**Applying Data Analytics & Machine Learning Methods for Recovery Factor Prediction and Uncertainty Modelling**

*Munish Kumar, Kanna Swaminathan, Aizat Rusli, Abel Thomas-Hy*

The estimation of recoverable hydrocarbons, or field recovery factor (RF), is a critical variable for Oil and Gas (O&G) companies to identify profitable investments, plan and optimise field development, manage/ monitor ongoing production, and rank commercial and technical decision making. However, determination of RF remains one of the greatest uncertainties in O&G projects, particularly in the early stages of field life.

RF prediction is difficult because a huge number of variables can affect the final RF of a field, some of which O&G operators have no control over, such as fluid and rock properties. Others important parameters include completion methods if enhanced oil recovery or waterflooding was done. In early field life, insufficient production data coupled with subsurface uncertainty make RF prediction more of an “art” than a science, and it is often the experience of the operator combined with analogue studies that is used to determine RF. However, operator experience can be biased, and in most cases, there is insufficient data from analogue fields to narrow the RF range.

Utilizing techniques of big data manipulation and machine learning, ERCE has studied 1447 sandstone and carbonate fields extracted from an extensive worldwide database to (a) deconstruct and identify key variables that impact production profiles for the field and (b) create machine learning models that can predict RF based on said key variables. Our data set consists of ~200,000 real world data points. This data set was ‘pythonically’ cleaned and sorted such that only variables where measurements are present for 80% of the cases are kept. Where missing data was present, ERCE utilized regression models to augment the data set. Big data techniques were also used to automatically remove, rename, and combine data sets to reduce data granularity, saving ~1000 man-hours were manual cleaning attempted for the database.

Our work aims to show that not all variables influence RF equally; any machine learning model should therefore be built with variables that have the greatest influence on RF yet have the lowest pairwise correlation. The paper will demonstrate the accuracy of the RF prediction using our machine learning tool and will also discuss recommendations on the applicability of the procedure in different field types, and with different data availability.