Application of Machine Learning to Predict CO_2 Enhanced Oil Recovery: A Case Study for the EOR33 Project

David Alexander
Energy Systems Engineering Unit, The University of Trinidad and Tobago
Email: david.alexander@utt.edu.tt

Laurice Phillips
Centre for Information and Communication Technology, The University of Trinidad and Tobago
Email: laurice.phillips@utt.edu.tt

Abstract

As Trinidad and Tobago (T&T) seeks to manage its carbon emissions while boosting oil production via different EOR techniques, it will be consequential to utilize machine learning techniques to enhance the predictability of reservoir performance. Machine learning presents the opportunity to develop proxy models which can be cost effective and can be used for making fast practical and more accurate decisions for field development. This paper presents a machine learning approach to predicting cumulative oil production when CO_2 is injected into an oil reservoir for enhancing oil recovery. A commercial simulator CMG's CMOST-AI was used to generate the data set required for analysis by performing a sensitivity analysis on the parameters that had uncertainty to determine their impact on the cumulative oil production when CO_2 is injected in the EOR 33 project located in South Trinidad. Supervised learning using linear regression, polynomial regression and Random Forest data analysis techniques were used analyze data generated by the commercial software so that the rate and accuracy of future predictions would be greatly enhanced. The linear regression and polynomial regression showed an R^2 of 0.988 and 0.999, and a MAPE of 0.0119 and 0.0031 respectively, indicating that the proxy models generated can be used for reservoirs with similar parameters for predicting oil recovery when CO_2 is injected. The Random Forest model can be used for prediction although it has a much lower R^2 of 0.91 and a MAPE of 0.0339.

Keywords: CO_2-EOR, machine learning, prediction, regression analysis

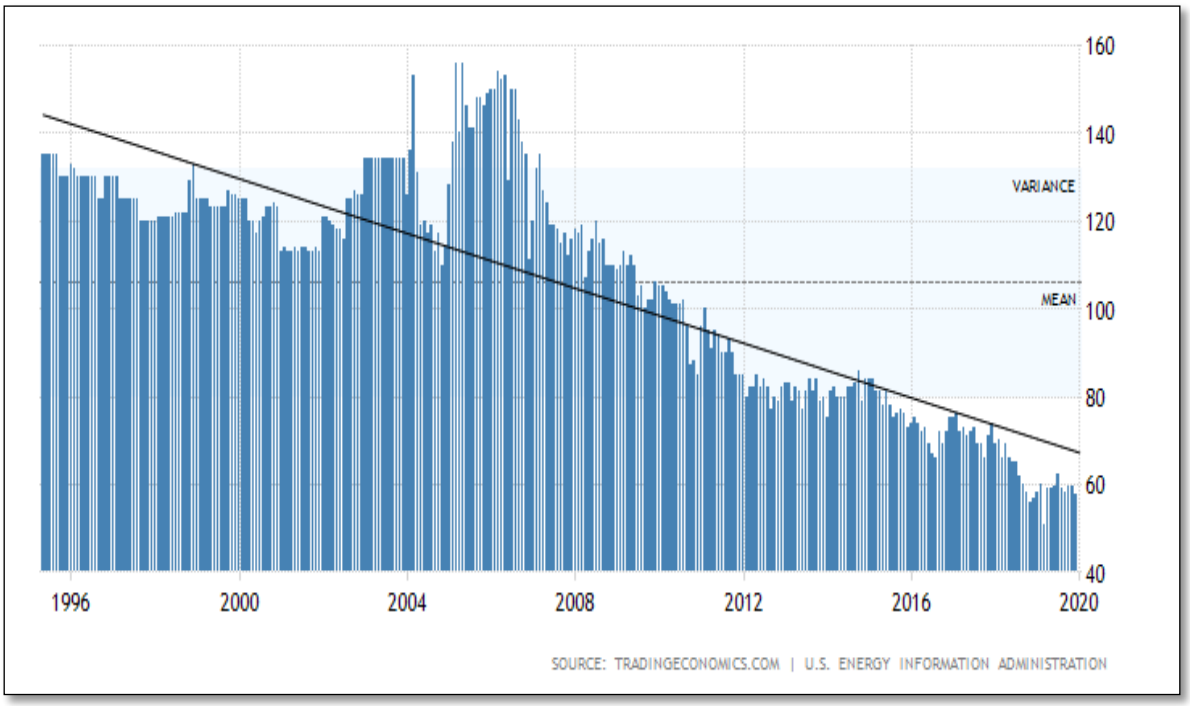
Dr David Alexander *is an Associate Professor in the Energy Systems Engineering Unit at the University of Trinidad and Tobago where he also serves as the Programme Leader. Dr Alexander holds a BSc in Chemistry/Analytical Chemistry and an MSc in Petroleum Engineering from the University of the West Indies. He also holds a Ph.D. in the field of Petroleum Engineering from the University of Trinidad and Tobago (UTT) in collaboration with the University of Texas at Austin. He has further training with the Computer Modelling Group (a reservoir engineering software), Judicial Writing from the University of Nevada and Environmental Mediation from Vermont Law School. Dr Alexander has close to (20) years of teaching/research and professional experience in science and engineering. His main areas of research involve reservoir engineering and waste management.*

Dr Laurice Phillips *is an Assistant Professor in the Centre for Information & Communication Technology at The University of Trinidad and Tobago where he also serves as the Programme Leader for the Masters in ICT. Dr Phillips holds a BSc in Computer Science & Management, an MSc in Computer Science and a PhD in Computer Science from the University of the West Indies. Dr Phillips's doctoral research specialised in digital fingerprint classification where he was awarded local and international patents for a novel technique in digital fingerprint classification using Regular Expression Machine Learning through the University of the West Indies. Dr Phillips has over (20) years of teaching, research and professional experience in computer science and information & communication technology. His main areas of research include Digital Image Processing, Biometric Recognition and Machine Learning techniques.*

Introduction

Trinidad and Tobago's (T&T) economy has relied heavily on the energy sector for several decades, however there has been a constant decrease in crude oil production as seen in Figure 1. T&T averaged 107 thousand barrels of oil per day (BOPD) from 1994 until 2019 to a record low of just over 50 thousand BOPD in March of 2019 (Medica, et al., 2020). To arrest this decline in crude oil production and to create a sustainable economy for T&T, enhanced oil recovery techniques should be implemented. T&T has produced over 320 MMbbls of oil utilizing various secondary and other EOR processes such as WASP, steamflood, and waterflood (Ministry of Energy and Energy Industries, n.d.).

Figure 1
Crude Oil Production Data from Trinidad and Tobago by Year from 1994 to 2019



Note. Source: Trading Economics, n.d.

T&T also has a history of injecting CO_2 for enhancing oil recovery with success (Mohammed-Singh & Singhal, 2005). Predicting the incremental recovery factor from an enhanced oil recovery (EOR) technique is a very crucial task that requires huge investment and expert knowledge to guide EOR laboratory experiments and reservoir simulation studies. Predictive tools based on

machine learning are gaining in popularity in the exploration and production (E&P) industry by enhancing conventional procedures and reducing operational cost under current unstable oil market (Belazreg, Mahmood & Aulia, 2020). The current ultimate oil recovery factor worldwide is about 35%, which means that two-thirds of the oil remain underground. Increasing the recovery factor from 35% to 45% would bring about 1 trillion bbl. of Oil (Labastie, 2011).

CO_2-EOR has gained a lot more traction over the past few years (Adel et al., 2018; Iino, Onishi & Datta-Gupta, 2019; Saira, Yin & Le-Hussain, 2020; Gajbhiye, 2021) to investigate modified CO_2 injection strategies, CO_2-Foam applications, and the effects of miscibility. Other studies focused on CO_2-EOR have shown that not only does CO_2 injection enhance oil recovery when the appropriate reservoirs are selected, but it can also serve as a strategy for sequestering CO_2, aiding the reduction of greenhouse gas emissions. Hosseini, Alfi, Nicot and Nuñez-Lopez (2018) investigated the distribution of CO_2 into the free and residually trapped oil, gas, and brine phases, and several contributing factors. Their numerical simulations encompassed a number of field development strategies, including continuous gas injection (CGI), water alternating gas (WAG), water curtain injection (WCI) and combinations thereof. The authors concluded that while the simulations revealed that CGI maximizes oil recovery and CO_2 storage in absolute volume terms, WAG offers a more balanced approach than other strategies or combinations, and has a better potential to be optimized for optimal performance in the field. WAG injection can produce large amounts of oil and store large volumes of CO_2 with the lowest gross utilization ratio. These types of analyses require further investigation.

