Reservoir Characterization in KG Deep Water using Data Analytics, Eastern Offshore Basin, India

Sanjai Kumar Singh, CS Bahuguna, Rajeev Tandon*
GEOPIC, ONGC, Dehradun, Uttarakhand, India

Abstract

Oil and gas industry is data driven industry with more than petabyte of seismic data, well data, reservoir and production data growing in size continuously. In current scenario there is an urgent need to deploy new technologies and approaches to integrate and interpret all these data to drive faster and accurate decisions leading to finding of new resources, increased recovery rates and reduced environmental impacts in a cost effective manner. In this study, data analytic approach of reservoir characterization was attempted in Eastern deep water offshore basin of India (Fig. 1), which is one of the most promising deep water block with several discoveries. Reservoirs in this area are sands within Godavari clay of Plio-Pleistocene age. These reservoirs are the slope channel sands which occur as high amplitude anomaly bursts within low to moderate value pertaining to encasing clay in the background. Capturing these high amplitude anomalies as geobodies bring out the geometry of NW-SE trending slope channel sands and the data from the drilled wells targeting them confirm that these high amplitude bursts to be reservoir facies. Most of these sand bodies are charged with hydrocarbons. However, as observed in some wells these high amplitude events, though correspond to reservoir sands, are not always hydrocarbon bearing. These surprises are a major challenge during exploration and development of field. To understand and overcome this challenge, a data analytic approach of reservoir characterization study was carried out. The data analytic approach for causal variables of seismic amplitude response and reservoir characterization using seismic were successfully implemented using open source tools. Statistical method was used for causal variable analysis. Causal variable analysis helped to understand the bright amplitude response in seismic. This also guided to choose the variable of importance from hydrocarbon exploration point of view. Out of several causal variables, water saturation is having significant control on seismic amplitude response. Water saturation and facies were selected as target variable for prediction using seismic and attributes. All the predictions at blind locations, recently drilled development wells along with a few exploratory wells, have corroborated the study outcome.

Introduction

Oil and gas industry is data driven industry with more than petabyte of seismic data, well data, reservoir data and production data which are growing continuously. These data are being generated through different phases of E&P business cycles of exploration, appraisal, development and production involving several disciplines like geophysics, geology, petro-physics, rock physics, reservoir engineering, drilling engineering and production. There is an urgent need to deploy new technologies and approaches to integrate and interpret all these data to drive faster and accurate decisions leading to finding of new resources, increased recovery rates and reduced environmental impacts in a cost effective manner. Big data analytics is an emerging technology which has been successfully implemented in other industries and walks of life. Since E&P industries have variety of voluminous data with varying scale and size, big data analytics technology is a good option to go far. In fact this data driven technology is already leading in some part of E&P business, like drilling optimization, production optimization etc. This approach of integrating all these data to derive faster and accurate decisions may definitely lead to finding of new resources, increased recovery rates and reduced environmental impacts in a cost effective manner.

The data analytical approach for reservoir characterization in KG deep water (Figure 1) was carried out which resulted in better delineation of reservoirs. Reservoirs in this area are sands within Godavari clay formation of Plio-Pleistocene age with heterogeneous distribution. These reservoirs are the slope-channel sands that occur as high amplitude anomaly bursts within low to moderate amplitude ranges pertaining to encasing clays in the background. Capturing these anomalies as geobodies bring out the geometry of NW-SE trending slope channel sands and these high amplitude bursts correspond to coarser clastics. However, these high amplitude seismic anomalies in some drilled wells have encountered reservoir sands devoid of hydrocarbons which poses major challenge in exploration and development in the area.

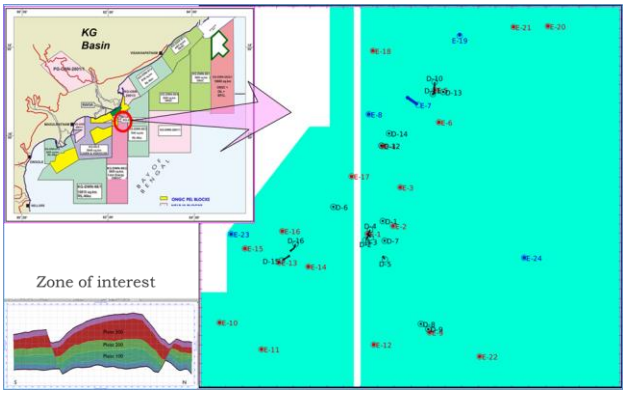


Figure 1: KG basin map showing area of study along with wells. Wells name starting with E and in red colour are used in the study. Remaining wells were kept blind.

