Artificial neural networks for predicting petroleum quality

Sandro da Silva Camargo

Universidade Federal do Pampa. Rua Travessa 45, 1650, 96413-170, Bagé, RS, Brasil.
sandro.camargo@unipampa.edu.br

Paulo Martins Engel

Universidade Federal do Rio Grande do Sul. Av. Bento Gonçalves, 9500, 91501-970, Porto Alegre, RS, Brasil.
engel@inf.ufrgs.br

Abstract: Due to limited understanding of many diagenetic processes which contributes to petroleum quality determination, mathematical models become a very useful tool to improve understanding of these processes and to improve reservoir quality predictions prior drilling. Especially for reservoir engineers and petrophysicists the distribution of porosity and permeability are very important in the formation evaluation and definition of recovery strategies and evaluation of reservoir quality. In this context, we have developed an artificial neural network based model to predict macroporosity of sandstones reservoir systems. We have used a score to quantify the importance of each feature in prediction process. This score allows creating progressive enhancement neural models, which are simpler and more accurate than conventional neural network models and multiple regressions. The main contribution of this paper is the building of a reduced model just with the most relevant features to macroporosity prediction. A dataset, containing petrographic and petrophysical characteristics, containing samples of the same formation sandstone reservoir was investigated. Study results show that progressive enhancement neural network is able to predict macroporosity with accuracy near 90%, suggesting that this technique is a valuable tool for reservoir quality prediction.

Key words: progressive enhancement neural model, reservoir quality prediction, macroporosity prediction.

Introduction

Nowadays, more than 85% of world energy consumption comes from fossil fuels and petroleum is dominant among them (International Energy Agency, 2010). This dominance has created a world's economic dependence on petroleum production. Furthermore, petroleum companies are among the biggest corporations in the world and have formed a key part of the global economy. However, fossil fuels are a limited resource, and reservoirs are consumed more rapidly each year. More accessible reservoirs, which exploration is low cost, are almost vanished.

Besides, remaining reservoirs are increasingly more technically difficult to extract and, therefore, more expensive. Later reservoirs will only be economically feasible to extract at extremely high costs. In this context, exploration costs are a very important variable in reservoir exploration decision. In order to know the costs of field exploration, and increase exploration success rates, oil companies must try to predict the quality of reservoirs. Accurate prediction of reservoir quality is, and will continue to be, a key challenge for hydrocarbon exploration and development (Kupecz *et al.*, 1997).

Due to limited understanding of the details of many diagenetic processes, there is a lack for new techniques and tools to support quality predictions. Despite its notable economic importance, relatively few papers illustrate the application of studies to reservoir quality prediction. The main difficulty to execute this task is that the creation of models to reservoir quality prediction is highly dependent on quality

and availability of calibration datasets (Kupecz *et al.*, 1997; Franchi, 2001). Biased datasets will generate poor models. Furthermore, lack of observations combined to a high amount of features describing each observation can become more difficult, or even prohibitive, to fit a multivariate model to forecast reservoir quality. This problem is known as "curse of dimensionality" (Bellman and Dreyfuss, 1962).

Regression analysis techniques have been extensively used to predict petroleum reservoirs quality. However, this technique has well known limitations and it is highly dependent on domain experts. Otherwise, when domain experts are not available, the dataset quality affects calibration and validation of these models. Latterly, soft computing techniques have been used in many areas, including reservoir characterization and modelling. Among these techniques, Artificial Neural Networks (ANNs) have been increasingly used.

In this paper, we have used the Progressive Enhancement Neural Model (PENM) to predict porosity in sandstones reservoir systems. The main contribution of our approach is to generate a reduced model with more relevant features to porosity prediction. The results show that our approach generates more accurate predictions than commonly used techniques, multiple regression analysis, and conventional ANNs. Another advantage is a lower dependency of domain expert than regression analysis.

The remaining of the paper is organized as follows: The next section presents related works. Neural Modelling section presents basic concepts of ANNs. Experiments and Results section presents the dataset, data preprocessing, experiment results and analysis. The last section concludes the paper and gives some directions of future work.

