Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran

Vahid Bolandi[1]; Ali Kadkhodaie[2,3*]; Reza Farzi[4]

1. Faculty of Geosciences, Shahid Chamran University, Ahwaz, Iran

2. Department of Earth Science, Faculty of Natural Science, University of Tabriz, Tabriz, Iran

3. Department of Petroleum Engineering, Curtin University, Perth, Western Australia

4. School of Geology, University College of Science, University of Tehran, Tehran, Iran

Abstract

Determination of TOC is critical to the evaluation of every source rock unit. Methods which are dependent upon extensive laboratory testing are limited by the availability and integrity of the rock samples. Prediction of TOC (Total Organic Carbon) from well Log data being available for the majority of wells being drilled provides rapid evaluation of organic content, producing a continuous record while eliminating sampling issues. Therefore, the ideal method for determining the TOC fraction within source rock units would utilize common well log data. So a model was developed to formulate TOC values in the absence of laboratory TOC measurements from conventional well log data. Consequently, with the assistance of FL (Fuzzy Logic), TOC estimated from well log data with an overall prediction accuracy of 0.9425 for the test set. Following that TOC content of the Kazhdumi formation optimally has been divided into 4 zones using K-means cluster analysis, since searching for patterns is one of the main goals in data mining. There is a general increase in TOC from zone 1 to zone 4. The optimal number of zones has been detected by means of the knee method that finds the

*Corresponding author: Tel: +98 912 638 3051

E-mail addresses: kadkhodaie_ali@tabrizu.ac.ir (A. Kadkhodaie), bolandi_v@yahoo.com (Vahid Bolandi), rezafarzi@akumni.ut.ac.ir (Reza Farzi)

"knee" in a number of clusters vs. Compactness, Davies-Bouldin and Silhouette values. In the last step, using SVM (Support Vector Machine) and ANN (Artificial Neural Network) algorithms, two commonly used techniques, classification rules developed to predict the source rock class-membership (zones) from well log data. The proposed method is found effective in directly extracting patterns from well log data after defining classification rules. Quantitative comparisons of the results from ANN and SVM depicts that for classification problem of source rock zonation SVM with RBF (Radial Basis Function) kernel readily outperforms ANN in term of classification accuracy (0.9077 and 0.9369 for ANN and SVM, respectively), reduced computational time and highly repeatable results. This method would enable a more elaborate assessment of Kazhdumi formation to be undertaken by providing a comprehensive quick look results derived directly from well log data while using conventional methods one can't define patterns within the data without grouping data manually.

Keywords: FL (Fuzzy Logic), K-means cluster analysis, SVM (Support Vector Machine), ANN (Artificial Neural Network), Source rock.

1. Introduction

Petroleum source rocks are the primary component of the petroleum system and constitute the precursors of oil and gas which, under favorable conditions, may ultimately migrate to reservoirs and be trapped to form accumulations (Magoon and Dow, 1994). In source rock evaluation point of view, TOC is the most important rock parameter that affects source rock concept. Thus, TOC quantification and its spatial distribution considered being one of the fundamental steps in source rock evaluation. This parameter is best measured using LECO and Rock-Eval pyrolysis apparatuses on cutting and core derived samples. These data are usually sparse since a few exploration wells intentionally penetrate into the source rock horizons so a limited number of samples can be achieved for laboratory analysis. This places a great emphasis on employing methods to estimate this parameter utilizing common data such as well log data, which are available almost in all wells. Before the advent of artificial intelligence the attempts to establish qualitative and quantitative relationships among well log data and TOC values has generally been in the form of empirical correlations (Carpentier et al.,

1989; Fertle and Ricke, 1980; Fertle, 1988; Herron, 1988; Mendelson and Toksoz, 1985; Meyer and Nederlof, 1984; Passey et al., 1990; Schmoker and Hester, 1983 and Schmoker, 1981, 1979), where Δ log R method of Passey et al. (1990) has been used with some success. Flowing that more recently intelligent techniques alternatively have been increasingly applied to estimate TOC values from well log data (Alizadeh et al. (2012), Bolandi et al. (2015), Huang and Williamson (1996), Kadkhodaie-Ilkhchi et al. (2009) and Kamali and Allah Mirshady (2004)). Moreover, previous investigations have revealed that intelligent techniques are superior to conventional methods in identifying the complex relationship among TOC and well log data even in highly heterogeneous formations because of their pattern recognition ability.