INTEG, GEOPIC, ONGC, Dehradun
sks175@gmail.com

Reservoir Characterization in KG Deep Water using Data Analytics, Eastern Offshore Basin, India

Reservoir characterization study in this area had been performed earlier too and a number of property volumes had been generated. In the present study, it was decided to integrate all available data and derive more insight out of it, using analytical and machine learning approach. Major challenge in the area is to identify the causal variables of high amplitude burst in seismic data and to understand the seismic amplitude response of hydrocarbon charged reservoirs.

Causal Variable Analysis:

For causal variables of seismic amplitude response analysis, mean values of logs (Vp, Vs, RHOB, P-impedance, Vcl, Resistivity, NPHI, PHIT, PHIE, Sw, DT, S-impedance, GR, DTS & Vp/Vs) and RMS values of seismic (near to far stack and full stack) of brine as well as hydrocarbon bearing zones were extracted. Total 244 samples corresponding to brine and hydrocarbon bearing zone of 20 wells are used in the study. All these data were reshuffled or randomized to remove any bias due to data of any particular well and zone. After compiling data, exploratory data analysis (EDA) was done. EDA encompasses an iterative approach and enhances the process towards consistent data integration, data aggregation and data management. EDA is achieved by adopting a suite of visualization techniques from univariate, bivariate, and multivariate perspective. The main objective of EDA is to maximize insight into a dataset. The histogram is a basic EDA tool for displaying the frequency distribution of a set of data. A mound-shaped histogram may indicate that the data follow a normal distribution. Like histogram, the box plot is also a coarse summary of the data. It allows comprehending at a glance the specific important features of the data.

Figure 2 shows the data distribution of far stack amplitude corresponding to water and hydrocarbon bearing zones. It is clear from the figure (box & density plot in both cases of coloured by facies as well as layer) that data is positively skewed. This type of plot were generated for all the data under analysis. Boxplots and density plot readily convey an impression of the extent of variability or scatter in the data. They provide a visual check on the assumption, common in many uses of statistical models, that variability is constant across treatment groups. For example, in any multi attribute regression it is assumed that the input data distribution is normal. If this assumption is not fulfilled, the generated model will fail in prediction on unseen data.

Box and density plots are methods of univariate data analysis. Bivariate data analysis explores the relationship between the two data sets. It becomes very important to explore the relationship between several inputs to be fed into any statistical modeling process. If inputs are highly correlated with each other, they may not provide extra information in the modeling process. Rather highly correlated data make the generated model numerically unstable.

Figure 3 shows the correlation matrix plots. These plots reveal lots of insight about the data in terms of distribution and correlation between them. Diagonal plots in Figure 3 shows distribution (density plot) of individual variables, below diagonal show cross plot between the variables and above diagonal show the correlation coefficients.

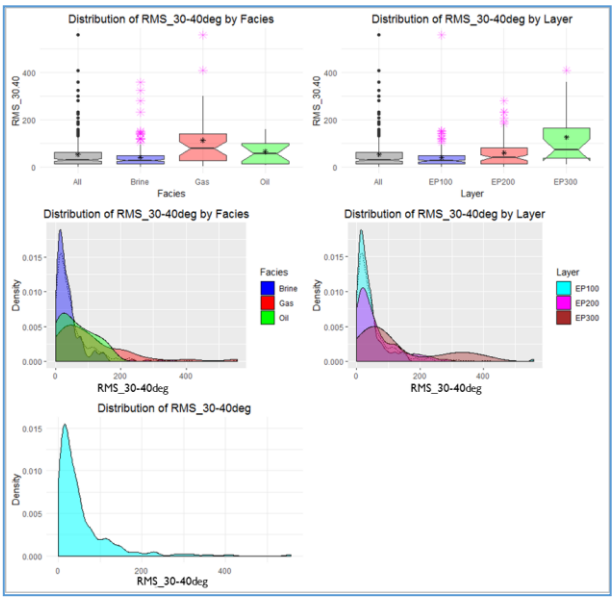


Figure 2: Box and density plot of far stack amplitude (30-40deg) coloured with facies (left) and layer wise (right). The amplitude distribution is positively skewed.