Related work

Two general approaches have been used for predicting reservoir quality in sandstones: empirical and process-oriented techniques. Empirical techniques use multiple regression analysis. Process-oriented techniques use chemical and mathematical models to understand diagenetic processes and their effects on the evolution of reservoir (Kupecz *et al.*, 1997). Despite the uncertainty associated with simulator-based forecasts, reservoir simulation continues to be the most reliable method for making performance predictions, particularly

for reservoirs that do not have an extensive history (Franchi, 2001).

Regression analysis is the most commonly used technique to predict reservoir quality (Kupecz *et al.*, 1997; Love *et al.*, 1997; Bloch, 1991). However, this technique has limitations and demands intense interaction with domain expert. Moreover, such models are sensitive to the limits imposed by the calibration dataset. Differences in depositional controls, depositional and sequence stratigraphic settings, and sequence stratigraphic concepts between sandstones and carbonates imply creating different models for predictions. Usually, if the dataset embraces data with two or more of these differences, the dataset must be divided in multiple subsets, which must encompass similar observations. This task is domain expert dependent. To achieve more accurate predictions, a model for each subset must be created (Bloch, 1991). Consequently, models are likely to be basin-specific, and may even be restricted to particular facies or stratigraphic horizons, thus inherently limited in their application.

Recently, soft computing techniques have been used in reservoir characterization and modelling (Nikravesh *et al.*, 2003). Among these techniques, ANNs have been used to identify relationships between permeability, measured logs and core data (Ligtenbert and Wansink, 2003), to predict water saturation from log data (Al-Bulushi *et al.*, 2009), to predict asphaltene precipitation in crude oil (Zahedi *et al.*, 2009) and to predict reservoir volume (Akin *et al.*, 2008). ANNs also have been used combined with other techniques, making part of hybrid models, like described in (Chao *et al.*, 2009) to predict borehole stability. Similarly to our work, ANNs and multiple regressions were compared when predicting trap quality (Shi *et al.*, 2004).

Neural modelling

To model this problem, we have selected ANNs technique. ANNs is a biologically inspired computing scheme. The study of ANNs started as an attempt to build mathematical models which worked in the same way that brains do. This study had lead to an abstract computer model of the human brain. ANNs are an adaptive, distributed, and highly parallel system which has been used in many knowledge areas and has proven to solve problems that require pattern recognition (Bishop, 1996). This

model uses a group of algorithms which are considered to implement the fundamental functional source of intelligence (Kupecz *et al.*, 1997). Nowadays, ANNs is a solid technique and became a powerful language for using large flexible non-linear models (Gersheinfeld, 1998).

Analogously to the brain, an ANN is composed of processing nodes, also called artificial neurons, which are interconnected by weighted edges, called synapses. Each neuron receives one or more inputs, multiplied by their weights, and sums these products to generate an output, which is adjusted by an activation function and sent to one or more neurons. Formally, a neuron has inputs x_1, x_2, …, x_m. Each input x_i is multiplied by its corresponding weight w_i. That is, the neuron evaluates $net = x_1 w_1 + x_2 w_2 + … + x_m w_m$. Finally, the neuron computes its output y as an activation function of net, i.e., $y = f(net)$.

Based on structure of the connections among neurons, two different classes of network architecture are identified: non-layered recurrent and layered feed-forward, which is the scope of this work. Regarding layered networks, neurons are organized forming layers. Neurons in a layer get input from the previous layer and feed their output to the next layer. The first layer is called input layer. Neurons at this layer just transmit their input to the next layer, with no computation. The last layer is called output layer. Typically, for regression problems, there is only a single neuron in this layer. For classification problems, there is a neuron for each category of the target variable. Between input and output layers, it can be one or more layers, called hidden layers.

Based on the amount of layers, two different classes of layered feed-forward neural networks are identified: single layer and multilayer. In single layer there is only one computing layer, the output layer. If there are one or more hidden layers, the network is called multilayer or MLP (Multi Layer Perceptron). Single layer networks can only learn linearly separable patterns. Otherwise, MLP can learn non-linear patterns. The universal approximation theorem for ANNs states that MLP can approximate any continuous function (Bishop, 1996). Figure 1 shows a typical MLP.