Over the past decades, studies have addressed the use of Artificial Intelligence in petroleum geoscience, but generally focused were on regression and clustering methods. By contrast, the classification method is less documented. Cluster analysis is a popular technique whose basic objective is to discover unknown sample groupings within data, while classification attempts to assign each input value to one of a given set of classes. So in this study, machine-learning techniques are used for supervised learning of classifiers. The goal is to construct a classifier that can correctly predict a label sequence subjecting a new input sequence. Here, the output classes are TOC zones (Class labels are given to FL predicted TOC log using k-means clustering algorithm), and the inputs are conventional well log data. In this paper, first of all, a detailed discussion on regression, clustering and classification are presented. After that, a comparative study is done experimentally. On the basis of the result found, a conclusion is then drawn for the comparison. Thus, we first used FL to estimate TOC values from well log data, validated with the measured geochemical data from Rock-Eval pyrolysis. The results show that TOC content of the Kazhdumi formation can be optimally classified into four classes using the K-mean algorithm. Then, SVM and ANN are evaluated for developing a classification rule that can accurately predict the class membership of unknown samples from well log data. SVM, outperform ANN for in extracting patterns directly from well log data without the need to do the tedious task of regression and clustering analysis for other wells, providing a comprehensive quick look results in reduced computational time, no iteration, and with highly repeatable results.

2. Geological setting

The study area is located in the northwest of the Persian Gulf (Fig. 1), geologically in the Dezful embayment. The basin is located at the junction of the Arabian Shield and Iranian continental block that belong to two different (Arabian and Eurasian) lithospheric plates. The collision of these plates at the Mesozoic/Cenozoic boundary produced the Zagros Fold Belt. The Persian Gulf basin is an approximately 2600 km long and 900-1800 km wide basin that includes the Persian Gulf as well as onshore areas, spanning the last 650 Ma and is the largest basin with active salt tectonism in the world. This basin is bounded to the SW by the Arabian shield, whereas to the NE it is bounded by the Zagros fold and thrust belt. The Persian Gulf basin exhibits a strongly wedge-shaped nature in a NE-SW cross section with sediment thicknesses ranging from 18 km just S of Zagros fold and thrust belt (Morris, 1977) to approximately 4500 m near the Arabian shield, which is in fact situated in the offshore area of Zagros Fold Belt (Edgell, 1996). The lithological variety of the Persian Gulf includes several potential source, reservoir (mainly carbonate rocks and some clastic ones) and cap rocks, which together with the occurrence of important stratigraphic and structural traps and the contribution of processes that completed the petroleum system (Edgell, 1996; Abdollahie Fard et al., 2006), makes the Persian Gulf basin the richest region of the world in terms of hydrocarbon resources, containing 55–68% of recoverable oil reserves and more than 40% of gas reserves of the world (Konyuhov and Maleki, 2006). By the way Organic-rich rocks, which are widespread at different levels in the rock sequences of the study area including Pabdeh (Paleocene- Oligocene), Kazhdumi (Albian-Cenomanian), Garau (L.Certaceous) and Sargelu (M.Jurrasic) formations, in this study Kazhdumi formation as a part of Bangestan group (James and Wynd, 1965) has been selected on estimation of TOC from petrophysical data. It is a proven Early Cretaceous source rock, which presumably produced the majority of the commercial hydrocarbons in this area (Bordenave, 2002; Bordenave & Huc, 1995) and stratigraphically positioned between the Cenomanian-Turonian Sarvak Formation and Aptian Dariyan Formation consisting of an alternation of shale and limestone where detailed

discussion are found in Motiei (1995). Also, the regional geology of the Persian Gulf and the adjacent area has been discussed in numerous publications (Bordenave and Burwood, 1990; James and Wynd).

3. Methodology

This study consists of five major steps:

1. Rock-Eval pyrolysis measurements on drill cutting samples
2. Synthesis of TOC log from conventional well log data using FL and
3. Calculating TOC applying Δ log R technique
4. Grouping the synthesized TOC log data into clusters using k-means algorithm
5. Training SVM and ANN classifiers to predict the source rock class membership (zones) from well log data
6. Correlation of estimated with the achieved source rock class membership (zones)

This integrated technique could be considered as a favored way for predicting TOC parameter and extracting source rock distribution pattern from well log data by machine learning techniques. The basic concepts are introduced briefly in this section since extensive information already exists in the literature.