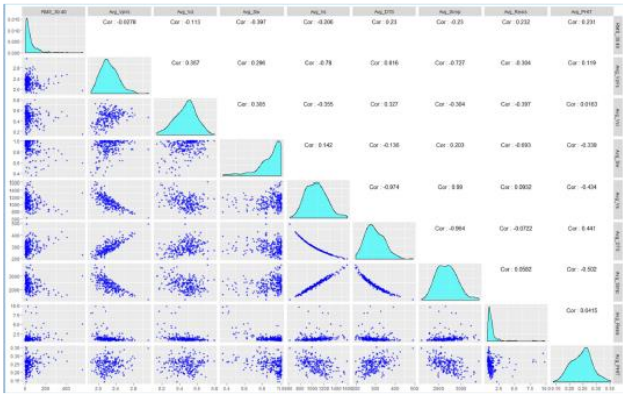


Figure 3: Data visualization & relationship between the data in terms of correlation

A correlation coefficient is a measure of the degree of relationship between two variables. It is usually a number between -1 and 1. The magnitude represents the degree of the correlation and the sign represents the trend of the correlation. A high degree of correlation (closer to 1 or -1) indicates that the two variables are very highly correlated, either positively or negatively. A high positive correlation indicates that observations with a high value for one variable will also tend to have a high value for the second variable. A high negative correlation indicates that observations with a high value for one variable will also tend to have a lower value of the second

Reservoir Characterization in KG Deep Water using Data Analytics, Eastern Offshore Basin, India

variable. Correlations of 1 (or -1) indicate that the two variables are essentially identical, except perhaps for scale (i.e., one variable is just a multiple of the other). Here a simple analysis of Pearson correlation of partial stacks amplitude with all available logs were done and shown in Figure 4 as correlation matrix. Brown to blue in correlation matrix show maximum negative (-1) to maximum positive (+1) correlation.

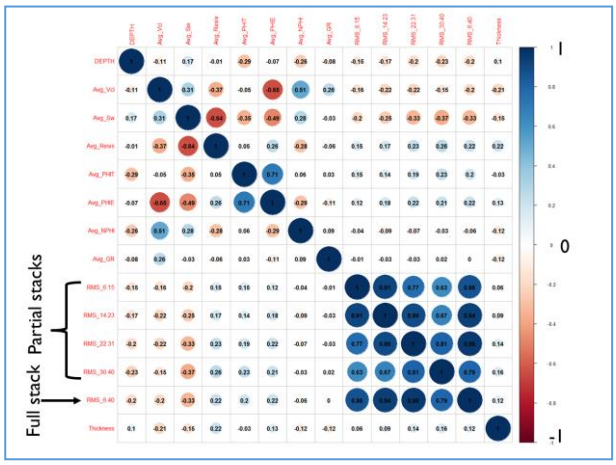


Figure 4: Correlation matrix showing correlation between logs and seismic amplitudes (near to far and full stack)

Figure 4 clearly shows that as the water saturation is decreasing, amplitude is increasing from near to far stack in seismic. The correlation values show significant changes in in a range of -0.2 to -0.37 for near to far stack. This indicates water saturation variation will have significant impact on seismic amplitude. Total porosity is having positive correlation with seismic amplitudes and showing increasing trend of correlation. Like total porosity, effective porosity is also having positive correlation with seismic amplitudes and showing increasing trend of correlation. Depth, Vcl and Sw are showing negative correlation, indicating that seismic amplitudes increase with decrease of these values, while resistivity, PHIT, PHIE and thickness are showing positive correlation. There is no significant correlation of seismic amplitudes with NPHI and GR as depicted in figure 4. Similarly correlation of amplitudes with elastic properties like P-impedance, Vp/Vs, S-impedance, and density are also analyzed. There is hardly any correlation between Vp/Vs and amplitudes but other properties like P-impedance, S-impedance and density (RHOB) show negative correlation.

Detail analysis in terms of correlation matrix, cross plot or scatter plot were made. There are several factors that are having control on seismic amplitudes response. Causal variable analysis of seismic amplitude response reveals several factors, like water saturation, depth & thickness, volume of clay, P-impedance, porosity and resistivity of reservoir interval, that are having control on seismic amplitudes (Figure 5). Several combinations of these causal variables may cause the same amplitude response in seismic. In such scenario, any decisions on the basis of amplitudes may only lead to uncertainty.