The need for computers to make educated decisions is growing exponentially. Various methods have been developed for decision making using observation vectors. Among these are supervised and unsupervised classifiers. Recently, there has been increased attention to ensemble learning – methods that generate many classifiers and aggregate their results. Breiman (2001) proposed Random Forests for classification and clustering (Lowe & Kulkarni, 2015).

Machine Learning

Over the past few years, machine learning (ML) has become an integral part in an extensive variety of day-to-day business (Albon, 2018). The oil and gas sector is no exception. ML has been applied to screening of reservoirs to prediction forecasting of reservoir production performance among other applications. ML is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. The term coined by Arthur Samuel (1959) fits into three broad categories: supervised ML, unsupervised ML, and reinforcement learning. Here we are concerned with supervised Learning (SL) which is the task of learning a function that maps an input to an output, based on example input-output pairs (Russell & Norvig, 2009). It falls into broad types: *classification* and *regression*, where *regression* assumes a response variable is quantitative and hence can ascertain a correlation between dependent and independent variables. On the other hand, *classification* is a more qualitative approach to categorizing a data set. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class (James et al., 2013).

This paper utilises three ML techniques to develop the best method for predicting the cumulative production of oil when CO_2 is injected. This will aid in determining the best strategies for improving oil recovery when CO_2 is utilized. Linear regression, polynomial regression and Random Forest algorithms were utilised to predict the production of oil (objective function). These

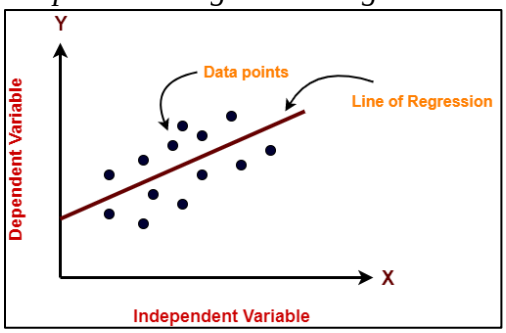
regression models were measured for accuracy using the coefficient of determination r-squared (R^2) and the mean absolute percentage.

Linear Regression (LR)

Regression is a ML technique where the model predicts the output as a continuous numerical value. Linear regression (LR) is a widely used ML algorithm under supervised learning. It is a statistical method that is used for predictive analysis. LR performs the task to predict a dependent variable (target) based on the given independent variable (s). So this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

Figure 2

Graphic Showing Linear Regression



Note. Source: *Linear Regression*. From https://www.gatevidyalay.com/linear-regression-machine-learning-examples

Polynomial Regression (PR)

Polynomial regression (PR) is an algorithm where the independent variable is increasing by higher powers and some coefficients scaling the variables.

$$y = a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n$$

PR is derived using concepts of linear regression with modifications to accelerate accuracy and is a form of linear regression or a special case of multiple linear regression that estimates the relationship to the n^{th} degree, polynomial providing the best relationship between the dependent and independent variable. There are different types of polynomials that differ by degree or order of variables. Simple or multiple linear equations are polynomial equations with a single order.

The models

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

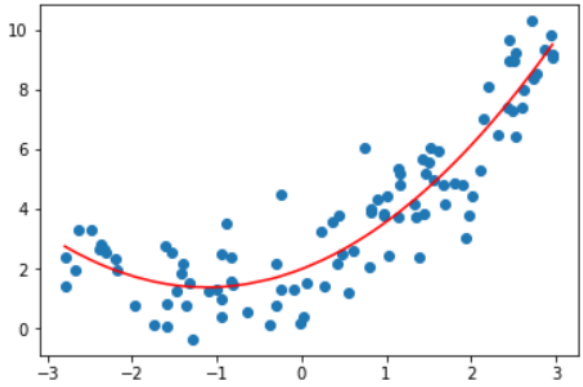
and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

are both examples of second order models or quadratic models. The coefficients of x^1 are linear effect parameter while the coefficients of x^2 are quadratic effect parameters. The original features

of polynomial regression are converted into polynomial features of required degree, then modeled using linear model.

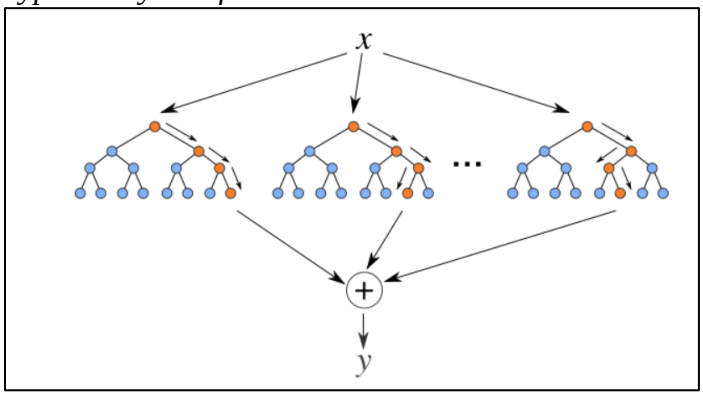
Figure 3:
Graphic Showing Polynomial Regression



Note. Source: *Polynomial Regression*. From https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models

Random Forest (RF)
Random Forests (RF) is an ensemble-based learning algorithm that maps predictors onto a categorical or numerical response. Random Forests consists of multiple decision trees. These decision trees can either be classification or regression trees. The goal is to find $XT * (i,)$ such that a partition score is optimized. A constructed Random Forests model can be used to rank the importance of each parameter (i.e., column) in $XT \subset \mathcal{D}T$. This is done by permutating one parameter at a time. Figure 4 demonstrates a typical layout of Random Forest model, which consists of multiple decision trees.

Figure 4:
Typical Layout of Random Forest Model



Note. Source: Bakshi, 2020

Related Works

One of the main challenges in the oil and gas industry is the time taken to perform hydrological modelling on reservoirs. Many researchers have turned to using machine learning techniques and meta-algorithms (Jordan & Mitchell, 2015) to improve the time and efficiency of analyzing oil and gas fields. Such analyses can involve hundreds of parameters of multivariate data and take several days when performed using conventional techniques on high performance clusters. The introduction of ML techniques has offered several advantages and researchers are better able to perform complex reservoir modelling, with much lower development costs (Mohaghegh et al., 2011). The advent of artificial intelligence, ML techniques and meta-algorithms have made it much easier to identify complex patterns amongst hyper-parameters, fast-track parameter tuning and improved accuracy of predictions in the models.

Li et al. (2022) used machine learning techniques to predict the effectiveness of CO_2-WAG (water alternating gas injection). Their research experimented with several different injection parameters by building over 200 numerical models using CMG numerical simulation software. The research technique used the randomized regression forest algorithm to sample several datasets using the bagging method. In Li's approach, several decision trees were trained by selecting features randomly, using stochastic subspace method, until the best features were selected for full branching. The research produced several different CO_2-WAG development scenarios with various injection parameters from which 70% of them were used as training sets and 30% as testing sets. The results of the experimentation showed a very high prediction accuracy such that the predicted value of the test set was very close to the true value. The research showed a storage efficiency of 1.10%, 3.04%, and 2.24% respectively for the average absolute prediction deviations of cumulative oil production, CO_2 storage amount, and CO_2. Li's ML technique also showed a high computational efficiency in their approach. Their ML techniques were able to predict results for over 200 experimentation scenarios in 10 seconds on average, whereas, conventional CMG simulation method took approximately 108 minutes. Overall, the ML method demonstrated rapid prediction of CO_2-WAG high accuracy and high computational efficiency with various injection parameters by using parameter optimization.