In order to train an ANN to perform a task, we must adjust the weights of each synapse in such a way as to create a model representing patterns expressed in data. This process is called learning. Although many learning algorithms for MLP have been proposed,
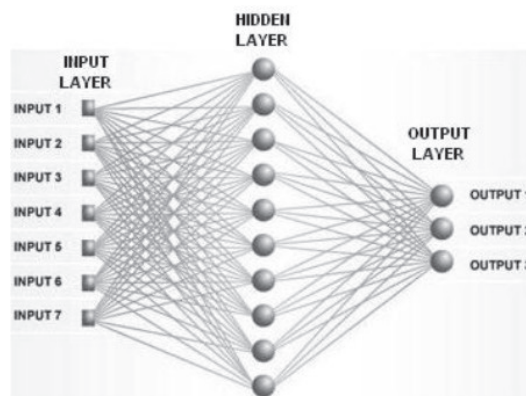

Figure 1. A typical MLP containing 3-layers, 7 neurons in input layer, 10 neurons in hidden layer and 3 neurons in output layer are utilized.

backpropagation algorithm is the most widely used (Munakata, 2008).

The main difficult in the use of MLP lies in the design process. There are no rules to choose the best ANN configuration, in terms of layers and neurons in each layer, and the best training parameters. Therefore, designing an MLP demands some iteration in order to find the best MPL settings to a particular problem.

Although ANNs is a powerful technique they are affected by the "curse of dimensionality" in two ways: (i) in high dimensional data the ANN can use almost all its resources to represent irrelevant portions of search space; (ii) even if ANNs could focus on important portions of search space, the higher the dimensionality of input space the more data may be needed to find out what is important and what is not. Moreover, according the Ockham's razor principle, the modeller should select the most simple model and grossest reservoir description that will allow the desired estimation of reservoir performance (Franchi, 2001).

Progressive enhancement neural model

In order to overcome the "curse of dimensionality" problem, we have used an approach called progressive enhancement neural model (Camargo and Engel, 2010). This approach includes creating regression models in which the choice of predictive features is carried out by an automatic procedure. To guide this automatic procedure, we have been used a neural importance score, which is computed by:

$$s_i = \frac{\sum_{k=1}^{n} |w_{ik}|}{n}$$

where:

- s_i represents the score for i-th feature;
- n represents amount of neurons in the first hidden layer;
- w_{ik} represents the weight of synapse between i-th neuron of input layer and k-th neuron of first hidden layer.

The main objective of this score is to quantify the importance of each feature regarding target prediction. After MLP training, patterns expressed in data were learned and are represented in synaptic weights. So, the largest synaptic weights are supposed to be linking the most important input features to the first hidden layer. If a synaptic weight tends to zero, its propagation to the first hidden layer, and consequently to following layers, will tend to zero too, denoting its little importance regarding target prediction. While synapses among input and first hidden layer are supposed to encode patterns expressed in data, synapses among first hidden layer and output layer are supposed to decode patterns expressed in data to reconstruct this pattern. For this reason, our approach is concentrated in synaptic weights before first hidden layers, and ignores the others.

In order to provide data to compute the scores of features, a conventional MLP must contain knowledge about the dataset. This way, before computing the scores, MLP must be previously trained with original training dataset, in which each sample must be described by all available features. After this training, scores can be computed.

In order to achieve the progressive enhancement of initial neural model, an iterative process is performed. This process is performed as follows: After score computing, it is executed a forward selection, which involves starting with the simplest model, containing just the largest score feature. At each stage, the next feature available, which has the largest score among unselected features, is inserted in the model, and the model is evaluated. This process continues iteratively until the measure is locally maximized, or when the available improvement falls bellow some threshold. Prediction error is the metric used to evaluate models. Model evaluation is performed by some cross-validation technique.

The PENM approach is similar to Effroymson's algorithm (Effroymson, 1960). However, while Effroymson's approach is f-test and multiple regression-based, our approach is neural network-based.