3.1 Rock-Eval Pyrolysis

For this study 31 cutting samples of Kazhdumi formation were taken to assess geochemical characteristics (Fig. 2). Accordingly, Rock-Eval 6 pyrolysis was performed following the method described by Espitalie´ et al. (1977) and Lafargue et al. (1997), under standard conditions with a temperature program of 25 °C min^{-1}, where the final temperature reaches 800 °C in the pyrolysis oven and 850 °C in the oxidation oven. For this purpose, each cutting sample was decontaminated (from micas that come from lost circulation material (LCM) iron filings from the drill bit), pulverized to a fine powder and all samples were weighed to 60-70 mg prior to subjecting to the apparatus. This provides a powerful technique that assesses the quantity, type and thermal maturity, which are the main parameters for characterizing a hydrocarbon source rock. The apparatus provides TOC (wt.%), S_1: volatile hydrocarbons (free hydrocarbons) [mg HC/g rock], S_2: hydrocarbons derived from

kerogen pyrolysis (hydrocarbons cracked) [mg HC/g rock], T_{max} (°C): the temperature at the highest yield of S_2 hydrocarbons, which are the main obtained parameters together with several calculated parameters from measured parameters encompassing HI (Hydrogen index): $(S_2/TOC)\times100$ [mg HC/g TOC]; OI (Oxygen index): $(S_3/TOC)\times100$ [mg CO2/g TOC] and PI (Production index): $S_1/(S_1+S_2)$ (Fig. 2). Summary of interpretive guidelines for Rock-Eval analytical procedure and discussion on Rock-Eval parameters are available in Espitalié et al. (1977), Langford and Valleron (1990), Peters and Cassa (1994) Lafargue et al. (1997).

3.2. Fuzzy inference system

A fuzzy inference system (FIS) is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made or patterns discerned. The basic concept of fuzzy logic, or fuzzy set theory, was first introduced by Zadeh (1965). Unlike crisp logic (Boolean logic), fuzzy logic differs from classical logic in that statements are no longer black or white, true or false, on or off. In traditional logic, an object takes on a value of either zero or one. In fuzzy logic, a statement can assume any real value between 0 and 1, representing the degree to which an element belongs to a given set. Therefore, FL is well suited to solve geosciences problem which is associated with uncertainty and vagueness (Kadkhodaie-Ilkhchi et al., 2006), capable of modeling nonlinear functions of arbitrary complexity (Matlab user's guide, 2012). The formulation between inputs and outputs is performed through a set of fuzzy if–then rules. Normally, fuzzy rules are extracted through a fuzzy clustering process. Subtractive clustering is one of the effective methods for constructing a fuzzy model. The effectiveness of a fuzzy model is related to the search for the optimal clustering radius, which is a controlling parameter for determining the number of fuzzy if–then rules. Fewer clusters might not cover the entire domain, and more clusters (resulting in more rules) can complicate the system behavior and may lead to lower performance. It is necessary to optimize this parameter for the construction of the fuzzy model. (a) Fuzzification (translate input into truth values), (b) Rule Evaluation (compute output truth values) and (c) Defuzzification (transfer truth values into output) are the main steps in FIS, where discussion are available in Kadkhodaie-Ilkhchi et al. (2009) and Guillaume (2001).

3.3. Feature descriptions between TOC and input petrophysical data

Generally speaking, a source rock is defined by high gamma-ray intensity (Dellenbach et al., 1983). The gamma ray tool measures the radioactivity of various formations. Organic-rich rocks have high concentrations of radioactive elements including potassium, thorium and uranium and increase the gamma-ray response.

Generally, Neutron log reading is a response to hydrogen atoms concentration in rocks. The volume of organic matter in the formation has a direct relationship with hydrogen atoms content and porosity of the rock. Thus, neutron porosity increases in the organic-rich intervals.

The sonic transit time (DT) is the reciprocal of the velocity of the compressional wave and is a function of formation lithology, porosity, type and distribution models of fluids (water, gas, oil, kerogen, etc.). as a result with apparent DT value increase TOC content tends to elevate (Kamali and Mirshady, 2004).

Density log measures the bulk density of the formation, a response of fluids and matrix constituent minerals density. Organic matters have a low density (about $1g/cm^3$) and their concentration tends to reduce the bulk density of the rock (Schmoker, 1979).

In addition, the resistivity increases dramatically in mature source rocks and presumably is related to the generation of non-conducting hydrocarbons (Nixon, 1973; Schmoker and Hester, 1989; and Meissner, 1978). Thus, the resistivity responses of a source rock will be affected by its TOC content.

