An Overview of Data Mining Applications in Oil and Gas Exploration: Structural Geology and Reservoir Property-Issues

Hamed Nikhalat Jahromi
LIAAD, INESC TEC
Portugal
hnj@inesctec.pt

Alpio M. Jorge
LIAAD, INESC TEC
Faculty of Science, University of Porto
Portugal
amjorge@fc.up.pt

Abstract

Low oil prices have motivated energy executives to look into cost reduction in their supply chains more seriously. To this end, a new technology that is experimentally considered in hydrocarbon exploration is data mining. There are two major categories of geoscientific problems in which data mining is applied: structural geology and reservoir property-issues. This research overviews these categories by considering a variety of interesting works in each of them. The result is an understanding of the specific geoscientific problems studied in the literature, along with the relative data mining methods. This way, this work tries to lay the ground for a mutual understanding on oil and gas exploration between the data miners and the geoscientists.

1 Introduction

There are two ways of increasing, or stabilizing, profit from a corporate finance perspective; either the revenues are increased or the costs are decreased. Given the fall in oil prices since late 2014, today, cost reduction is particularly the centre of attention for many executives in the energy industry.

Oil and gas supply chain consists of four main phases: extraction (including exploration), refinement, transport, and distribution/sale; for a discussion on these phases see Nikhalat-Jahromi et al. [NJC16]. In the context of cost reduction, exploration is the act of minimizing the expenses associated with finding commercial oil and gas deposits. These expenses rise from the following main exploratory activities (in the order of being carried out): satellite infrared and/or radar and/or microwave surveying; aerial imaging; geobotany prospecting and geochemical exploration; aerial magnetic and/or electromagnetic and/or gravity surveying; (surface) seismic surveying; and exploratory wells (generating well logs). Kearey et al. [KBH13] and Crum [Cru14] debate these activities thoroughly. The key is minimizing uncertainty, i.e., increasing the accuracy with discovering omitted knowledge, in the results generated from these activities. Note that more accurate results might decrease the number of needed activities and, therefore, aside from cost reduction, quicken the exploration and ultimately the oil extraction [Feb12].

Copyright © 2017 by the paper’s authors. This volume is published and copyrighted by its editors.

In: A. Jorge, G. Larrazabal, P. Guillen, R.L. Lopes (eds.): Proceedings of the Workshop on Data Mining for Oil and Gas (DM4OG), Houston, Texas, USA, April 29th, 2017 (arXiv:1705.03451).
Data mining is a new technology applied in oil and gas exploration that can create cost reductions and bring about considerable financial gains [Ami00]. It permits enhanced analysis and more effective decision-making by leveraging large volumes of data gathered in exploration [AS+12]. In other words, data mining’s potential lies in its ability to economically extract value from bulk volumes of a wide variety of data by enabling high-velocity discovery and analysis. Feblowitz [Feb12] states that (pp. 7-8):

“By applying advanced analytics, such as pattern recognition, to a more comprehensive set of data collected during seismic acquisition, geologists may be able to identify potentially productive seismic trace signatures that have been overlooked in newly acquired or archived data. Using data from other disciplines could enhance exploration efforts. For example, historical drilling . . . data from a nearby well could help geologists and geophysicists verify their assumptions in their analysis of a field. This becomes especially important where environmental regulations restrict new surveys.”

Baaziz and Quoniam [BQ15] state that application of data mining in oil and gas exploration is in the experimental stage with much of the efforts focused on data-intensive computing. Academic institutions, oil and gas companies, and business analytics service providers, are working on various applications.

This paper provides an up-to-date overview of data mining applications in oil and gas exploration, and tries to create a common understanding between the geoscientists and data miners. It considers the two major categories of geoscientific problems in which data mining is applied and, however briefly, reviews a set of interesting works in each of them. The categories are: structural geology and reservoir property-issues. In the former, data mining assists in creating and understanding of accurate images of complicated geological structures, in the reservoir and beyond, in time and depth by exploiting attributes of seismic surveying data (seismic attributes); while in the latter, this technology helps in predicting reservoir characteristics such as lithology and permeability using both seismic attributes and well logs.