The final model, generated by our approach, is supposed to be enhanced regarding the initial MLP. This enhancement is obtained through selection of a subset of available features. This subset must contain just the most important features regarding target prediction. Formally, the progressive enhancement process will choose a subset of M features from the original set of N features, where ($M \leq N$). Furthermore, if the enhanced model contains just the most important features, it is expected to be more accurate than the original model, which contains the original features.

Experiments and results

Available data

For this study, we have used an existing dataset published by (Lima and De Ros, 2002). This database contains observations of petrographic and petrophysical characteristics of Devonian sandstone reservoirs of the Uerê Formation, which is an important oil exploration target of the Solimões Basin. Further, exploration of the Uerê sandstones is complicated by the heterogeneous quality of these reservoirs, which range from highly porous to extremely tight. Moreover, although the Solimões Basin has been explored throughout the past three decades, little is known about the reservoir quality control.

Another challenge is dataset dimensionality. This makes application of modelling techniques harder. The dataset contains 59 samples and 88 features, which describes the petrographic and petrophysical characteristics of sandstones.

Data pre-processing

Before model creation, the dataset must be prepared to improve modelling process. The pre-processing phase includes some tasks like cleaning and transformation.

During cleaning task, one sample was deleted. This sample was considered an outlier by domain expert. Features containing exactly the same value for all samples were deleted. A well known pattern involving prediction target was eliminated from input data. Macroporo-

sity can be directly computed through a sum of intergranular, intragranular in feldspar, intragranular in quartz grain, intragranular in mica, intragranular in heavy mineral, dissolution of pseudomatrix, dissolution of cement, mouldic, fracture and oversizes characteristics. So, these features were deleted, in order to allow finding non-trivial relationships between macroporosity and other characteristics. Some features holding the sum of others were deleted.

Regarding transformation task, input data were normalized by decimal scale technique (Han and Kamber, 2001). Original data, which were in [0,100] range, were divided by 100. Subsequently, data were transformed to fall in [–1,1] range, in order to be used by the neural network.

Thus, after pre-processing, the dataset contained 58 samples and 60 features.

The ANN was developed and tested on a Windows Based PC using MatLab software. Several MLP network structures were automatically tested and evaluated by cross-validation process. These structures had one or two hidden layers and from one to 60 neurons in each hidden layer. Different learning algorithms were tested too. After these several tests, the best performance ANN was adopted.

Adopted neural network

After testing and evaluating many architectures and learning algorithms, it was selected an ANN with the following characteristics:

Network Architecture
• Multi Layer Perceptron;
• Activation strategy: Feedforward;
• Totally connected nodes;
• 3 layers;
• 60 neurons in input layer;
• 4 neurons in hidden layer;
• 1 neuron in output layer.

Information processing in nodes
• Hidden layer(s) information processing: Sum is the input/propagation function. Hyperbolic tangent sigmoid activation function. BIAS node.
• Output layer information processing: Sum is the input/propagation function. Linear output function. BIAS node.

Learning algorithm
• Form of learning algorithm: Gradient descent with momentum and adaptive learning backpropagation;

• Learning parameters: Learning rate: 0.01, ratio to increase learning rate: 1.05, ratio to decrease learning rate: 0.7, momentum constant: 0.0. Epochs:
• Dynamic strategy of learning parameter change: If the new error exceeds the old error by more than a predefined ratio (defined as 1.04), the new weights and biases are discarded. In addition, the learning rate is decreased (multiplied by 0.7). Otherwise, the new weights, etc., are kept. If the new error is less than the old error, the learning rate is increased (multiplied by 1.05).
• 20 initializations/retraining;
• Synaptic weights were randomly generated and were normalized to a length of 1;
• Stopping method: minimum gradient reached or maximum epoch reached;
• Error measure: Mean Squared Error (MSE);

Model Selection
• MSE was the error measure used for evaluation;
• Selection of the best model was based upon lowest error on test set.

Model evaluation

Unfortunately, prediction methods are susceptible to overfitting the learning examples at the cost of decreasing generalization accuracy over unseen examples. For small training sets this problem is severer. Due to the lack of samples to perform an early stopping approach, and prevent overfitting, models evaluation must be carefully planned. One of the most successful methods for evaluating performance accuracy is the leave-one-out cross-validation technique, in which the set of m training instances is repeatedly divided into a training set of size m-1 and test set of size 1, in all m possible ways.