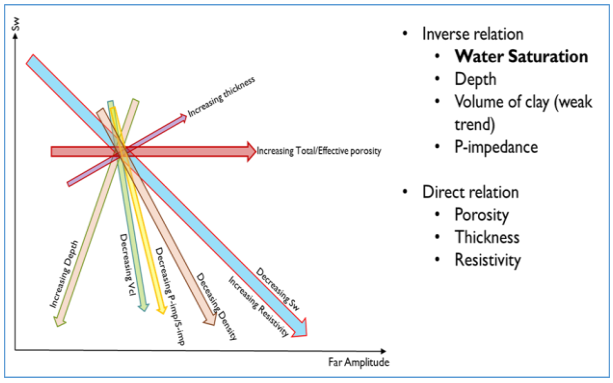


Figure 5: Causal variable of seismic amplitude response

Model Building & Validation:

The main approaches of analytics are statistical and artificial intelligence based. Statistical methods are well tested but in case of large and varied data it may have limitations. It will not be possible to decipher hidden pattern in the data though statistical methods. Artificial intelligence (AI) methods are good in finding hidden pattern in the data and is possible to train complex model. There is no limitation of handling data size and variety but it needs considerable computing resources. AI is a broad field that include machine learning (ML) and deep learning (DL). A machine-learning system is trained rather than explicitly programmed. Learning, in the context of machine learning, describes an automatic search process for better representations of the input data.

Different seismic attributes are generated to aid in the characterization objective. A brief list of seismic attributes and its benefits are given in table below.

Seismic Attributes	Information inferred
Instantaneous phase	Event continuity
RMS amplitude	Highlights geometry of importance
P-impedance	Lithology indicator
Vp/Vs	Fluid discriminator
Spectral decomposition	Infers tuning of geological feature of interest
Sweetness	Highlights presence of gas
AVO attribute volumes (Intercept, Gradient, Rp, Rs, Fluid factor, P*G etc)	Help to infer hydrocarbon presence (Gas)
Many more	

List of a few attributes along with help or information derived out of these, are given in the table. Even this list may be long. The point we want to emphasize is that the seismic data is same but several attributes are derived which provide different information about the geological features of interest. This means that the seismic data is manifold. Separate pieces of information are derived by applying some mathematical or geometrical transformation to the original seismic data. These

Reservoir Characterization in KG Deep Water using Data Analytics, Eastern Offshore Basin, India

information are derived and interpreted individually. In spite of all known mathematical or geometrical transformation of seismic data to derive seismic attributes for inferring useful information, there are chances that some hidden useful information in the data may still be left out. It is better to use data analytical framework and allow the machine to integrate and learn all these pieces of information from the data and find out the hidden useful information out of it as well as provide the realistic picture of geological feature in the sub surface. Individually these attributes may provide partial information but together they may supplement each other and also may get enhanced by the hidden pattern or information in the data that can be mined using machine learning/deep learning techniques. With this idea of seismic data manifold and machine/ deep learning based analytical framework, several volume of reservoir properties (like water saturation, porosity, facies and facies probabilities) volumes can be computed.

Out of several causal variables, water saturation is having significant control on seismic amplitude response. Water saturation and facies were selected as target variables for prediction using seismic and its attributes. The list of seismic attributes considered for modeling and property volume generation are given in Figure 6.

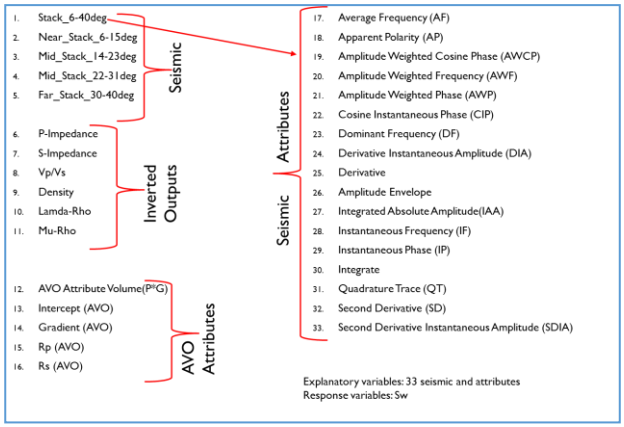


Figure 6: List of seismic attributes (33 nos.) considered as input for modeling and water saturation volume prediction.