Owing to the above-mentioned relationships and considering the direct and reverse relationships between well log responses with TOC values and (Fig. 3) and observing plots of measured TOC on top of the logs, which is the most compiling presentation for showing log correlations (Fig. 4), sonic (DT), density (RHOB), neutron porosity (NPHI), gamma ray (GR), deep resistivity (LLD) and shallow resistivity (LLS) are good indicators and have been utilized in TOC estimation modeling. In contrast, the potassium (POTA), uranium (URAN) and thorium (THOR) logs were discarded due to their very weak association with TOC values and trial and error input selection.

3.4 Pattern recognition

In short, pattern recognition is defined as the recognition of patterns and regularities in data. Generally, pattern recognition is a subfield of Artificial Intelligence that deals with classification (Supervised learning) and clustering (Unsupervised learning) problems according to the learning process used to generate the output values. Supervised learning attempts to assign each input value to one of a given set of classes, while unsupervised learning deals with a dataset that has not been labeled, and tries to discover unknown patterns.

3.4.1. Clustering

Cluster analysis is a popular unsupervised categorizing technique recognized as an important area of data mining (Jain et al., 1999) encountered in many fields, such as geology and geochemistry. In fact, the basic objective of clustering is to discover uncovered relationships within data. It gets the data and divides data elements into different groups (known as clusters) in such a way that the elements within a group possess high similarity while they differ from the elements in a different group. The whole clustering process should follow the maximizing the Intracluster property and minimize the inter-cluster property. The similarity between samples is assessed by measuring the distances between the points in the measurement space. Samples that are similar will lie close to one another, whereas dissimilar samples are distant from each other. There are some clustering techniques available. The focus here is on K-means method, considered to be one of the popular unsupervised data clustering algorithm (MacQueen, 1967; Khedairia and Khadir, 2008).

The K-means algorithm takes the input parameter k, the number of clusters, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point, we need to recalculate k new

centroids. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, we may notice that the k centroids change their location step by step until no more changes are done. In other words, centroids do not move anymore. Finally, this algorithm aims at minimizing an objective function, in this case, a square error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{k} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the n data points from their respective cluster centers.

The Formal Algorithm (Ayesha et al., 2010) is:

1. Select K points as initial centroids

2. Repeat

3. Form k clusters by assigning all points to the closest centroid

4. Recompute the centroid of each cluster

5. Until the centroids do not change

The K-Means is a simple algorithm that has been adapted to many problem domains and it is a good candidate to work for a log data points.

K-Means is based on the minimization of the average squared Euclidean distance between the data items and the cluster's center (called centroid).

3.4.2. Classification

A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the

features or observed variables). The goal is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes

Support Vector Machine classifier

The SVM algorithm is a nonlinear classification algorithm based on kernel methods, first developed by Vapnik and Lerner (Vapnik and Lerner, 1963) and Vapnik and Chervonenkis (Vapnik and Chervonenkis, 1964). This algorithm is firmly grounded in the framework of statistical learning theory - Vapnik Chervonenkis (VC) theory, which improves the generalization ability to learn machines to unseen data (Smola and Schollköpf, 1998). Its algorithm can construct a maximum marginal hyperplane in the training sample by a set of particular training samples called support vectors (Fig. 5). The maximum marginal hyperplane can separate two types of datasets as much as possible and maximize the difference between them. The SVM finds the maximum marginal hyperplane (optimal-classification hyperplane) through kernel functions.

Given a training set of instance-label pairs $(x_i, y_i), i = 1, \dots, l$ where $x_i \in R^n$ and $y \in \{1, -1\}^l$, The optimal hyperplane for Linear SVM, is found by solving the following constrained primal Quadratic Programming Problem

$$\min_{w,b,e} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{l} \xi_i \ , \qquad (2)$$

Subject to $y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,3, \dots, l,$

Where C is a penalty parameter for misclassified points in a training set and with slack variables $\xi_i \geq 0$. The goal is to define an optimal separating hyperplane

$$w^T x + b = 0 \qquad (3)$$

where $x_i \in R^n$. Using Wolfe's dual formulation, eq. (2) can be expressed as:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \left(x_i . x_j\right) \alpha_i \alpha_j - \sum_{j=1}^{l} \alpha_j \qquad (4)$$

Subject to $\sum_{i=1}^{l} y_i \alpha_i = 0$, where $0 \leq \alpha_i \leq C, \quad i = 1,2,3, \ldots, l,$