Model error is the sum of absolute errors for each one of m tests performed during leave-one-out cross-validation process.

Comparison and analysis

After selecting the best ANN architecture and algorithms, and after its training, patterns found in data are expressed in ANN synaptic weights. Figure 2 shows the scores from trained ANN. Few features, which are high scored, are identified as important and the most part of features are low scored. Low
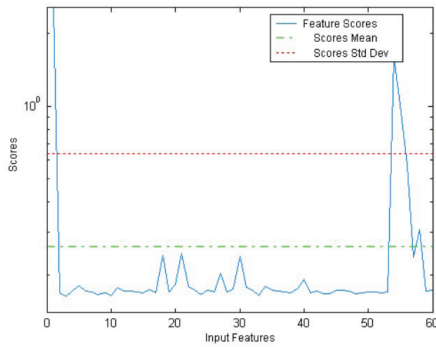
Figure 2. Scores for each input feature.

scored features are perceived as noise in conventional ANN. Higher scored features are supposed to be most important in target feature prediction.

Table 1 shows the top 10 ranking containing the scores of each feature. These scores are the basis for PENM creation. Features, which their scores are greater than mean plus a standard deviation, were emphasized.

Table 1. Top 10 feature scores for macroporosity prediction.

Position	Feature	Score
1	Quartz Monocrystalline	2.5538
2	Intergranular Volume	1.5992
3	Cement Total	1.0115
4	Carbonate Total	0.5677
5	Grain Replacement Total	0.3078
6	Quartz Overgrowth	0.2446
7	Clay Ooid	0.2396
8	Silicified Secondary Matrix	0.2379
9	Silica Total	0.2375
10	Microquartz Rims	0.2030

Let n as the amount of features in original dataset, the iterative process performed during PENM creation can generate until n models. Table 2 shows evaluation results for each one of these models. Clearly, some models, which contain just a subset of all available features, are more accurate than the full model, which contains 60 features of original dataset. In this case, the model containing the 3 largest scored features is the enhanced neural model generated by our approach.

In order to evaluate the results obtained with PENM approach, other approaches were applied to this dataset: Multiple Regression,

which is the most commonly used technique to predict reservoir quality (Kupecz *et al.*, 1997; Love *et al.*, 1997; Bloch, 1991), and conventional neural networks.

Figures 3, 4 and 5 show a plotting of predicted and measured macroporosity using Multiple Regression, conventional ANN, and PENM, respectively. In each figure, correlation coefficients between measured macroporosity and predicted macroporosity are shown.

Table 2. Macroporosity prediction errors.

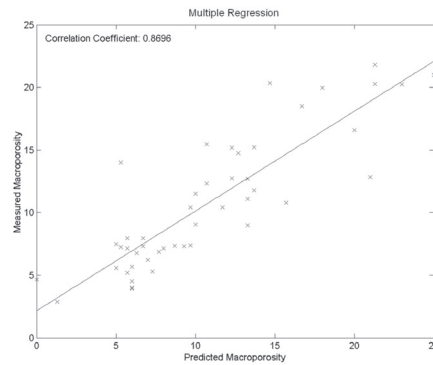| Input features	Absolute error	% Error
1	2.7173	26.17
2	2.7110	26.11
3	1.8637	17.97
4	1.9578	18.86
5	2. 0533	19.78
...	...	...
60	2.1403	20.61

Figure 3. Correlation coefficient between predicted and measured macroporosity using multiple regression.
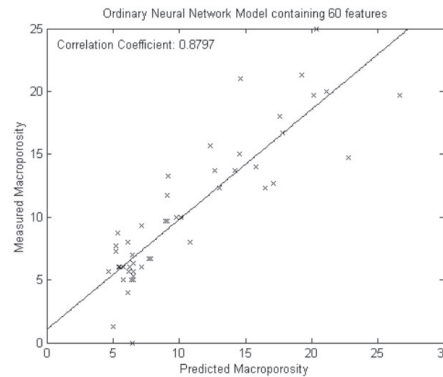