The rest of this paper is organized as follows: Sections 2-3, respectively, cover the structural geology and reservoir property-issues. Each section reviews the relative works chronologically and begins with the earliest research, but groups with that research all the later works which address the same geoscientific problem. Section 4 concludes the paper and makes a few suggestions for further research.

2 Structural Geology

In oil and gas exploration, accurately mapping the top, bottom and lateral extent of geological structures is important as these structures might yield hydrocarbon. This positioning, in presence of oil and gas, assists in determining the thickness of the hydrocarbon bearing (rock) layers and consequentially in deciding the economic viability of the reserve. Furthermore, using accurate mapping, by the optimal placement of the wells, production in multi-layered reservoirs, where shale and sand layers are laminated, can be carried out from and combined into a single well, and, therefore, result in cost savings. To the knowledge of the authors, Sriram et al. [SSK75] carried out the first application of intelligent analysis techniques in oil and gas exploration. They used non-linear mapping and a non-parametric scheme to estimate the probability density function that can produce a distribution of clusters in the data. Sriram et al. [SSK75] use their techniques for mapping laminated sand and shale sequences based on seismic data for a reservoir in the Gulf of Mexico. Yin et al. [YWY96] trained a backpropagation (BP) neural network regressor for predicting the thickness of an oil deposit in China. They correlate 10 seismic attributes, such as wavelet and Hilbert Transformation, and the known thicknesses from the exploratory wells. For introduction to and discussion on various seismic attributes see Brown [Bro96].

Later, Yu et al. [YZD08] propose a dynamic all parameters adaptive BP neural network regressor for predicting the deposit's thickness by fusing genetic algorithms (GAs), simulated annealing (SA), and the BP neural network. The fusion offsets the disadvantages of one technique by the merits of the other and eventually results in a more robust neural network. Yu et al. [YZD08] use a GA for designing the network structure. Next, a hybrid SA-BP is implemented for finding the optimal weights and biases of the network. These two steps are repeated recursively for various seismic attribute combinations of the training data set until the optimal network is obtained. The method is applied to an oil reservoir in Central China. The optimal network is made of six seismic attributes, out of 50 available ones, that are correlated to the known reservoir thickness from 15 exploratory wells. Comparison with other methods in the literature, such a GA-SA algorithm, proves the superiority of Yu et al.’s [YZD08] on this data, as it generates a smaller mean squared error (MSE) in evaluation using the test data set.

Since faults have great significances for the storage and migration of hydrocarbon resources, fault detection is a central component of petroleum exploration. Thus, accurately imaging faults’ displacements, and determining their upthrown (the side of the fault in which rocks move upward) and downthrown sides is important. Meldahl
et al. present a key method for fault detection. It comprises an iterative processing workflow using directive seismic attributes (attributes steered in a user-driven, or data-driven, direction) a BP neural network, and image processing techniques. The network transforms all the attributes into one new attribute which indicates the probability of the fault at that position studied; it is trained on these attributes at example fault and non-fault positions. The trained network is applied to the entire target seismic area to produce a new section with high probability values at places where fault is present. Knowledge of fault shape and orientation can be fed back to the process either to increase the detection strength or to increase the resolution of the highlighted faults. In an example, Meldahl et al. fed 22 seismic attributes to a network. These were the reference time, dip, several dip-variance, energy, similarity, and amplitude calculations. The network made an accurate classification of the entire 3-dimensional (D) seismic section into fault and non-fault. Next, Meldahl et al. applied image enhancement filters and special purpose extraction algorithms to extract fault planes in a user-controlled way. Tingdahl and Rooji demonstrate the added value, in terms of fault continuity and contrast, of the combination of seismic attributes rather than a single attribute in Meldahl et al.’s method. Zheng et al. apply this method successfully to a reservoir with a complex fault system in south Texas. Results reconfirm the method’s ability in suppressing the surrounding noise and highlighting the faults. Meldahl et al.’s method is instrumental in both prospect evaluations (determining whether an area of exploration contains hydrocarbon in economic quantity) and research projects. In industry, it is used as a risk reduction tool and for ranking the prospects.