Where $\alpha \in R^l$ are lagrangian multipliers. The optimal separating hyperplane of eq. (3) can be given by

$$w = \sum_{i=1}^{l} \alpha_i^* y_i x_i, \qquad b = \frac{1}{N_{sv}} \left(y_j - \sum_{i=1}^{N_{sv}} \alpha_i^* y_i \left(x_i . x_j \right) \right), \qquad (5)$$

Where α^* is the solution of the dual problem eq. (5), N_{sv} represents the number of support vectors satisfying $0 < \alpha_i < C$. A new sample is classified as +1 or –1 according to the final decision function. In order to deal with non-linearity of the classification problem, one can define the mapping of examples to a so-called feature space of very high dimensions. The basic idea of this mapping into high dimensional space is to transform the non-linear case into the linear one. The kernel methods map the original parameter vectors into a higher dimensional feature space, without the need to compute the nonlinear mapping explicitly. The commonly used kernel functions include Polynomial, RBF and Sigmoid. For SVM classifier er, choice of kernel function is significantly important. Thus, SVM is capable of solving complex nonlinear classification problems since classes which are nonlinearly separable in the original space can be linearly separated in the higher dimensional feature space, dominant feature which makes SVM a powerful tool.

In this paper, the least squares version of SVM (Suykens et al., 2002) were applied where the least squares term added in the cost function, which transformed the problem from solving a QP problem to solving a set of linear equations (Suykens, 1999).

Artificial Neural Network classifier

Generally speaking, ANN can be considered as an information processing system composed of so-called neurons which are networks of interconnected simple processing elements. Neural networks are very good at pattern recognition problems. A neural network with enough elements (called neurons) can classify any data with arbitrary accuracy. They are particularly well suited for complex decision boundary problems over many variables. Two-layer (i.e. one-hidden-layer) feed-forward neural networks can learn any input-output relationship given enough neurons in the hidden layer (Matlab

user's guide, 2012). These efficient networks are widely used to solve complex problems by modeling complex input-output relationships. You can train a neural network (The learning process) to perform a particular function by adjusting the values of the connections (weights) between neurons. Processing of artificial neural network is done in three phase: training, validating, and testing. Firstly network is trained using input dataset. In this process, the weights are adjusted such that the mean squared error obtained between the experimental and obtained result can be minimized. The output of a single neuron p can be written as:

$$p = g\left(w^T x - b\right) = g\left(\sum_{k=1}^{n} w_i x_i - b\right) \qquad (6)$$

where $x = (x_1, x_2, \dots, x_n)^T$ denotes the network input vector and $w = (w_1, w_2, \dots, w_n)^T$ denotes the weight parameters to be adjusted. $g(.)$ is the activation function used in the hidden layer among which logistic sigmoid function ($f(x) = (1 + e^{-x})^{-1}$) is the most well-thought-of activation function. Activation functions for the hidden units are needed to subject non-linearity into the networks.

4. Application to the one of the offshore Iranian oilfield

4.1. TOC Prediction using FIS

One of the most important factors affecting source rock evaluation is TOC. This parameter best derived by Rock-Eval pyrolysis. However, geochemical data are just available in limited numbers. Due to lack of geochemical data from the entire sequence, a Takagi–Sugeno fuzzy inference system (TS-FIS) (Takagi and Sugeno, 1985) was created to model this parameter from well log data, which enables us to detail TOC distribution prospect of the sequence. For this purpose, DT, GR, NPHI, RHOB, LLS and LLD were taken for TOC prediction. Fuzzy if-then rules and inputs and output membership functions were extracted by subtractive clustering method as the powerful and efficient method for FL modeling (Chiu, 1994 & Kadkhodaie-Ilkhchi et al. 2010). Cluster radius which controls the cluster numbers and fuzzy rules is the significant parameter. Clustering process was accomplished gradually with cluster radius from 0.005 to 1 (with 0.005 intervals) to achieving optimal cluster radius. Thus, 200 fuzzy models with the different number of fuzzy if-then rules were

established. Then, the fuzzy model with the highest overall accuracy was selected as the optimal model. The results show that taking the clustering radius of 0.5 leads to the highest performance. As regard, choosing the value of 0.5 for clustering radius is associated with the highest R^2 value (0.9425) for the test data (Fig. 6). After construction of the fuzzy model, well log data of entire sequence was exposed to the model in order to synthesize the TOC log (Fig. 6).