Figure 4. Correlation coefficient between predicted and measured macroporosity using a conventional ANN.
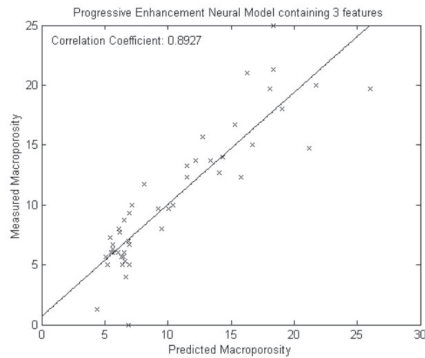
Figure 5. Correlation coefficient between predicted and measured macroporosity using PENM.
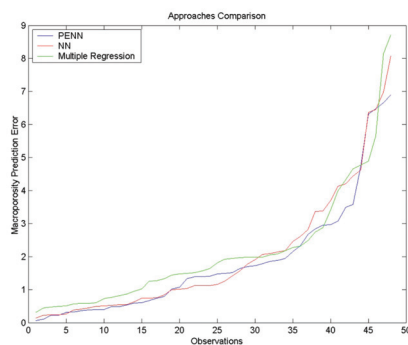

Figure 6. Residual error comparison obtained through application of three different approaches.

In Uerê formation dataset (Lima and De Ros, 2002), comparing correlation coefficients, it is possible to conclude that macroporosity values predicted by PENM are the closest values of measured macroporosity.

Furthermore, Figure 6 shows a comparison among residual errors obtained through application of different predictive approaches employed in this work. In order to improve results visualization, residual errors of each approach were sorted in ascendant order. This figure shows that PENM achieves less residual errors than the other approaches used.

Conclusions

In this paper, we have used the Progressive Enhancement Neural Model to predict reservoir quality in sandstones. The main contributions of this paper include the following topics:

(i) The Progressive Enhancement neural model successfully predicted macroporosity with a correlation coefficient of 0.8927 on 58 samples in a well in the Uerê formation.

The conventional neural model predicted macroporosity with a correlation coefficient of 0.8789 in the same dataset. The commonly used multiple regression predicted macroporosity with a correlation coefficient of 0.8696. Thus, the progressive enhancement neural model has proven to be a powerful approach to predict sandstones macroporosity.

(ii) Despite a small accuracy gain obtained with PENM, the main contribution of our approach is generating simpler models, which contain just a small subset of features, without accuracy loss. Obeying Ockham's razor principle, models generated by PENM approach must be preferred due to simplicity.

(iii) Our approach can generate more explainable models, because it ranks the importance of each feature regarding target feature prediction. One of the greatest criticisms to ANN use is the generation of unexplainable black box models. Hence, this study can indicate a way to open the black box.

(iv) We have done several experiments to predict rocks permeability. PENM has created simpler and more accurate models. However, results were unsatisfactory, probably because petrographic and petrophysical characteristics are weakly related with permeability prediction.

(v) We also applied PENM to create models to classify observations according its petrofacies in this same dataset. In these experiments, small accuracy gains were repeated and simpler models were generated.

References

AKIN, S.; ROSS, C.M.; KOVSCEK, A.R. 2008. Combination of well log and pore-scale data to predict pretrophysical properties of diatomite. *Journal of Petroleum Science and Engineering*, **60**:133-149.

http://dx.doi.org/10.1016/j.petrol.2007.05.013

AL-BULUSHI, N.; KING, P.R.; BLUNT, M.J.; KRA-AIJVELD, M. 2009. Development of artificial neural network models for predicting water saturation and fluid distribution. *Journal of Petroleum Science and Engineering*, **68**:197-208.

http://dx.doi.org/10.1016/j.petrol.2009.06.017

BELLMAN, R.; DREYFUSS, S. 1962. *Applied dynamic programming*. Princeton, Princeton University Press, 366 p.

BISHOP, C.M. 1996. *Neural networks for pattern recognition*. Oxford, Oxford Press, 504 p.

BLOCH, S. 1991. Empirical predictions of porosity and permeability in sandstones. *AAPG Bulletin*, **75**:1145-1160.

