Do you Trust ML Data? OSDU Compatible File Lineage and QA/QC on ML-Derived Datasets from Rock Images

Jonathan Dietz, Raghd Gadrbouh, Edward Jarvis, Matthew Fry, Andrew Mills, Haoyi Wang

CGG Services (UK)

Abstract

Advances in data analytical techniques are streamlining processes, enabling ever larger datasets to be generated and synthesised. While these techniques can be extremely useful to quickly gain insight into data trends it is important to understand the limitations of such approaches and have in place a robust process for data QC. A QC process needs to work in conjunction with the new technologies at the same scales and timeframes and not act as the next bottle neck in data analytical processes. QC processes are needed if large, collated datasets are to be used with confidence in different use cases. In this instance machine learning was utilised in the processing of a large number of core (>1,300) and thin section (>3,000) images. Techniques were employed in image segmentation and object classification with the goal of predicting grain type and microfacies in addition to determining pore count and property statistics from thin section photomicrographs. In addition processing of core images looked to identify core colour and predict lithofacies, lithology and UV fluorescence at a pixel level scale from white and ultra violet light images.

The results of both approaches demonstrated an excellent correlation with independently collated datasets including: x-ray diffraction spectroscopy, helium porosity, total organic carbon and both core and wireline gamma. While correlation indicated a potentially reliable and rapid proxy for many geological and petrophysical parameters, a series of QC rules were also run over the data set to identify outliers and data inconsistencies and also act to flag potential different use cases based on the quality of the data. Running the QC processes over the greater than 100,000 row dataset resulted in several million QC records. While a large proportion of the data (>80%) was suitable for general reservoir screening purposes a much lower proportion of data (60%) was deemed appropriate for the purpose of petrophysical calibration where a greater degree of precision is required.

In light of the need for a universal and open environment for the sharing of common processes and practices the OSDU (Open Subsurface Data Universe) provides a common shared framework for lineage tracking, reference data, QC rules and technical use cases. The approach taken in this study utilised the OSDU data modal ensuring that the processes and steps employed are compatible and therefore recognised and accessible to the wider geological community.