Lima and Lin (2019) used ML techniques to predict the CO_2 and brine leakage in a 200 years' duration geological carbon sequestration (GCS) reservoir project. Lima and Lin's research integrated seismic and geological data using several models containing an injection well, a legacy well in three geological layers. Their research approach generated over 500 simulations to model underground water flow in investigating the effects at GCS sites attributed to CO_2 injection. The research utilized convolution neural networks (CNN) to analyze pressure data, while the inception model was used to analyze the seismic datasets. Less than 100 simulations were used for testing while the rest were used for training. The performance of the ML techniques was compared and it was concluded that there were more accurate predictions of CO_2 and brine leakage on GSC sites when compared between the model using only seismic data and other using both seismic data and well pressure.

Gholami et al. (2019) used ML techniques to predict the values for several properties at CO_2-EOR project. The research involved coupling grid-based surrogate reservoir models (SRM_G) and well-based surrogate reservoir models (SRM_W) to construct a simulation of CO_2-EOR projects at an oilfield. The research investigated SRM_W models for flooding front and simulated several properties changes in the reservoir including pressure, phase saturation, as well as composition of reservoir fluid components at desired time step. The SRM_W models were also used to simulate related well production data, such as oil rate, water rate and water oil ratio, and other data. The research utilized an artificial neural network (ANN) model with one hidden layer to train the SRM_W which was used to estimate response of the reservoir at the well level (rate) with various reservoir parameters and operational constraints. The values of each property at each time step were predicted using one trained SRM. A total of 60 neural networks were used for the SRM_G to predict the interested properties at each time step with 15 models per property. The well productivity index was obtained by integrating SRM_G and SRM_W models by first calculating pressure, phase saturation and other parameters. Traditional numerical modeling took more than 48 hours to run one test, however, with machine learning, the reported total time for running 60 neural network models to deploy the SRMs' calculation took approximately 800 seconds which represented a significant improvement (Yan et al., 2021).

Lv et al. (2021) used ML techniques to optimize CO_2 flooding parameters based on the ANN and the particle swarm optimization (PSO) algorithm. Firstly, a typical ¼ five-point well pattern characteristic model was established. Several variable parameters and ranges were determined for Monte Carlo sampling methods to generate over 3000 geological models. Then, a relation model was established using an ANN and particle swarm optimization (PSO) algorithm was used to optimize the CO_2-WAG production system with given geological parameters. The research used a total of 2100 sets of data where 70% were used for training while the remaining data sets were used for validation. The neural network model and the particle swarm algorithm were combined to optimize the parameters of CO_2-WAG flooding. The results of the research show that the established model is applicable in potential evaluation of enhancing the oil recovery and optimization for parameters in the CO_2-WAG well group and has a high prediction accuracy of 97. Therefore, it can be used to predict the enhanced recovery factor by CO_2-WAG.

Krasnov et al. (2018) used ML techniques to enhance oil recovery predictions. Their research explored proxy models based on multidimensional linear interpolation and showed that using regression models with Random Forest methods to construct a set of decision trees were more efficient. The main principle used was ensemble training where a subsample of a training set was chosen, after which, a decision tree was constructed. With each construction and splitting of a decision tree, separate random features were defined and observed. Key parameters were used to limit the height of the tree, the number of objects in the leaves, as well as the number of objects in the subsample, in which the splitting is performed. Finally, predetermined criteria were used to determine the best features for the model. The research investigated several parameters such as oil properties (density, viscosity, and saturation pressure) as well as heterogeneity of permeability, relative phase permeability, and oil saturation, and generated over 300 simulations. In the traditional approach, a proxy model using linear multidimensional interpolation was used to construct a multidimensional cube of parameters in which each dimension was formed by vectorizing parameters from a resulting function., In the process, new parameters were formed and an interpolation function was applied on the new parameters. In the machine learning approach, regression-building tasks were made using a regressor such as Random Forest method. Multidimensional regression modelling using the Random Forest method provided several advantages over the traditional methods, such as parameter selection and parameter tuning, as well

as computational reduction when generating hydrodynamic models and improved accuracy and predictability of the simulated functions. The research concluded that the computations using machine learning algorithms proved more productive when compared to the traditional computational experiments.

Zeeshan et al. (2021) explored ML applications in the oil and gas industry. Their research offered a comprehensive review of data sciences and ML roles in different petroleum engineering and geosciences segments such as petroleum exploration, reservoir characterization, oil well drilling, production, and well stimulation on newly emerging field of unconventional reservoirs. Their research showed that ML techniques can be used in several ways, including precise drilling, product optimization, well correlation and reservoir characterization. In precise drilling, ML techniques were used to reduce or minimize the high levels of risk and uncertainty in drilling operations. Smart sensors were used to record drilling parameters such as pressure, temperature, and seismic activity in real time to control drilling rate as well as identify and predict risks in real time. In product optimization, machine learning techniques were used to generate analytics and prediction models from reservoir data to improve production efficiency. In well correlation, machine learning techniques were used to analyze geophysical well-log data and construct sophisticated models for geologists. The approach significantly reduced the operational time and complexity for geologists when analyzing hundreds of wells in large oil and gas fields. In reservoir characterization, machine learning techniques were used in developing optimal production and reservoir management strategies. Several materials balance calculations were dependent on accurate determination of properties such as permeability, direction of oil flow, porosity, and pressure. Zeeshan et al.'s (2021) research concluded that AI offered significant advantages in solving problems in several oil and gas industry operations involving prediction, classification and modelling huge amounts of data and hundreds of parameters. Significant operational risk can be reduced by coupling machine learning methods with real-time data to generate more effective models with improved accuracy.