CAMARGO, S.S.; ENGEL, P.M. 2010. A progressive enhancement neural model to predict reservoir quality in sandstones. *In:* SOUTHERN CONFERENCE ON COMPUTATIONAL MODELING, 3, Rio Grande, 2010. Proceedings... Rio Grande, p. 100-106.

http://dx.doi.org/10.1109/MCSUL.2009.25

CHAO, W.; MIAN, C.; YAN, J. 2009. A prediction method of borehole stability based on seismic attribute technology. *Journal of Petroleum Science and Engineering*, **65**:208-216.

http://dx.doi.org/10.1016/j.petrol.2008.12.033

EFFROYMSON, M.A. 1960. Multiple regression analysis. *In:* A. RALSTON; H.S. WILF (eds.), *Mathematical Methods for Digital Computers*. New York, Wiley, p. 191-203.

FRANCHI, J.R. 2001. *Principles of applied reservoir simulation*. 2nd ed., Burlington, Elsevier, 294 p.

GERSHEINFELD, N. 1998. *The nature of mathematical modelling*. Cambridge, Cambridge University Press, 356 p.

HAN, J.; KAMBER, M. 2001. *Data mining concepts and techniques*. San Diego, Morgan Kaufman, 550 p.

INTERNATIONAL ENERGY AGENCY. 2010. Strategic cost world energy outlook. Available at: http://www.eia.gov/oiaf/ieo/. Accessed: 01/10/2010

KUPECZ, J.; GLUYAS, J.; BLOCK, S. 1997. Reservoir quality prediction in sandstones and carbonates: An overview. *In:* J. KUPECZ; J. GLUYAS; S. BLOCK (eds.), *Reservoir quality prediction in sandstones and carbonates*. Tulsa, USA, American Association of Petroleum Geologists, p. VII-XXIV. AAPG Memoir 69.

LIGTENBERT, J.H.; WANSINK A.G. 2003. Neural network prediction of permeability in the El Garia Formation, Ashtart Oilfield, Offshore Tunisia. *In:* M. Nikravesh; F. Aminzadeh; L.A. Zadeh (eds.), *Developments in Petroleum Science,* **51**:397-411.

http://dx.doi.org/10.1016/S0376-7361(03)80023-7

LIMA, R.D.; DE ROS, L F. 2002. The role of depositional setting and diagenesis on the reservoir quality of Devonian sandstones from the Solimões Basin, Brazilian Amazonia. *Marine and Petroleum Geology*, **19**:1047-1071.

http://dx.doi.org/10.1016/S0264-8172(03)00002-3

LOVE, K.M.; STROHMENGER, C.; WORONOW, A.; ROCKENBAUCH, K. 1997. Predicting reservoir quality using linear regression models and neural networks. *In:* J. KUPECZ; J. GLUYAS; S. BLOCK (eds.), *Reservoir quality prediction in sandstones and carbonates*. Tulsa, American Association of Petroleum Geologists, p. 47-60. (AAPG Memoir 69).

MUNAKATA, T. 2008. *Fundamentals of the new artificial intelligence: Neural, evolutionary, fuzzy and more*. 2nd ed., London, Springer, 260 p.

NIKRAVESH, M.; AMINZADEH, F.; ZADEH, L.A. 2003. *Soft computing and intelligent data analysis in oil exploration*. Berkeley, Elsevier, 754 p.

SHI, G.; ZHOU, X.; ZHANG, G.; SHI, X.;LI, H. 2004. The use of artificial neural network analysis and multiple regression for trap quality evaluation: a case study of the Northern Kuqa Depression of Tarim Basin in western China. *Marine and Petroleum Geology*, **21**(3):411-420.

http://dx.doi.org/10.1016/j.marpetgeo.2004.01.003

ZAHEDI, G.; FAZLALI, A.R.; HOSSEINI, S.M.; PAZUKI, G.R.; SHEIKHATTAR, L. 2009. Prediction of asphaltene precipitation in crude oil. *Journal of Petroleum Science and Engineering*, **68**:218-222.

http://dx.doi.org/10.1016/j.petrol.2009.06.023