All the data, seismic and seismic attributes as well as water saturation and facies at 20 wells, used in the study, were gathered. There are 72197 data points in total comprising of all 20 wells. The facies wise data distribution in zone of interest is shown in Fig. 7. Figure 7 shows the data contribution to be imbalanced; indicating maximum contribution from shale compared to that of brine and hydrocarbon. In such scenario, modeling is a challenging task as 80% accurate prediction may predict shale only. Cross validation with resampling method was adopted in the study to tackle such scenario.

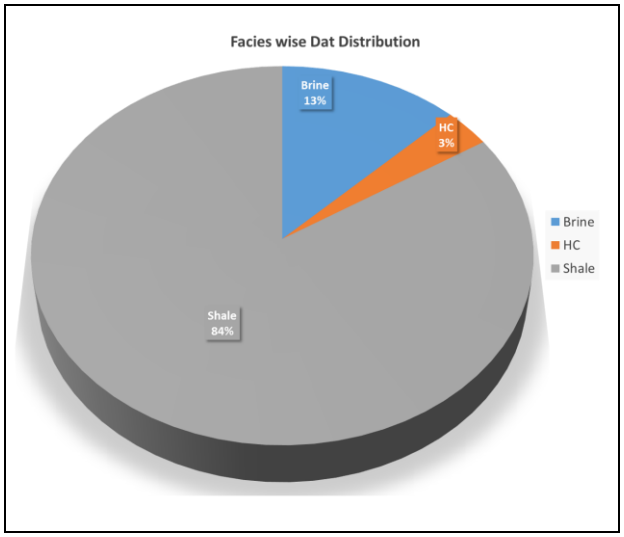


Figure 7: Facies wise data distribution in zone of interest shows imbalanced data contribution.

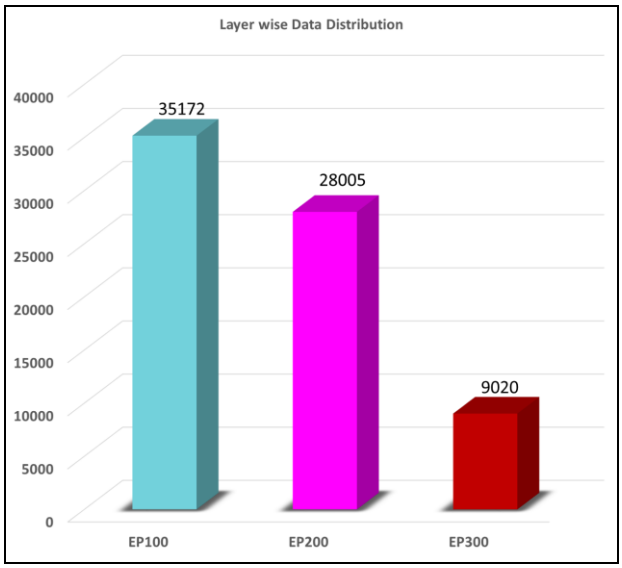


Figure 8: Data contribution details from different layers in zone of interest.

Figure 8 is showing data contribution layer wise in zone of interest. Maximum data contribution is from EP100 and EP200. EP300 is having least contribution of data. Prediction at unknown location on unseen data depends totally on data representation during model training. Interpretation of the results, where less data representation is there, should be done cautiously.

Reservoir Characterization in KG Deep Water using Data Analytics, Eastern Offshore Basin, India

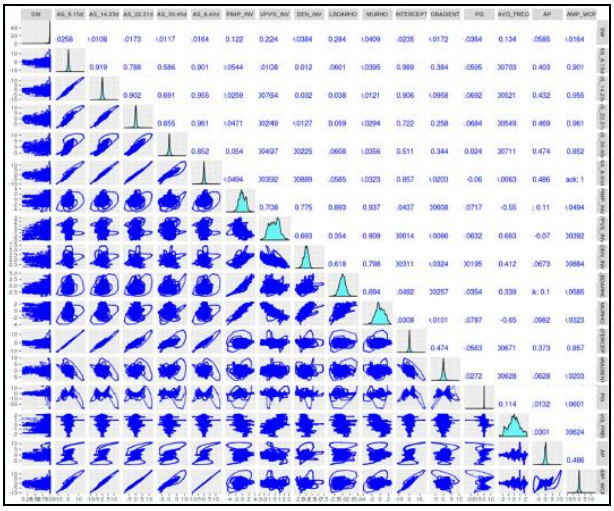
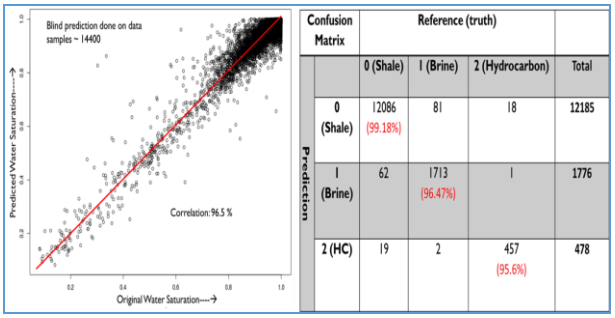