3 Reservoir Property-Issues

The earliest discoveries of oil and gas deposits were based on finding structural traps. But soon a different type of trap became important, stratigraphic. Marr defines the difference between the two as:

“Structural trap is one where … [impermeable] rocks have been folded and possibly faulted to produce a … closure … that is capable of trapping hydrocarbons in [below] porous and [permeable layers] …”, while stratigraphic traps accumulate oil due to changes of rock character rather than faulting or folding; (permeability is the character of rock that is an indication of the ability for fluids to flow through it; porosity is the percentage of void space in the rock; and the term stratigraphy refers to the study of rocks and their variations). For further discussion on traps see Marr.

In finding stratigraphic traps one must go beyond structure and deduce possible lithologies (grossly meaning identifying rock layers in the subsurface) and probable presence of oil and gas. Horizontal lithologic analysis is known as lithofacies analysis. Justice et al. provided system-supported human interaction with access to automated pattern recognition procedures for comprehending rock variations. They cluster the points of the seismic section, considering the seismic attributes, resulting in groups of lithofacies in the section. ITERATE, a technique that sorts data and uses an iterative redistribution operator in making clusters, was proposed by Biswas et al. as a method for clustering lithofacies in seismic sections. To quantify the quality of the clusters generated, two probabilistic measures are introduced: cohesion (intra-cluster similarity), and distinctness (inter-class dissimilarity). Cohesion is measured as the increased predictability of each feature value of the objects in the data set given the assigned cluster structure. Distinctness is quantified using a probability match measure. Unlike deterministic evaluation measure like MSE, measures in Biswas et al.’s (1998) work are probabilistic, since ITERATE’s cluster definitions are probabilistic.

Later, Chandra et al. presented an application of unsupervised self-organizing maps (SOM) neural networks in lithofacies analysis. Their clustering of the lithofacies for a reservoir in India is carried out in two steps: first, the shape of the traces within the horizontal interval of the 3-D seismic section of the field studied are analysed by an SOM which constructs a series of model traces that best represent the diversity of the shapes observed. Once these model traces are available, in the second step, each trace in the interval is compared to all the model traces and is assigned to the one with which it scores the best correlation. The result is a coloured seismic section with similar lithofacies coloured alike. Olorunsola et al. use the Generative Topographic Mapping (GTM) technique for clustering lithofacies of a complex granite wash formation in Texas Panhandle, United States. GTM, a non-linear projection technique, provides a probabilistic representation of the data-vectors in a corresponding lower latent dimensional space. The features that describe the data of the 3-D seismic section in Olorunsola et al.’s method are eight geometric, textural, and inversion seismic attributes: coherence, coherent energy, curvature, peak frequency, gray-level co-occurrence matrices (GLCM) entropy and heterogeneity, p-impedance, and reflector convergence.

Finally, two clustering techniques were used by Moraes et al. for vertical lithologic analysis in a
Brazilian reservoir: fuzzy c-means, and SOM. The authors extract three attributes (amplitude, cosine phase, and Hilbert Transformation) from the points of the 3-D seismic section of the field. Both algorithms were run for 2-10 partitions to identify the best number of vertical lithologic groups in the seismic section and found three as the optimal value. The quality of the groups made is assessed using PBM (acronym constituted of the initials of the names of its inventors Pakhira, Bandyopadhyay and Maulik), a cluster validation index for discussion on PBM see Pakhira et al. [PBM04]. The values of the PBM were largely similar for the fuzzy c-means and SOM techniques for each group number (2-10) tested, indicating that on this data these techniques have comparable quality.

Xiao and Zhu [XZ90] analysed the fuzzy relations of, collectively: the amplitude, phase, frequency, structure curvature, interval velocity, apparent polarity, and the thickness of low velocity layer; with the nature of the hydrocarbon deposit. They established the membership functions of these seismic attributes, classifying high-, low-, and non-productive reserves, thus properly placing the wells in the field and minimizing the number of dry (without any hydrocarbon recovery) and low-productive ones. Yao et al. [YLZ04] propose the feature expansion and selection in classification of high-, low-, and non-productive deposits. It is applied in two hydrocarbon prospects in China. Here, explicit nonlinear functions such as spline, radial basis function (RBF), and polynomial are used for feature expansion. Features are extracted from the seismic sections of the prospects. Expanded features tend to be more linearly related with the classification result intended. Next, a linear support vector machine (SVM) is used for feature selection where useless attributes are removed. An algorithm for recursive feature elimination (RFE) has been developed which trains the linear SVM and computes the weights for different attributes. The features with smaller weights, i.e., those contributing less information, are removed and the process is repeated until no useless attribute remains. Finally, a linear SVM is trained with selected seismic features that classifies the deposits.