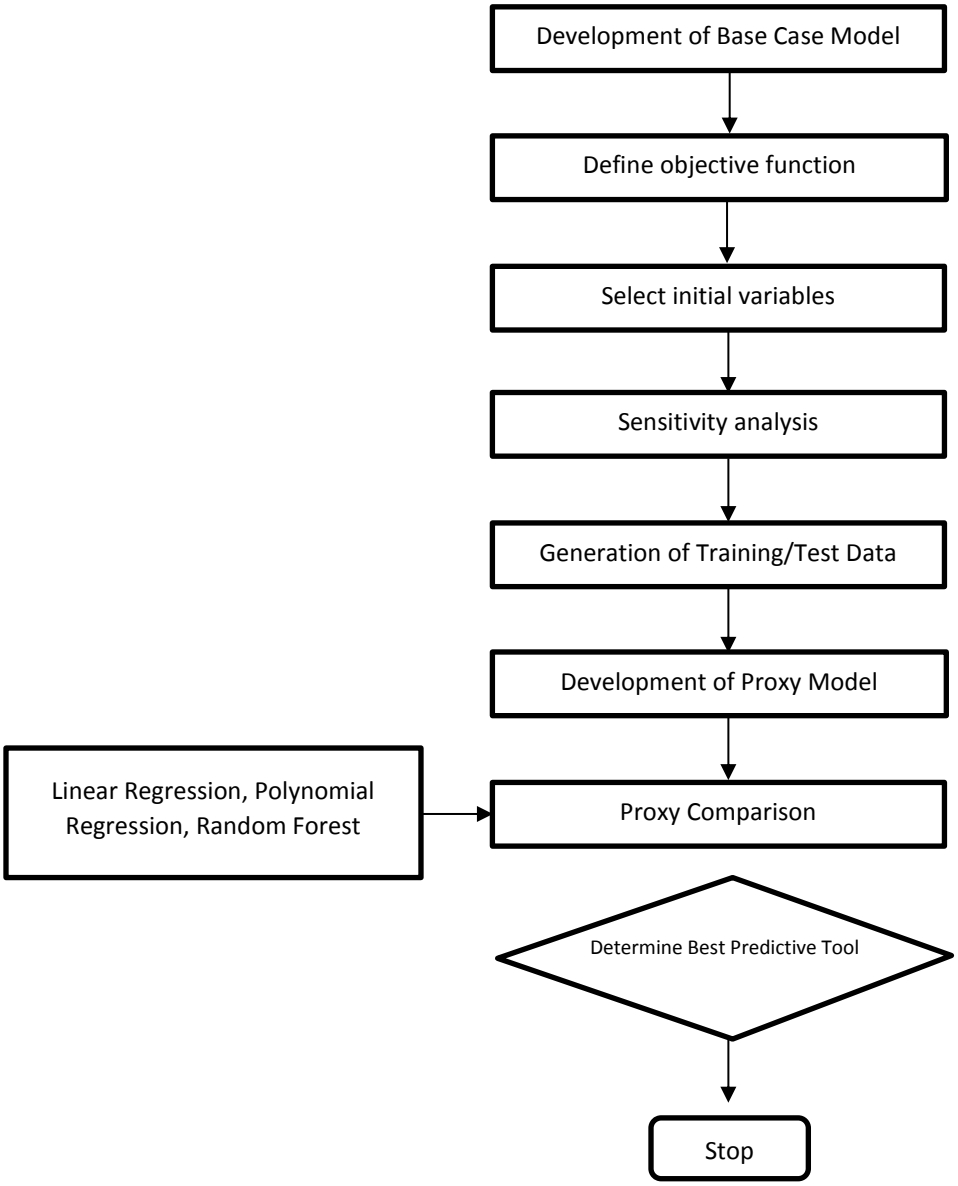
Methodology

The flow chart in Figure 5 outlines the general workflow utilised in this research to conduct this research.

Field Description

The reservoir being investigated is the EOR 33 located in the Forest Reserve field in the southwest peninsula of the island of Trinidad (Figure 6). The field started production in 1961 and had a brief period of water injection in 1976. CO_2 was injected in 1976 after 17.4% of the OOIP was produced. For this study, EOR 33 was selected since it was proven that the field had a favourable response to CO_2 injection. The Forest Reserve field is located on the southern flank of the east-northeast trending Fyzabad Anticline. The reservoirs were deposited under deltaic conditions and are highly heterogeneous and complex (Mohammad-Singh & Singhal, 2004). Shale outs and faults and water-oil contacts define limits of individual oil accumulations and are illustrated in Figure 7. Mohammed-Singh and Singhal (2004) stated that EOR 33's geologic environment is of deltaic deposition where distributary channel fills of lower delta plain environment (that is, it is within the realm of a river-marine interaction) create a highly heterogeneous and complex structure as characterized by levees, crevasse splays, and over bank mud. Moreover, due to changes in fluvial channel paths that were present in the Pliocene depositional period, there are numerous shale lenses.

Figure 5:
Workflow for Proxy Model Development



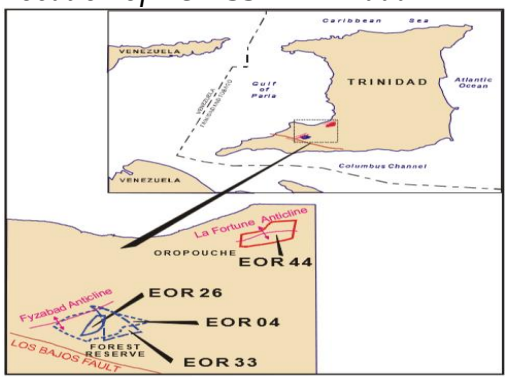
The shale-outs and faults which provide a trapping mechanism, and the water-oil contacts present, define the limits of individual oil accumulation. It is assumed that the faults are sealing; however, flow across them is sometimes identified where there are juxtaposed porous pays on the two sides (that is, the porous segments are close enough to provide transmissibility between them).

The data for EOR 33 used for the simulation modeling was obtained from work done by Mohammed-Singh and Singhal (2004) because the field demonstrated enhanced production when CO_2 was injected. The commercial software CMG-IMEX and computer modelling intelligent

optimization and analysis tool (CMOST-AI) were utilized to develop the reservoir model, and conduct the sensitivity analysis required to generate the training and test data.

Figure 6

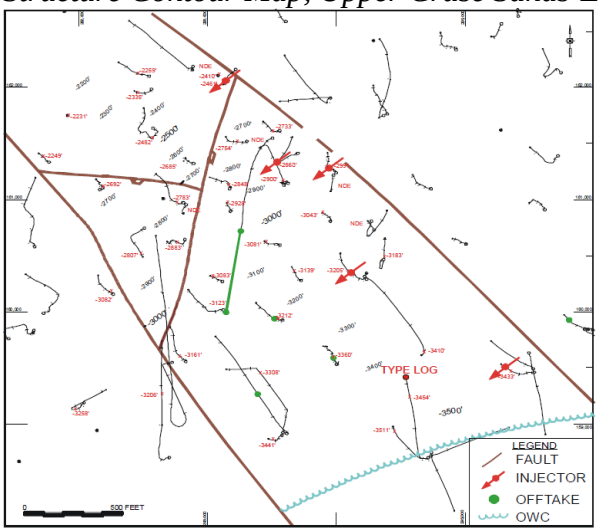
Location of EOR 33 in Trinidad



Note. Source: Mohammed-Singh & Singhal, 2004

Figure 7:

Structure Contour Map, Upper Cruse Sands-EOR 33



Note. Source: Mohammed-Singh & Singhal, 2004

Numerical Modelling

The map for EOR 33 had to be digitized using the Didger software using the structure contour map was obtained from Mohammed-Singh and Singhal (2004). This map, along with the field data presented in Table 1, was used to design the base simulation model using commercial software CMG-IMEX as seen in Figure 8.

Table 1

Reservoir Data for EOR 33 Well- Forest Reserve

ROCK PROPERTIES	**EOR33**
Area (acs)	67
Pay Zone	L. Forest
Depth (ft)	3000
Thickness (ft)	144
Porosity (%)	32
Permeability (mD)	125
Oil Saturation (%)	75
Temperature (^{o}F)	120
Transmissibility (mD-ft/cp)	536
FLUID PROPERTIES	
Initial Conditions	
Reservoir Pressure (psi)	1400
Solution Gas Oil Ratio (scf/bbl)	193
Oil Formation Volume Factor (bbl/bbl)	1.1
Oil Gravity (^{o}API)	19
Oil Viscosity (cp)	16
At CO_2 Flood Start	
Reservoir Pressure (psi)	500
Solution Gas Oil Ratio (scf/bbl)	50
Oil Formation Volume Factor (bbl/bbl)	1.04
Oil Viscosity (cp)	32

Note. Source: Mohammed-Singh & Singhal, 2004

The model was created using 3,750 grid blocks with 50 blocks in the *i* direction, 25 blocks in the *j* direction and 3 blocks in the k direction representing an area of 60 acres. A petrophysical analysis of the log Figure 9 identified sand shale sequence which was taken into consideration. According to the structure map provided in Mohammed-Singh and Singhal (2004), EOR33 consisted of five injection and six production wells. It was assumed that all the injector wells operated at a max bottom hole pressure (BHP) of 2000psi, whilst all producers operated at 100psi min BHP.

Once the base case was developed, the CMOST-Al was used to investigate the impacts of uncertainty parameters on the oil production from the simulation results to determine the best injection rate with the implementation of a CO_2-EOR project.

