**Using Data Analytics on Dimensionless Numbers to Predict the Ultimate Recovery Factors for Different Drive Mechanisms for GoM Fields**

**Abstract**

The ultimate recovery factor is strongly affected by petrophysical parameters, engineering data, structures, and drive mechanisms. The knowledge of the recovery factor is needed for multiple decision makings and it should be known as early as possible. Classical ways of obtaining recovery factor need significantly amount of production data or high quality of analogy studies, and this is usually either too late for a reservoir or too much dependent on operator’s knowledge of other fields.

We propose to aim at the same problem with a different perspective. We capitalize on existing database from the same basin, but explore their parametric relationships between different reservoirs and wells using data analytics. Given the fact that there are hundreds of attributes in a database to quantify a reservoir, and some of records may not be accurate or contradictory to each other, we propose to use dimensionless quantity first to categorize them based on similarity theorems, i.e. geometry similarity, kinematic similarity, and dynamic similarity. These independent dimensionless variables not only reduce the number quantity for data analytics, but also they have particular physical meanings.This paper presents a comparative study of different data mining techniques such as linear regression, ridge regression, least absolute shrinkage and selection operator regression, decision trees, random forests and neural networks for predicting ultimate oil recovery factor. Additionally , statistical significance of various geological, reservoir and engineering parameters is also studied. A public dataset related to oil fields in Gulf of Mexico is used for this study. This dataset consists of 4000 oil reservoirs and each reservoir has 80 attributes.

Initial data cleaning was carried out on this dataset to remove reservoirs with erroneous data entries. Dimensionless reservoir parameters are defined and used for the study to make the models consistent to other reservoirs. 80: 20 training and test datasets are separated and different models defined above were trained using 80% training data set. 20% test dataset was used to evaluate the trained models. A few models based on their intrinsic design predicted the ultimate recovery factor with an accuracy of 5-7%, and a few other models predicted the same with an accuracy of 10-12%. Interestingly Artificial neural network model prediction error is around 1%. In addition to predicting ultimate recovery factor, relative importance of various dimensionless parameters, geological, reservoir and engineering parameters is listed based on these models. This kind of study uses already available reservoir data to provide a quick means to evaluate new oil reservoirs even with limited data. Discuss what mechanisms that you are going to study, please think about.

The innovative of this study is applying similarity theorems by using dimensionless variables to reduce the freedom of variables.

**Objective:**

The ultimate recovery factor is strongly affected by petrophysical parameters, engineering data, structures, and drive mechanisms. The knowledge of the recovery factor is needed for multiple decision makings and it should be known as early as possible. Classical ways of obtaining recovery factor need significantly amount of production data or high quality of analogy studies, and this is usually either too late for a reservoir or too much dependent on operator’s knowledge of other fields. We capitalize on existing database from the same basin, but explore their parametric relationships between different reservoirs and wells using data analytics.

Given the fact that there are hundreds of attributes in a database to quantify a reservoir, and some of records may not be accurate or contradictory to each other, we propose to use dimensionless quantity first to categorize them based on similarity theorems, i.e. geometry similarity, kinematic similarity, and dynamic similarity. These independent dimensionless variables not only reduce the independent variable quantity for data analytics, but also they have particular physical meanings. This paper presents a comparative study of different data mining techniques such as linear regression, ridge regression, least absolute shrinkage and selection operator regression, decision trees, randomforest and artificial neural networks for predicting ultimate oil recovery factor. Additionally, statistical significance of various geological, reservoir and engineering parameters is also studied. A public dataset related to oil fields in Gulf of Mexico is used for this study. This dataset consists of 4000 oil reservoirs and each reservoir has 80 attributes.

Initial data cleaning was carried out on this dataset to remove reservoirs with erroneous data entries. Dimensionless reservoir parameters are defined and used for the study to make the models consistent to other reservoirs. 80: 20 training and test datasets are separated and different models defined above were trained using 80% training data set. 20% test dataset was used to evaluate the trained models. A few models based on their intrinsic design predicted the ultimate recovery factor with an accuracy of 5-7%, and a few other models predicted the same with an accuracy of 10-12%. Interestingly Artificial neural network model prediction error is around 1%. In addition to predicting ultimate recovery factor, relative importance of various dimensionless parameters, geological, reservoir and engineering parameters is listed based on these models.

This kind of study uses already available reservoir data to provide a quick means to evaluate new oil reservoirs even with limited data. The innovative of this study is applying similarity theorems by using dimensionless variables to reduce the freedom of variables.