Permeability estimation in exploratory wells is an important issue in determining the production potential of the deposit and therefore its economic feasibility. Mohaghegh et al. [MAA+94] proposed a method for assessing permeability by analysing three well logs, including gamma ray, bulk density, and deep induction log response. For introduction to and debate on well logs see Boyer and Mari [BM97]. They trained, using permeability of core samples from several exploratory wells of a hydrocarbon deposit in West Virginia, a BP neural network regressor. But before that a general regression neural network is used for identifying the optimum network design to be used in the BP network. Wong et al. [WFOG05] modelled estimation of permeability based on eight well logs (gamma ray, deep resistivity, shallow resistivity, flushed zone resistivity, bulk density, neutron porosity, photoelectric factor, and sonic travel time) as an SVM regression problem; permeability and the relative well logs are available for a few hundred data points for an exploratory well in a reservoir offshore Western Australia. The regression problem turns to the optimization of a convex formulation. Wong et al. [WFOG05] optimized this formulation by applying Lagrange multipliers. The SVM regressor performs better than Mohaghegh et al.’s [MAA+94] method in this case in that it generates a lower MSE in evaluation.

In an interesting project, the Correlations Company [Web01] was commissioned by the US government to perform data mining and determine the effectiveness of water injection in enhancing hydrocarbon recovery in an oil well in Nebraska. The effectiveness was defined as the primary to secondary (after water injection) oil extraction (P/S) ratio. The company used fuzzy logic for ranking and determining the important data attributes, out of 30 available ones from previous water injection experiences in other wells in the same region, in estimating the ratio. Six attributes were identified as central: water to oil ratio in the crude oil extracted from the well, gas to oil ratio, acreage dedicated to the well in the region, average porosity in the deposit beneath, average permeability, and the initial well pressure. A neural network regressor was trained on the correlation between these six and the P/S ratio of the previous water injection experiences. Next this network was used for prediction of the injection result in the well under question.

Estimation of shear wave velocity in wells is a key subject which is necessary in petrophysical studies of oil and gas deposits; (shear wave is a seismic body wave that shakes the ground perpendicular to the direction the wave is moving, and petrophysics is the study of the physical and chemical properties of rocks and their contained fluids). For many wells this velocity is not measured appropriately due to broken instruments, bad well conditions, and difficulty in accessing the equipment; or further not measured at all for cost reasons. Eskandari et al. [ERM04] suggest two methods for determining the shear wave velocity in wells from well logs: a multi regression method and a BP neural network. The authors had access to the shear wave velocity of core samples for a carbonate oil field in the Zagros Basin, South of Iran. Using the well logs of the carbonate oil field; compressional wave velocity, bulk density, and neutron porosity; and; neutron porosity, bulk density, transit time, gamma ray, deep laterolog, and X and Y coordinates; respectively, for the multi regression and the neural network methods, the regression
is carried out and the network is trained. Evaluation shows that the neural network here predicts the shear wave velocity slightly better, i.e., generates a better correlation coefficient than the multi regression method. Later, a neural network with Bayesian regularization (NNBR) along with stratigraphic constraints was recommended by Srisutthiyakorn [Sri12] for prediction of the shear wave velocity in wells from well logs. Shear wave velocity is available for many data points from nearly 30 exploratory wells in an oil prospect onshore Thailand. The author trains the NNBR with three well logs (gamma ray, neutron porosity, and resistivity) as attributes. In comparison to the conventional BP neural networks, NNBR has a smaller weight and bias leading to a smoother output. Srisutthiyakorn [Sri12] imposes stratigraphic constraints by separating the data into geologic formations so the NNBR is trained and predicts velocities that share the same geologic settings. This results in faster training times. The NNBR outperforms Eskandari et al.'s [ERM04] multi regression method in the Thai prospect.