4.2. TOC Prediction using Δ log R technique

It is important to use a high **accuracy** method for prediction of TOC. In this section, a comparison is made between FL and empirical method known as Δ log R, relies on the separation of the resistivity and sonic (or density) curves to identify areas containing organic matter (Passey et al., 1990). The sonic log was used as a porosity log along with the resistivity log. Results show that despite the fact that the Δ log R method fails to predict TOC in the study formation, but it provides an approximate TOC trend when compared to measured TOC values (Fig. 6). In all, it can be concluded that the FL estimations are much more accurate than Δ log R method.

4.3. Grouping the TOC log by cluster analysis

In cluster analysis, a fundamental problem is to determine the best estimate of the number of clusters, which is usually taken as a prior in most clustering algorithms. Choice of optimal cluster number generally is known as cluster validity (Bezdek, 1973) and assessed using different types of validity measurements. Various measurements are available for cluster validation including so-called internal validity indices, which evaluate the goodness of a data partition using only quantities and features inherited from the dataset (Jain et al., 1999). In this paper, we analyze the labeling strategy based on the application of Compactness (Nguyen and Caruana, 2007), Davies-Bouldin (Davies and Bouldin, 1979), and Silhouette value (Rousseeuw, 1987). In this dissertation, we tackle the choice of optimal cluster numbers utilizing evaluation graph where the y-axis values can be any evaluation metric, such as distance, similarity, error, or quality. Accordingly, Compactness, Davies-Bouldin and Silhouette values versus a number of clusters were used to obtain the optimal number of clusters by means of the knee method that finds the “knee” in a number of clusters vs. clustering evaluation graph.

4.3.1. Evaluation methods of cluster analysis

Compactness (CP)

It is one of the commonly used measurement criteria, which employ only the information inherent to the dataset. According to the description given by Nguyen and Caruana (2007), CP measures the average distance between every pair of data points, which belong to the same cluster. More precisely, it is defined as

$$\mathrm{CP}(\pi^*) = \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{n_k} \left(\frac{\sum_{\mathrm{x_i}, \mathrm{x_j} \in \mathrm{C_k}} \mathrm{d}(\mathrm{x_i}, \mathrm{x_j})}{(\mathrm{n_k}\,(\mathrm{n_k}-1))/2} \right) \qquad (7)$$

where K denotes the number of clusters in the clustering result, $\mathrm{n_k}$ is the number of data points belonging to the Kth cluster, $\mathrm{d}(\mathrm{x_i}, \mathrm{x_j})$ is the distance between data points $\mathrm{x_i}$ and $\mathrm{x_j}$, and N is the total number of data points in the dataset. Ideally, the members of each cluster should be as close to each other as possible. Thus, lower value of CP means better cluster configuration.

Davies-Bouldin (DB)

The DB index makes use of similarity measure $\mathrm{R_{ij}}$ between the clusters $\mathrm{C_i}$ and $\mathrm{C_j}$, which is defined upon a measure of dispersion ($\mathrm{s_i}$) of a cluster $\mathrm{C_i}$ and a dissimilarity measure between two clusters ($\mathrm{d_{ij}}$). According to Davies and Bouldin (1979), $\mathrm{R_{ij}}$ is formulated as

$$\mathrm{R_{ij}} = \frac{\mathrm{s_i} + \mathrm{s_j}}{\mathrm{d_{ij}}}, \qquad (8)$$

where $\mathrm{d_{ij}}$ and $\mathrm{s_i}$ can be estimated by the following equations. Note that ν_x denotes the center of the cluster $\mathrm{C_x}$ and $|\mathrm{C_x}|$ is the number of data points in the cluster $\mathrm{C_x}$.

$$\mathrm{d_{ij}} = \mathrm{d}(\nu_i, \nu_j) \qquad (9)$$

$$\mathrm{s_i} = \frac{1}{|\mathrm{C_i}|} \sum_{\forall \mathrm{x} \in \mathrm{C_i}} \mathrm{d}(\mathrm{x}, \nu_i) \qquad (10)$$

Following that, the DB index is defined as

$$DB(\pi^*) = \frac{1}{k}\sum_{i=1}^{k} R_i, \qquad (11)$$

Where $R_i = \max_{j=1\ldots k, i\neq j} R_{ij}$. (12)

The DB index measures the average of similarity between each cluster and its most similar one.

Silhouette index

Silhouette width is defined as the average value of all observations' Silhouette values (Brock et al., 2008; Rousseeuw, 1987). This has the value between -1 and 1: a value near 1 indicates that the point partitioned to the right cluster. The Silhouette width of observation i is defined as follow (Rousseeuw, 1987):

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (13)$$

Where $a(i)$ is the distance measured as follow

$$a(i) = d(i, A) \qquad (14)$$

This is defined as the average distance between i and all other observations in cluster A. Also, $b(i)$ is the average distance of i to the observations in the nearest neighbor cluster.