Figure 9: Distribution and correlation between some of 33 input variables

Figure 9 is showing the linear correlation matrix between the input variables. High correlation between the input variables indicates the redundant information contribution in model building. Most of the input variables are showing normal distribution (diagonal of Figure 9) but having different data value ranges. Some of the machine learning models require data to be on the same scale. But decision tree based models have no restriction of data scale and distribution but there is no harm if data are transformed to common scale. Before model building, all the aspect of data, like distribution, scale, outliers, were taken care.

Out of several methods attempted, deep learning method was implemented and predicted results were found to be very good even on unseen data with high accuracy. The predicted results like sand probability, hydrocarbon probability, brine sand probability, water saturation (hydrocarbon saturation) using deep learning methods are very good. Deep learning offers better performance on many problems. Deep learning also makes problem-solving much easier, because it completely automates what used to be the most crucial step in a machine-learning workflow that is feature engineering. Shallow machine learning techniques involve transforming the input data into one or two successive representation spaces, usually via simple transformations. But the refined representations required by complex problems, like in the present case, generally can't be attained by such techniques. Deep learning, on the other hand, completely automates this step. With deep learning, all features are learnt itself rather than having to be engineered by someone else. This has greatly simplified machine-learning workflows, often replacing sophisticated multistage pipelines with a single, simple, end-to-end deep learning model. It allows a model to learn all layers of representation, at the same time, rather than in succession. With joint feature learning, whenever the model adjusts one of its internal features, all other features that depend on it automatically adapt to the change, without requiring human intervention.

Water saturation, facies modeling and volume prediction were carried out using 33 input attributes using deep learning. Resampling techniques with 4 fold validation was implemented to overcome the issue of data imbalance. Out of 72197 data samples, 20 % (i.e. 14400 aprox.) samples were taken aside for blind testing of trained model. 80% data (i.e. 57797) samples were used to train the model by 4 fold validation technique. Before partitioning the data samples, data pre-processing steps were applied making data amenable for learning. With above data partitioning schemes, several scenario of deep model were tested with varying number of layers as well as varying number of neurons per layers. Optimum number of layers were decided by testing different deep learning scenario, keeping other parameters same. The correlation between actual and predicted water saturation on blind data (nearly 14400 samples) is nearly 96.5%, which is quite good (Figure 10). Hydrocarbon saturation was computed by subtracting water saturation from 1.



Confusion Matrix	Reference (truth)			
	0 (Shale)	1 (Brine)	2 (Hydrocarbon)	Total
0 (Shale)	12086 (99.18%)	81	18	12185
1 (Brine)	62	1713 (96.47%)	1	1776
2 (HC)	19	2	457 (95.6%)	478

Prediction

Figure 10: Prediction accuracy of water saturation (left) and facies (right) on blind data.

Similarly right table in Figure 10 shows the prediction accuracy of facies in terms of confusion matrix. The overall accuracy of facies prediction was 97 %. Facies probability volumes were also calculated. There was a very good validation of all these outputs at blind locations of recently drilled development wells in the area.

Results & Conclusions:

The data analytic approach for causal variable of seismic amplitude response and reservoir characterization using seismic were successfully implemented using open source tools. Statistical method was used for causal variable analysis. Causal variable analysis helped to understand the bright amplitude response in seismic. This also guided to choose the variable of importance from hydrocarbon exploration point of view. Out of several causal variables, water saturation is having significant control on seismic amplitude response .All the outputs were generated using 33 inputs of seismic and seismic derived attributes. Even more number of inputs can be integrated for model building and predictions. All the above predictions at blind locations, recently drilled development

wells along with a few exploratory wells, as indicated in map in Figure 1, have corroborated the study outcome as shown in Figure 11, 12 & 13.