Figure 8
3D Reservoir Simulation Model for EOR 33 generated using the CMG Software

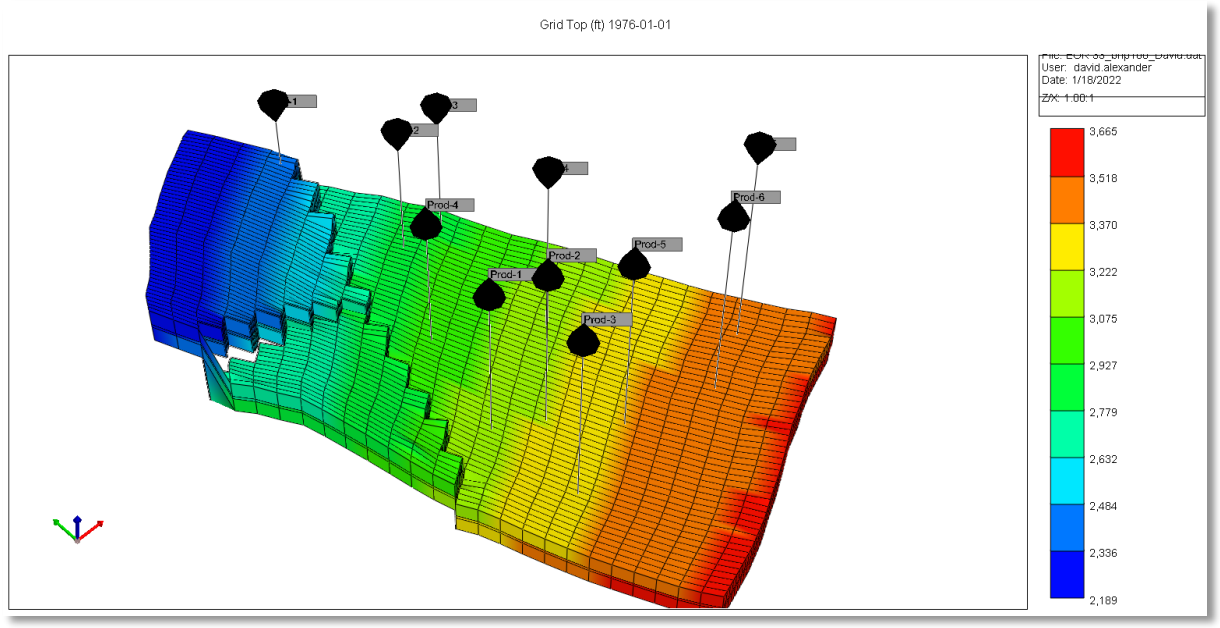
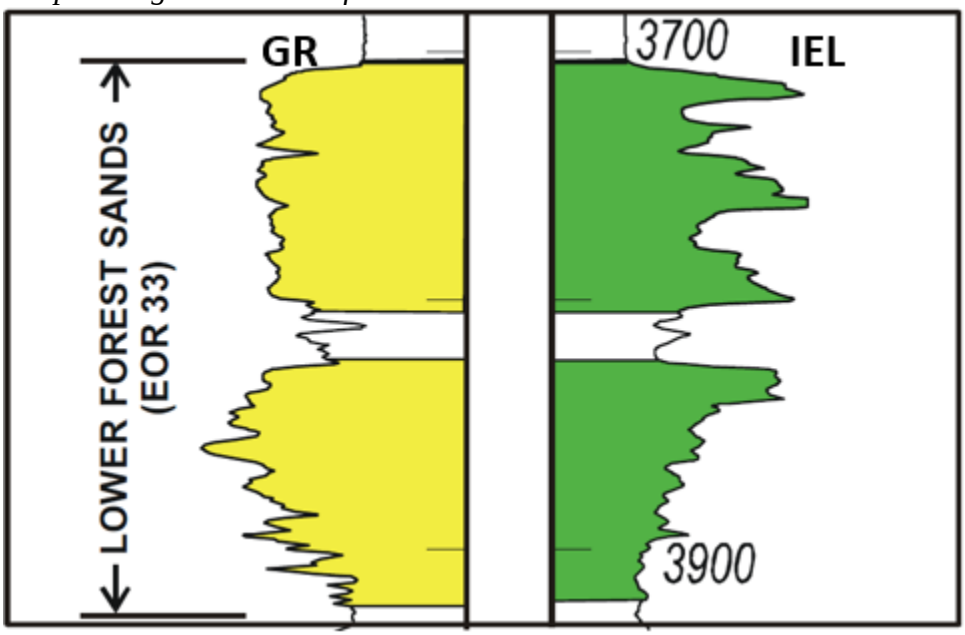


Figure 9
Adapted Log Illustration of EOR 33



Note. Source: Mohammed-Singh & Singhal, 2005

To determine the uncertain parameters that had the greatest impact on the model, a sensitivity analysis was conducted in the CMOST program by applying ranges to each parameter identified for evaluation. The parameters were as follows: relative permeability of connate water, horizontal permeability, vertical permeability, porosity, bottom hole pressure, rock compressibility, connate water saturation, and critical water saturation. Table 2 identifies the parameters, ranges, and actual values applied in the master and base datasets.

Table 2
Parameter Ranges for Sensitivity Analysis

Parameters	Range	Actual Value
KROCW	0.6 – 1	0.8
Horizontal Permeability	75 – 175	125
Vertical Permeability	0.075 – 0.2	0.1
Porosity	0.24 – 0.36	0.32
Bottom Hole Pressure	75 – 150	100
Rock compressibility	3E-06 – 5E-06	4E-06
SWCON	0.15 – 0.25	0.2
SWCRIT	0.15 – 0.25	0.2

After this process, the parameter type was set as ‘continuous real’ to enable CMOST with the ability to correct the identified parameters to any value within the specified range. The objective functions were then selected, and edits were made in the control center to define how the simulations for the sensitivity analysis were to be run in the CMOST engine. Once this was

completed, the model was history-matched using the information presented by Mohammed-Singh and Singhal (2004). For simplicity, one key assumption in this study was CO_2 was injected after primary production. This was done to ascertain the impacts of CO_2 injection on the reservoir.

Generation of Training and Test Data

The commercial software CMG's CMOST-AI was utilized to develop over one thousand experimental results that were used in this study after developing the base numerical simulation model in IMEX. In this study, the response surface methodology was used to generate the base dataset which will be split into training and test data. In CMOST-AI, the software performed the combinations using a random seed, generated based on computer clock time to indicate how the "engine" will be performing, resulting in the experimental tables generating values for each simulation result. CMOST-AI generated multiple simulation runs by changing in each one the different values of the parameters, and calculating how the objective function (cumulative oil production) was altered by the candidate values of each parameter. Subsequently, CMOST-AI then developed the coefficients for each parameter. CMOST-AI then generated the proxy model for the objective function using the simulation results (training data). The software then checked to see if the proxy model created was a good proxy to the actual simulation results. CMOST generated some additional runs that are called "verification" or "test data" in the following quality control plot. The proxy models generated in CMOST-AI were not utilized since each run created 47 new datapoints and its own proxy model. However, for this research, we utilized a combination of all the datasets generated to form a larger database. In this case, we were utilizing over 1000 datapoints to generate our own proxy model which was then tested using R^2 and MAPE. The dataset created was divided into 80% for the training and 20% for the test after it was generated.

Results and Discussion

Results for the Pearson's Correlation Coefficients for each parameter utilised as seen in Figure 10 indicated that there were no strong relationships between the parameters, hence all of them were used in the analysis to develop the proxy models initially.

Figure 10

Correlation Matrix of Key Correlation Parameters

Index	CPOR	GTARGET_STW	Krogcg	Krwiro	MAX_STO_PRODUCERS	Ng	Nog	Now	Nw	PERMK	POROSITY	Sor	Swc
CPOR	1	0.0543634	0.0366001	0.061413	0.0319738	0.0227563	0.0110258	0.0148581	0.0142744	0.036572	0.0269741	0.0538184	0.0268882
GTARGET_STW	0.0543634	1	0.0100783	0.00670969	0.0635388	0.0425708	0.00209428	0.0318525	0.0365143	0.0302023	0.0238934	0.0106578	0.000128373
Krogcg	0.0366001	0.0100783	1	0.0178005	0.0227008	0.0302919	0.0327023	0.0299251	0.0399149	0.0085197	0.0462142	0.0302793	0.0687643
Krwiro	0.061413	0.00670969	0.0178005	1	0.0517949	0.0522347	0.0281131	0.0330488	0.00639186	0.0235909	0.0237975	0.00032987	0.00575849
MAX_STO_PRODUCERS	0.0319738	0.0635388	0.0227008	0.0517949	1	0.00848401	0.00735383	0.0389489	0.000483877	0.00915268	0.0102134	0.00486012	0.0142945
Ng	0.0227563	0.0425708	0.0302919	0.0522347	0.00848401	1	0.00625348	0.0101031	0.0104066	0.0092138	0.0443012	0.00317132	0.0118589
Nog	0.0110258	0.00209428	0.0327023	0.0281131	0.00735383	0.00625348	1	0.0552079	0.0285385	0.0419933	0.01397	0.0370818	0.029572
Now	0.0148581	0.0318525	0.0299251	0.0330488	0.0389489	0.0101031	0.0552079	1	0.0075299	0.0139442	0.0146156	0.0092103	0.0152884
Nw	0.0142744	0.0365143	0.0399149	0.00639186	0.000483877	0.0104066	0.0285385	0.0075299	1	0.001774	0.0161646	0.0252731	0.0154124
PERMK	0.036572	0.0302023	0.0085197	0.0235909	0.00915268	0.0092138	0.0419933	0.0139442	0.001774	1	0.0453384	0.0309102	0.00290581
POROSITY	0.0269741	0.0238934	0.0462142	0.0237975	0.0102134	0.0443012	0.01397	0.0146156	0.0161646	0.0453384	1	0.0233122	0.0290416
Sor	0.0538184	0.0106578	0.0302793	0.00032987	0.00486012	0.00317132	0.0370818	0.0092103	0.0252731	0.0309102	0.0233122	1	0.000301054
Swc	0.0268882	0.000128373	0.0687643	0.00575849	0.0142945	0.0118589	0.029572	0.0152884	0.0154124	0.00290581	0.0290416	0.000301054	1