The mean of the silhouette widths for a given cluster Ck is called the cluster mean silhouette and is denoted as ς_k:

$$\varsigma_k = \frac{1}{n_k}\sum_{i\in I_k} s(i) \qquad (15)$$

Finally, the global silhouette index (Petrović, 2006) is the mean of the mean silhouettes through all the clusters given by:

$$S = \frac{1}{K}\sum_{k=1}^{K} \varsigma_k \qquad (16)$$

both a cluster's silhouette and the global silhouette take values between -1 and 1.

Supervised learning attempts to assign each input value to one of a given set of classes. Therefore, prior to classification task, class membership have to be known. Consequently, K-mean algorithm utilized to automatic data labeling. In regard to above mentioned discussion about the optimal number of cluster determination, clustering process was accomplished gradually with cluster number from 2 to 100 (with 1 intervals) to achieving optimal cluster number by means of the knee method that finds the "knee" in a number of clusters vs. Compactness, Davies-Bouldin and Silhouette values (Fig. 7). The results depict that taking the clustering number of 4 leads to the best data labeling and pattern segregation through TOC values (Fig. 8). TOC values tend to increase from zone 1 to zone 4. The zone 1 contains a pile of sediments with TOC contents in the range of 0.42–2.1 wt%, deposited in the little more oxygenated environment. The zone 2, 3 and 4 is characterized by TOC quantities in the range of 2.1–3.56, 3.56-5 and 5-6.5 wt%, respectively.

4.4. Prediction of source rock zones

So far, only exploratory data analysis techniques, cluster analysis, have been discussed. These techniques attempt to analyze data without directly using information about the class assignment of the samples. Although cluster analysis is powerful methods for uncovering relationships in large multivariate data sets, they are not sufficient for developing a classification rule that can accurately predict the class membership of an unknown sample. Accordingly, two different algorithms, SVM and ANN applied to predict the sequence of class membership for the Kazhdumi formation from conventional well log data. Firstly, a three-layered Feed-Forward network was designed using MATLAB software. The network was trained with a number of the neurons in the hidden layer from 1 to 50 trained to achieving the best model. Totally 50 ANN model were established, which shows that taking the number of the neurons in the hidden layer of 35 leads to the highest performance (Fig. 9). The transfer function chosen to be TANSIG from layer one to two and PURLIN from layer two to layer three. The aforementioned three-layered ANN was trained by using the Levenberg Marquardt training algorithm (TrainLM), which details of their computation process and training can be found in Bishop (1995) and Boadu (1997, 1998). A common setup to fairly measure classification performance is by dividing the total number of data into a training set and a test set. The test set provides a

completely independent measure of network accuracy. A trade-off should be taken to divide the dataset into a training set and a test set, e.g. 2/3 of the dataset are used as training data while the rest is used later as test data. The test data are only used for the training by means of which an independent classification performance can be measured. The performance of the model was measured by counting the number of correctly labeled data points, in comparison to their known class labels, and dividing by the total number of data in the dataset (Classification Accuracy (CA)) (Nguyen and Caruana, 2007), which was applied to set the optimal parameters. Maximizing CA is a key criterion in selecting estimators. The CA for optimal ANN model was 0.9077. In the following step, we performed our comparative study by applying SVM classifier as a class-membership predictive model. In this paper, the RBF kernel has been adopted in SVM implementation which is performing quite well in comparison to other kernels (Bhattacharya et al., 2016; Camps-Valls and Bruzzone, 2009; Hsu et al., 2003). The other important parameters for the SVM classifier include the coefficient of RBF kernel γ and σ^2. These are determined via cross validation by selecting various combinations of the parameters values. Optimal values of γ and σ^2 that yields the maximum overall classification accuracy are 2.1201 and 0.60481, respectively. The CA for SVM model was 0.9369, which outperforms the ANN model. Consequently, taking the CA into account SVM have shown better results than those from ANN (Fig. 10). A significant advantage of SVMs is that whilst stochastic gradient descent isn't guaranteed to find the optimal set of parameters when implemented in NN, any decent SVM implementation is going to find the optimal set of parameters. So, ANNs can suffer from the existence of multiple local minima solutions in comparison with SVMs that finds a global and unique solution (Suykens et al., 2002). Simple geometric interpretation and giving a sparse solution are two more advantages of SVMs. Unlike ANNs, the computational complexity of SVMs does not depend on the dimensionality of the input space. The reason that SVMs often outperform ANNs in practice is that they are less sensitive to the presence of outliers in the training set and hence they are less prone to overfitting (Olson et al., 2008). Furthermore, the performance of the SVMs is slightly affected by training data set size which is an advantage for it as a machine learning method (Sebtosheikh and Salehi, 2015).