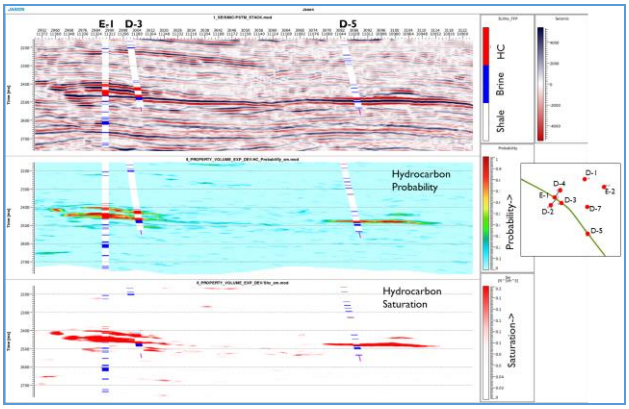


Figure 11: Sections showing predicted hydrocarbon probability and water saturation passing through blind wells D-3 & D-5.

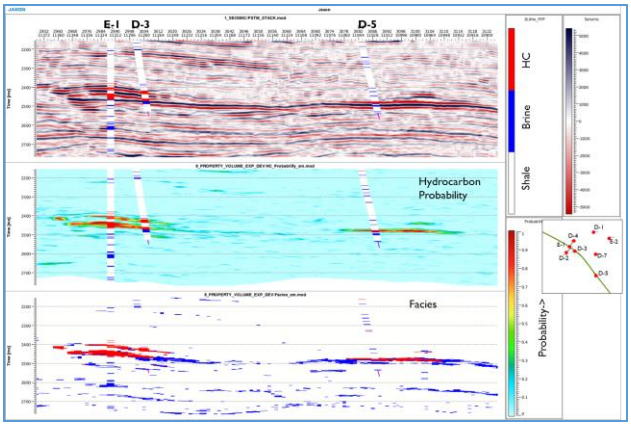


Figure 12: Sections showing predicted hydrocarbon probability and facies passing through blind wells D-3 & D-5

The data analytic approach of reservoir characterization through reservoir property predictions is an effective approach and it can decipher hidden pattern in the input data that is not possible manually through conventional approach.
Creating a system integrating all available inputs in a single platform and process with advanced machine learning methods may help better understanding of subsurface with faster and efficient decision making for exploration & development of hydrocarbon fields.

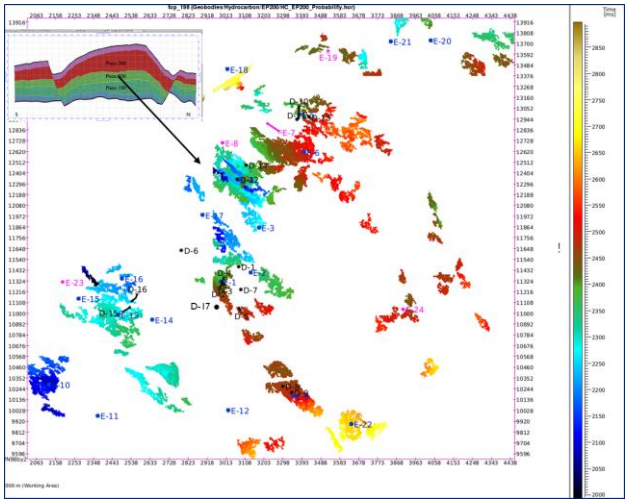


Figure 13: Hydrocarbon probability map for middle unit (EP200) validating the well observations at blind development wells with name starting with D and some exploratory locations.

References

1. Data Mining: Concepts and Techniques, Third Edition by Jiawei Han, Micheline Kamber & Jian Pei, The Morgan Kaufmann Series in Data Management Systems.

2. Data Analysis and Graphics Using R – an Example-Based Approach, Third Edition, by John Maindonald and W. John Braun, Cambridge University Press.

3. R machine learning essentials by Michele Usuelli, PACT Open source

4. Data Mining with Rattle and R by Graham Williams

5. Deep Learning with R by Francois Chollet & J.J Allaire

6. Applied predictive modelling by Max Kuhn and Kjell Johnson

Acknowledgements

This technical paper is a part of the project work carried out at INTEG, GEOPIC. The authors are thankful to ONGC Management for allowing to publish the paper in SPG-Kochi 2020 conference. We are also thankful to KG Basin interpretation group for providing geo-scientific data and technical support as and when required. We acknowledge the support from software and hardware group of GEOPIC.
The views expressed in this paper are solely those of the authors and need not necessarily be that of ONGC.