After removing the parameters with the least impact, the proxy model developed from the machine learning algorithms for the linear regression was:
$0.669\ POR + 0.44\ K_{rogcg} + 0.377\ S_{wc} + 0.184\ N_g - 0.343\ N_{og}$
Whereas the proxy model for the polynomial regression was:
$0.0671\ POR + 0.443\ K_{rogcg} + 0.375\ S_{wc} + 0.181\ N_g + 0.052\ K_{rogcg} * POR + 0.052\ S_{wc} * POR + 0.034\ Ng * POR + 0.032\ N_{ow} * S_{wc} + 0.3558\ N_{og} + 4.3 * 10^9$

All the computed statistical indicators of the three ML models are listed in Table 3. According to the table, the three predictive models can generate a reasonable prediction for LR, PR, and RF. However, the performance of PR models is better than that of other models developed using regression analysis utilizing RF, especially those applied to estimate the Solubility Trapping Index (STI) and Residual Trapping Index (RTI). The LR model is slightly less than the predictive PR model performance but much higher than that of the RF models as can be seen in Table 3 and Figures 11-14.

Table 3
The Statistical Indicator of Predicted and Actual Values for Three Different Machine Learning Algorithms

Statistical Indicator	**LR**			**PR**			**RF**		
	Training	Test	Results	Training	Test	Results	Training	Test	Results
R-squared	0.985	0.989	0.988	0.999	0.998	0.999	0.99	0.92	0.913
MAPE			0.0119			0.0031			0.0339

Figure 11

Cross-Plot of Predicted Vs Actual for Linear Regression, Polynomial Regression and Random Forest Algorithms (Training Data vs. Test Data)

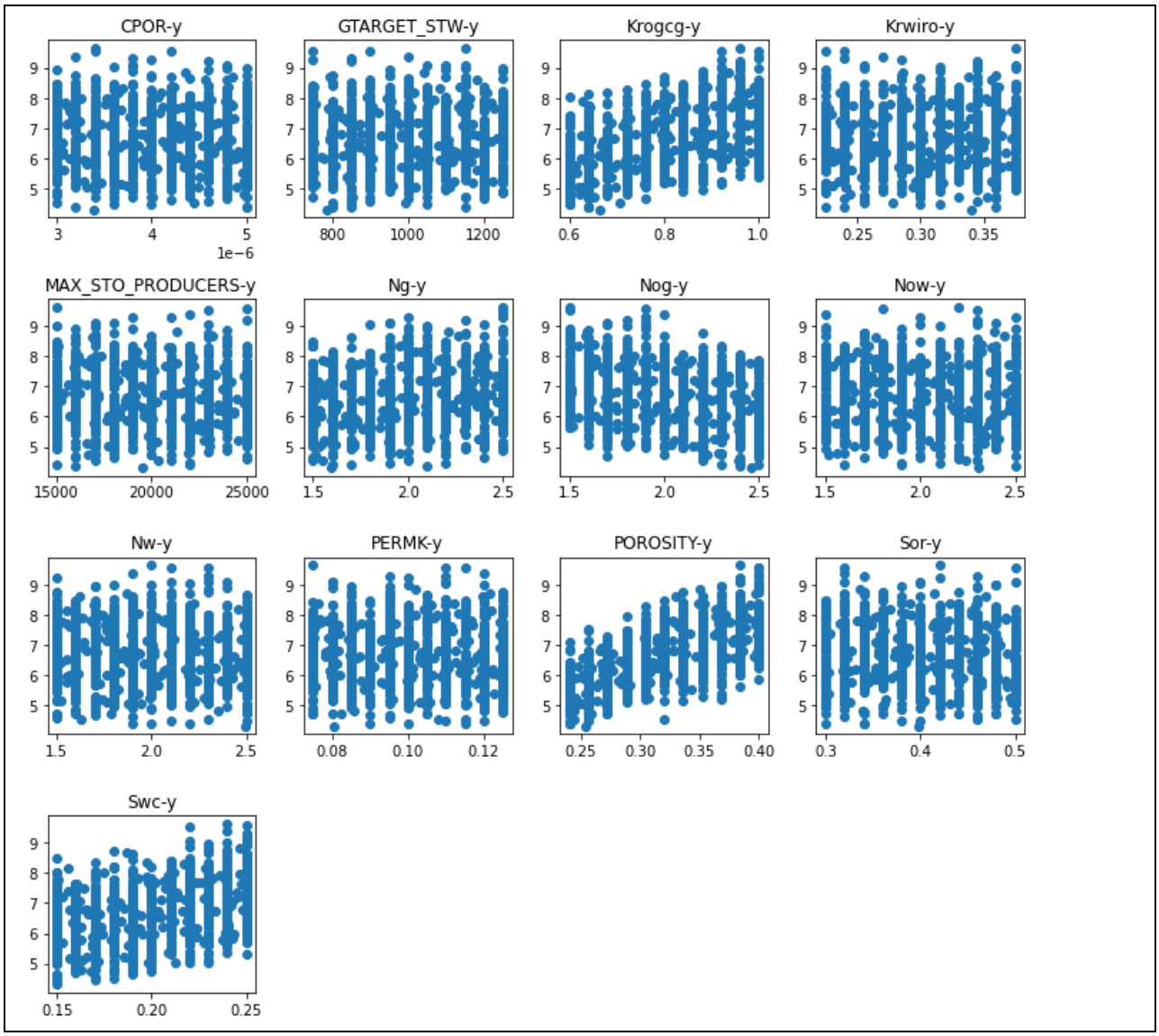


Figure 12

Cross-Plot of Predicted vs Actual for Linear Regression, Polynomial Regression and Random Forest Algorithms (Training Data vs. Test Data)