4 Conclusions and Further Research

In this paper an overview of data mining applications in oil and gas exploration was carried out. Two major categories of geoscientific problems, which have been the focus of data mining efforts, are studied: structural geology and reservoir property-issues. Table 1 presents these categories and provides the key information deduced on them.

| Category                 | Problem                              | General Data Mining Method | Type of Attributes | Algorithm(s)                                |
|--------------------------|--------------------------------------|---------------------------|--------------------|---------------------------------------------|
| Structural Geology       | Mapping, Laminated Sand and Shale Sequences | Clustering                | Seismic            | Non-Linear Mapping and a Non-Parametric Scheme |
|                          | Predicting Reservoir’s Thickness      | Regression                |                    | BP Neural Network                          |
|                          | Fault Detection                      | Classification           |                    |                                             |
| Reservoir Property-Issues| Lithologic Analysis                  | Clustering                |                    | ITERATE; SOM Neural Network; GTM; C-Means   |
|                          | Predicting Hydrocarbon Deposits       | Classification           |                    | SVM                                         |
|                          | Determining the Permeability in Wells | Regression                |                    | BP Neural Network; SVM                       |
|                          | Estimating the Secondary Recovery Ratio |                           |                    | Neural Network                             |
|                          | Predicting the Shear Wave Velocity in Wells |                           |                    | Multi Regression; BP Neural Network; NNBR   |

* In case of "Estimating the Secondary Recovery Ratio", in addition to well logs, a few uncommon attributes are used as well. For details see Section 3.

There are two ways of expanding this paper:

(i) There is another category of geoscientific problems in which data mining is applied. Here, an array of ancillary data which provide high-level information in hydrocarbon exploration are covered. Remote sensing (including satellite infrared, radar, and microwave surveying; and aerial magnetic, electromagnetic, and gravity surveying) is the most important topic in this category. Although there are cases of data mining application to geobotany prospecting and geochemical exploration as well. In remote sensing, swaths of earth are covered, resulting in an understanding of the geological megastructures, with the aim of narrowing down to locations with the potential for hydrocarbon. After that further surveying in these places provides precise details on the subsurface formations. This category of problems, although less popular in academia than those two considered in this paper, is interesting to be overviewed too.

(ii) This overview given its novel way of categorizing data mining applications in oil and gas exploration is worthy of being expanded to a comprehensive review. Such a review will be informative to both geoscientists and data miners and deepen their mutual understanding.
References

[Ami00] Fred Aminzadeh. Challenges direct future of geophysics. *American Oil and Gas Reporter (Invited Paper for the Special Millennium Edition)*, 43(1):123–132, 2000.

[AS+12] Ahmed Abou-Sayed et al. Data mining applications in the oil and gas industry. *Journal of Petroleum Technology*, 64(10):88–95, 2012.

[BM97] Sylvain Boyer and Jean-Luc Mari. *Seismic surveying and well logging*. Technip Editions, 1997.

[BQ15] Abdelkader Baaziz and Luc Quoniam. How to use big data technologies to optimize operations in upstream petroleum industry. 2015.

[Bro96] Alistair R Brown. Seismic attributes and their classification. *The leading edge*, 15(10):1090–1090, 1996.

[BWF98] Gautam Biswas, Jerry B Weinberg, and Douglas H Fisher. Iterate: A conceptual clustering algorithm for data mining. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(2):219–230, 1998.

[Cru14] Shannon Crum. Aerial photography and remote sensing, 2014.

[CSS03] Mahesh Chandra, AK Srivastava, and V Singh. Lithostratigraphic interpretation of seismic data for reservoir characterization. In *SEG Technical Program Expanded Abstracts 2003*, pages 1473–1476. Society of Exploration Geophysicists, 2003.

[ERM04] H Eskandari, MR Rezaee, and M Mohammadnia. Application of multiple regression and artificial neural network techniques to predict shear wave velocity from wireline log data for a carbonate reservoir south-west iran. *CSEG recorder*, 42:48, 2004.

[Feb12] Jill Feblowitz. The big deal about big data in upstream oil and gas. *Paper & presentation, IDC Energy Insights*, 2012.

[JHW85] JH Justice, DJ Hawkins, and G Wong. Multidimensional attribute analysis and pattern recognition for seismic interpretation. *Pattern Recognition*, 18(6):391–399, 1985.

