding the Canning Basin to generate features that better represent the true underlying relationship of the model which we are trying to create. Figure 3 shows that by splitting the data according to its facies, we can increase the correlation between the individual logs and TOC severalfold. Additional feature engineering techniques are still being employed at the moment, which will be published in the final paper.

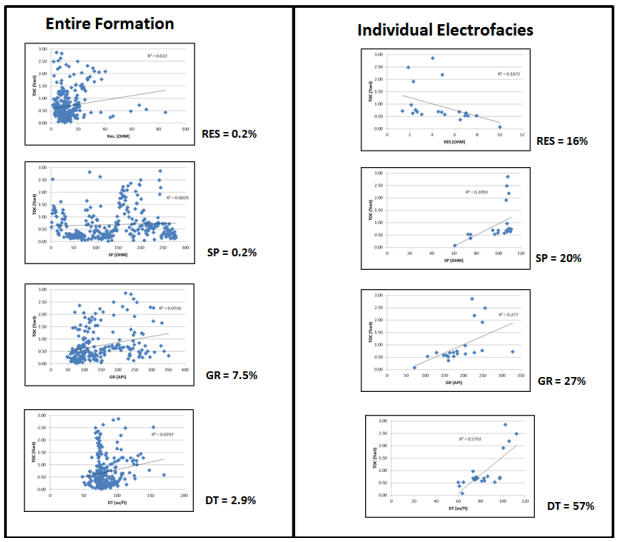

Figure 3. Figure 9b: Cross plots showing relationships between TOC and RES, GR, SP, and DT logs in the Laurel Formation before splitting them according to their facies

Model Training

I employed five algorithms to create our model. Three regularized regression algorithms, i.e. Lasso, Ridge and Elastic-Net and an ensemble of tree-based algorithms, i.e. Random Forests and Boosted Trees.

Regularization (in the context of linear regression) is a technique used to prevent overfitting by artificially penalizing model coefficients. It adds a penalty factor to the cost function, thereby dampening coefficients or removing features entirely. These algorithms suffer from two main flaws: they are prone to overfitting with many input features and they cannot easily express non-linear relationships. Random Forests is an ensemble technique in which large numbers of 'strong' decision trees are trained in parallel and their predictions are combined through bagging. Boosted trees train a sequence of 'weak', constrained decision trees and combine their predictions through boosting

I will standardize the input data and tune the hyperparameters using a cross-validation loop to optimize the model.

Results

As this study has only progressed to the feature engineering stage so far, no concrete results are available at this moment. Some preliminary model training was attempted without any feature engineering and the results obtained seemed promising. Table 1 shows the regression coefficients (R^2) and Mean Absolute Errors (MAE) obtained for the five algorithms which were trained on data from 10 wells in the Canning Basin. These results will be improved significantly in time for the final paper submission.

Table 1. Regression Coefficients (R2) and Mean Absolute Errors (MAE) for the algorithms employed in this study.

Lasso		Ridge		Enet	
R^2	*MAE*	R^2	*MAE*	R^2	*MAE*
0.05	0.32	0.03	0.32	0.04	0.32

RF		GB	
R^2	*MAE*	R^2	*MAE*
0.22	0.26	0.12	0.28

CONCLUSIONS

From the preliminary results, it is evident that it is possible to estimate TOC using a model-based, data-driven approach. Even though these results require a lot of improvement, we can still see that the Random Forests and Gradient Boosting algorithms perform significantly better due to their ability to model non-linear relationships. Additional conclusions will be added at the time of submitting the final paper after the study has concluded.

REFERENCES

Dellenbach, J., J. Espitalie, and F. Lebreton, 1983, Source rock logging: Transactions of the Eighth European Symposium, SPWLA, paper D.

Emelyanova, Irina & Pervukhina, Marina & Clennell, Michael & Dewhurst, David. (2016). Applications of standard and advanced statistical methods to TOC estimation in the McArthur and Georgina basins, Australia. The Leading Edge. 35. 51-57. 10.1190/tle35010051.1.

Passey, Q. R., K. Bohacs, W. L. Esch, R. Klimentidis, and S. Sinha, 2010, From oilprone source rock to gas-producing shale reservoir — Geologic and petrophysical characterization of unconventional shale gas reservoirs: International Oil and Gas Conference and Exhibition, SPE, http://dx.doi.org/10.2118/31350-MS.

Schmoker, J. W., 1979, Determination of organic content of Appalachian Devonian shales from gamma-ray logs: AAPG Bulletin, 63, no. 9, 1504–1537.

Tan, M., X. Song, X. Yang, and Q. Wu, 2015, Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: A comparative study: Journal of Natural Gas Science and Engineering, 26, 792–802, http://dx.doi.org/10.1016/j.jngse.2015.07.008.

Zhu, Y., A. Martinez, E. Liu, C. Harris, S. Xu, M. A. Payne, and M. Terrell, 2010, Understanding geophysical responses of shale gas plays: Shale Workshop, EAGE, http://dx.doi.org/10.3997/2214-4609.20145376.