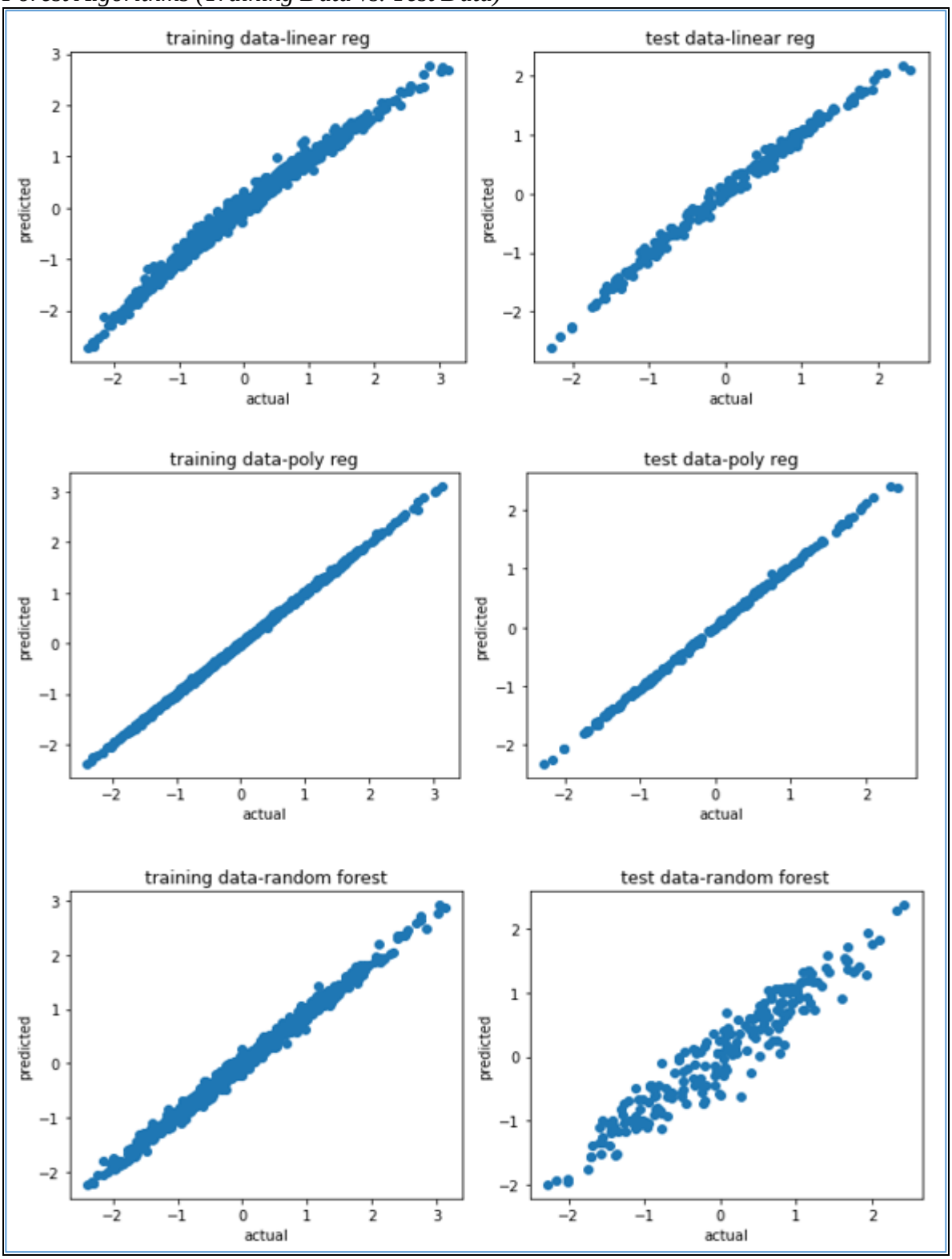


Figure 13
Sample Size vs. R2 For Linear Regression, Polynomial Regression and Random Forest Algorithms

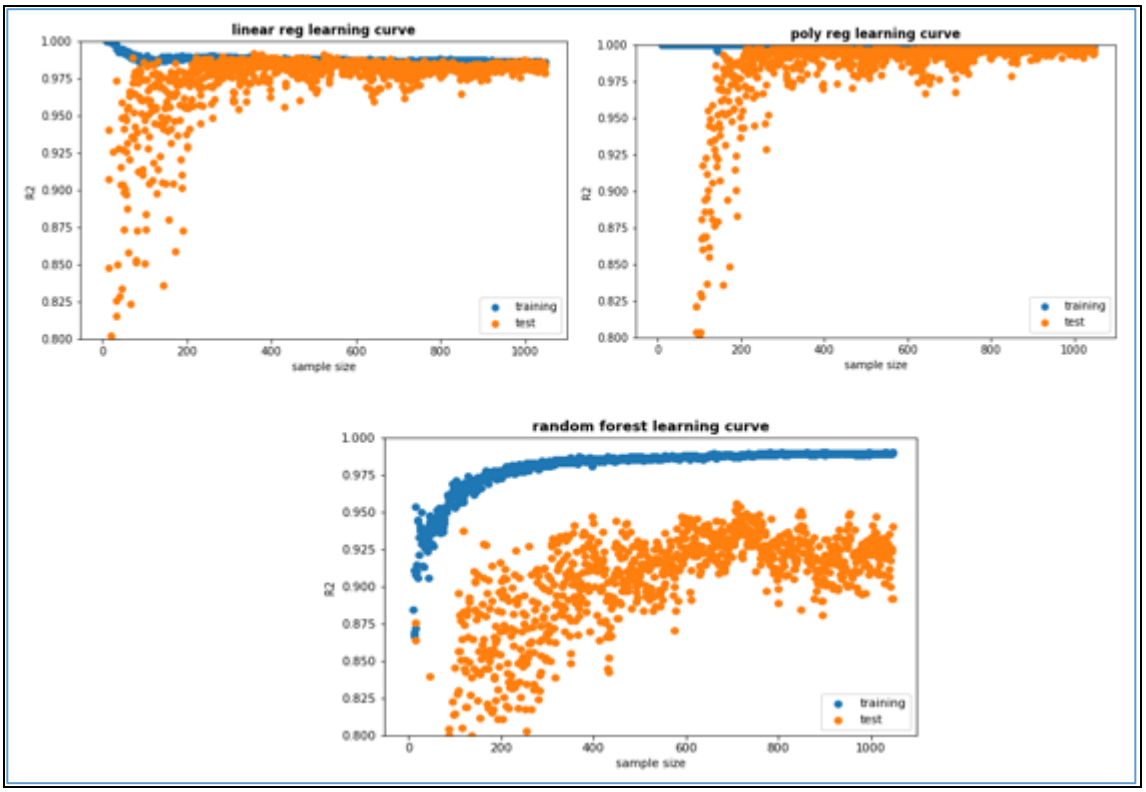
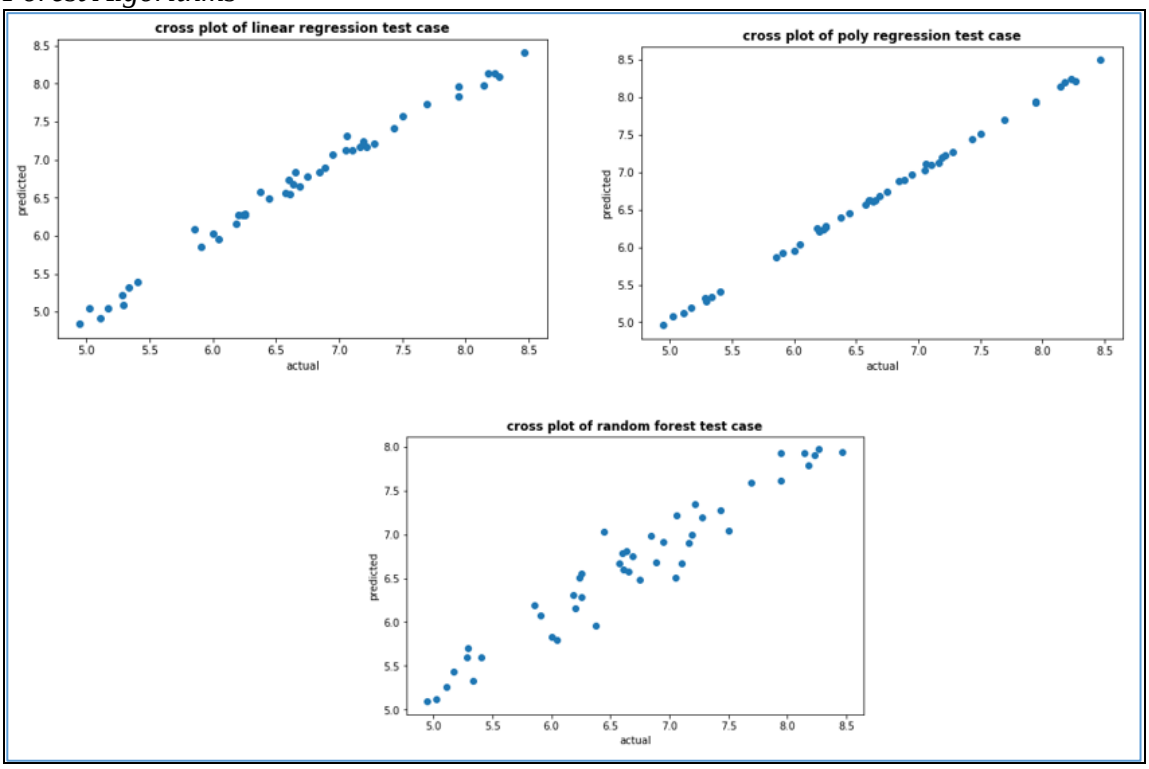