[KBH13] Philip Kearey, Michael Brooks, and Ian Hill. *An introduction to geophysical exploration*. John Wiley & Sons, 2013.

[MAA+94] Shahab Mohaghegh, Reza Arefi, Samuel Ameri, MH Hefner, et al. A methodological approach for reservoir heterogeneity characterization using artificial neural networks. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 1994.

[Mar71] John D Marr. Seismic stratigraphic exploration - part i. *Geophysics*, 36(2):311–329, 1971.

[MEEE06] DRS Moraes, RP Espindola, AG Evsukoff, and NFF Ebecken. Cluster analysis of 3d seismic data for oil and gas exploration. *WIT Transactions on Information and Communication Technologies*, 37, 2006.

[MHBdG01] Paul Meldahl, Roar Heggland, Bert Bril, and Paul de Groot. Identifying faults and gas chimneys using multitudinaries and neural networks. *The Leading Edge*, 20(5):474–482, 2001.

[NJFC16] Hamed Nikhalat-Jahromi, Dalila BMM Fontes, and Robert A Cochrane. Future liquefied natural gas business structure: a review and comparison of oil and liquefied natural gas sectors. *Wiley Interdisciplinary Reviews: Energy and Environment*, 2016.

[OQI+16] Oluwatobi Olorunsola, Jie Qi, Lennon Infante, Bryce Hutchinson, and Kurt Marfurt. Multiaattribute seismic-facies expressions of a complex granite wash formation: A buffalo wallow field illustration. In *SEG Technical Program Expanded Abstracts 2016*, pages 1884–1888. Society of Exploration Geophysicists, 2016.
[PBM04] Malay K Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. Validity index for crisp and fuzzy clusters. Pattern Recognition, 37(3):487–501, 2004.

[Sri12] Nattavadee Srisutthiyakorn. Redefining the standard of missing log prediction: Neural network with bayesian regularization (nnbr) with stratigraphic constraint—a case study from laminated sand-shale reservoir. In SEG Technical Program Expanded Abstracts 2012, pages 1–5. Society of Exploration Geophysicists, 2012.

[SSK75] KP Sriram, ET Stoessel, and BR Kowalski. Pattern recognition in hydrocarbon exploration. In Decision and Control including the 14th Symposium on Adaptive Processes, 1975 IEEE Conference on, volume 14, pages 118–119. IEEE, 1975.

[TDR05] Kristofer M Tingdahl and Matthijs De Rooij. Semi-automatic detection of faults in 3d seismic data. Geophysical Prospecting, 53(4):533–542, 2005.

[Web01] James R Weber. Data mining at the nebraska oil & gas commission. Technical report, Coral Production Corporation (US), 2001.

[WFOG05] Kok Wai Wong, Chun Che Fung, Yew Soon Ong, and Tamas D Gedeon. Reservoir characterization using support vector machines. In Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on, volume 2, pages 354–359. IEEE, 2005.

[XZ90] Ciyuan Xiao and Baiwen Zhu. A fuzzy mathematical method for predicting hydrocarbon accumulation area by comprehensively analysing multiple kinds of seismic information [j]. Oil Geophysical Prospecting, 2007, 1990.

[YLZ+04] Kaifeng Yao, Wenkai Lu, Shanwen Zhang, Huanqin Xiao, and Yanda Li. Hydrocarbon reservoir prediction using support vector machines. In International Symposium on Neural Networks, pages 537–542. Springer, 2004.

[YWY96] Xing-yao Yin, Guo-chen Wu, and Feng-li Yang. Predicting oil and gas reservoir and calculating thickness of reservoir from seismic data using neural network. In The 1996 3 rd International Conference on Signal Processing, ICSP’96. Part 2(of 2), pages 1601–1604, 1996.

[YZD08] Shiwei Yu, Kejun Zhu, and Fengqin Diao. A dynamic all parameters adaptive bp neural networks model and its application on oil reservoir prediction. Applied mathematics and computation, 195(1):66–75, 2008.

[ZKD14] Zhi Hong Zheng, Payam Kavousi, and Hai Bin Di. Multi-attributes and neural network-based fault detection in 3d seismic interpretation. In Advanced Materials Research, volume 838, pages 1497–1502. Trans Tech Publ, 2014.