Figure 14

Cross-Plot of Predicted vs Actual for Linear Regression, Polynomial Regression and Random Forest Algorithms



Conclusion

This work demonstrates that machine learning algorithms can be used to develop predictive models for oil recovery when CO_2 is injected was development with a high level of accuracy. For the EOR33 project, utilizing over 1000 (1047) data points, the polynomial regression models were the most effective for predicting cumulative oil production in reservoirs where CO_2 is injected with and R^2 and MAPE of 0.999 and 0.0031 respectively. The linear regression algorithm is also capable of prediction with the R^2 and a MAPE of 0.989 and 0.0119 respectively. The Random Forest, like the polynomial regression, was able to predict using a non-linear prediction algorithm the cumulative oil production however in this case, the linear regression model gave better predictions. Further improvements can be made to the Random Forest algorithm to improve the accuracy of its prediction since this tool has the advantage of requiring less tuning and can give results relatively quick for larger datasets.

References

Adel, I., Tovar, F., Zhang, F., & Schechter, D. (2018). The impact of MMP on recovery factor during CO_2 – EOR in unconventional liquid reservoirs. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers. https://doi.org/10.2118/191752-MS

Albon, C. (2018). *Machine learning with python cookbook: Practical solutions from preprocessing to deep learning* (1st ed.). O'Reilly Media Inc.

Belazreg, L., Mahmood, S., & Aulia, A. (2020). Random Forest algorithm for CO_2 water alternating gas incremental recovery factor prediction. *International Journal of Advanced Science and Technology*, *29*(1), 168-188.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Gajbhiye, R. (2021). Improving CO_2-foam performance for EOR at reservoir condition. In *OMC Med Energy Conference and Exhibition*. Society of Petroleum Engineers. https://onepetro.org/OMCONF/proceedings/OMC21/All-OMC21/OMC-2021-224/473169

Gholami, V., Mohaghegh D. S., & Maysami, M. (2019). Smart proxy modeling of SACROC CO2-EOR. *Fluids, 4*(2), 85. https://doi.org/10.3390/fluids4020085

Hosseini, S., Alfi, M., Nicot, J., & Nuñez-Lopez, V. (2018). Analysis of CO_2 storage mechanisms at a CO_2-EOR site. *Greenhouse Gases: Science and Technology*, *8*(3), 469-482. https://doi.org/10.1002/ghg.1754

Iino, A., Onishi, T., & Datta-Gupta, A. (2019). Optimizing CO_2- and field-gas-injection EOR in unconventional reservoirs using the fast-marching method. *SPE Reservoir Evaluation & Engineering*, *23*(01), 261-281. https://doi.org/10.2118/190304-PA

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R* (1st ed.). Springer.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*, 255-260. https://doi.org/10.1126/science.aaa8415

Labastie, A. (2011). En route: Increasing recovery factors: A necessity. *Journal of Petroleum Technology*, *63*(08), 12-13. https://doi.org/10.2118/0811-0012-JPT

Krasnov, F., Glavnov, N., & Sitnikov, A. (2018). A machine learning approach to enhanced oil recovery prediction. International Conference on Analysis of images, social networks and texts. AIST 2017. *Lecture Notes in Computer Science*, 10716. Springer, Cham. https://doi.org/10.1007/978-3-319-73013-4_15

Li, H., Gong, C., Liu, S., Xu, J., & Imani, G. (2022). Machine learning-assisted prediction of oil production and CO_2 storage effect in CO_2-water-alternating-gas injection (CO_2-WAG). *Applied Sciences*, *12*(21) 10958. https://doi.org/10.3390/app122110958

Lowe, B., & Kulkarni, A. (2015). Multispectral image analysis using Random Forest. *International Journal of Soft Computing*, *6*(2), 1-14 https://airccse.org/journal/ijsc/papers/6115ijsc01.pdf

Lv, W., Tian, W., Yang, Y., Yang, J., Dong, Z., Zhou, Y., & Li, W. (2021). Method for potential evaluation and parameter optimization for CO_2-WAG in low permeability reservoirs based on machine learning. IOP Conference Series: Earth and Environmental Science, *651*, 032038. https://iopscience.iop.org/article/10.1088/1755-1315/651/3/032038

Medica, K., Maharaj, R., Alexander, D., & Soroush, M. (2020). Evaluation of an alkali-polymer flooding technique for enhanced oil recovery in Trinidad and Tobago. *Journal of Petroleum Exploration and Production Technology*, *10*(8), 3947-3959. https://doi.org/10.1007/s13202-020-00981-7

Ministry of Energy and Energy Industries. Historical facts on the petroleum industry of Trinidad and Tobago. http://www.energy.gov.tt/historical-facts-petroleum/

Ministry of Energy and Energy Industries. (2019, April). Trinidad & Tobago: MEEI.

Mohammed-Singh, L., & Singhal, A. (2004). Lessons from Trinidad's CO_2 immiscible pilot projects 1973-2003. In *PE/DOE Symposium on Improved Oil Recovery*. Society of Petroleum Engineers. https://doi.org/10.2118/89364-MS

Mohammed-Singh, L., & Singhal, A. (2005). Lessons from Trinidad's CO_2 immiscible pilot projects. *SPE Reservoir Evaluation & Engineering*, *8*(05), 397-403. https://doi.org/10.2118/89364-PA

Mohaghegh, Shahab D. (2011, April 19-21). Reservoir simulation and modeling based on pattern recognition. [Conference session]. SPE 2011 Digital Energy Conference and Exhibition, The Woodlands, Texas, USA. doi: https://doi.org/10.2118/143179-MS

Rafael Pires de Lima & Youzuo Lin. (2019). Geophysical data integration and machine learning for multi-target leakage estimation in geologic carbon sequestration. SEG Technical Program Expanded Abstracts: 2333-2337. https://doi.org/10.1190/segam2019-3215405.1

Russell, S., & Norvig, P. (2009). *Artificial intelligence* (3rd ed.). Pearson Education.

Saira, S., Yin, H., & Le-Hussain, F. (2020). Using alcohol-treated CO_2 to reduce miscibility pressure during CO_2 injection. In *SPE Asia Pacific Oil & Gas Conference and Exhibition*. Virtual: Society of Petroleum Engineers. https://doi.org/10.2118/202341-MS

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. https://doi.org/10.1147/rd.33.0210

Sharma, G. (2021). 5 regression algorithms you should know. https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/

Trading Economics. Trinidad and Tobago crude oil production. Retrieved from https://tradingeconomics.com/trinidad-and-tobago/crude-oil-production

Yan, Y., Borhani, T. N., Subraveti, S. G., Pai, K. N., Prasad, V., Rajendran, A., Nkulikiyinka, P., Asibor, J. O., Zhang , Z., Shao , D., Wang , L., Zhang , W., Yan , Y., Ampomah , W., You , J., Wang , M., Anthony , E., Manovic , V., & Clough , P. T. (2021). Harnessing the power of machine learning for carbon capture, utilisation, and storage (CCUS) – a state-of-the-art review. *Energy Environment. Science*, *14*, 6122-6157. https://doi.org/10.1039/D1EE02395K

Zeeshan, T., Murtada, A. S., Amjed, H., Mobeen, M., Emad, M., Ammar, E., Sulaiman, A., Mohamed, M., & Abdulazeez, A. (2021). A systematic review of data science and machine learning applications to the oil and gas industry. *Journal of Petroleum Exploration and Production Technology*, *11*, 4339–4374. https://doi.org/10.1007/s13202-021-01302-2