5. Conclusions

In general, estimation of TOC from well log data plays important roles in early diagnosis of whether source rock is present in the area of interest or not. The TOC manipulates well log responses variable. Nevertheless, relating the well log data to TOC within the formation, it is possible to estimate the weight percent of TOC present. Thus, FL was used to extract this relationship in order to estimate TOC values from well log data. The results were validated with the measured geochemical data from Rock-Eval pyrolysis (R^2 values of 1 and 94.25% for training and testing sets, respectively). The Passey Δ log R method was unsuccessful in predicting TOC when compared to measured TOC values. The TOC content of the Kazhdumi formation was grouped optimally using K-means cluster analysis into four classes by means of the knee method that finds the "knee" in a number of clusters vs. Compactness, Davies-Bouldin and Silhouette values. TOC values respectively tend to increase from zone 1 to zone 4. Thus, assuming TOC content the Kazhdumi formation has been divided into 4 zones. In this paper for developing a classification rule that can accurately predict the class membership of an unknown sample, the performances from SVM and ANN are evaluated in the classification of well log data. The proposed method is found effective in specifying the TOC zonation from well log data. Quantitative comparisons of the results from ANN and SVM depicts that for classification problem of source rock zonation SVM with RBF kernel readily outperforms ANN in term of classification accuracy (0.9077 and 0.9369 for ANN and SVM, respectively). The final step in every study is to recognize patterns. Exploiting supervised learning enables us to extract these patterns directly from well log data without the need to do the tedious task of regression and clustering analysis for other wells. Also, this method would enable a more elaborate assessment of formation to be undertaken by providing a comprehensive quick look results.

Acknowledgements

The authors wish to acknowledge the National Iranian Oil Company (NIOC) for funding and logistical support for this project.

References

Abdollahie Fard, I., Braathen, A., Mokhtari, M., Alavi, A., 2006. Interaction of the Zagros fold-thrust belt and Arabian-type, deep-seated folds in the Abadan plain and Dezful embayment, SW Iran, Petroleum Geosciences 12 (4), 347-362.

Alizadeh, B., Najjari, S., Kadkhodaie-Ilkhchi, A., 2012. Artificial neural network modeling and cluster analysis for organic facies and burial history estimation using well log data: a case study of the South Pars gas field, Persian Gulf, Iran. Comput. Geosci. 45, 261–269.

Ayesha, S., Mustafa, T., Raza Sattar, A. & Inayat Khan, M., 2010. "Data Mining Model for Higher Education System ",European Journal of Scientific Research, ISSN 1450-216X Vol.43 No.1 , pp.27.

Bezdek, J.C., 1973. Cluster validity with fuzzy sets, Journal of Cybernetic, vol. 3, no. 3, pp. 58-73.

Bhattacharya, S., Carr, T.R. and Pal, M., 2016. Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA. Journal of Natural Gas Science and Engineering.

Bishop, C.M., 1995.Neural Networks for Pattern Recognition. Clarendon Press, Oxford, pp. 670.

Boadu, F.K, 1998. Inversion of fracture density from field seismic velocities using artificial neural networks Geophysics 63 534–45

Boadu, F.K., 1997. Rock properties and seismic attenuation: neural network analysis. Pure and Applied Geophysics 149, 507–524.

Bolandi, V., Kadkhodaie-Ilkhchi, A., Alizadeh, A., Tahmorasi, J. and Farzi, R., 2015. Source rock characterization of the Albian Kazhdumi formation by integrating well logs and geochemical data in the Azadegan oilfield, Abadan plain, SW Iran, Journal of Petroleum Science and Engineering 133, 167–176.

Bordenave, M.L., 2002. The Middle Cretaceous and Early Miocene petroleum system in the Zagros domain of Iran and its prospect evaluation. In: AAPG Annual Meeting, Houston, American Association of Petroleum Geologists, pp. 1–9.

Bordenave, M.L., Burwood, R., 1990. Source rock distribution and maturation in the Zagros orogenic belt: provenance of the Asmari and Bangestan reservoir oil accumulations. Organic Geochemistry 16, 369–387.

Bordenave, M.L., Huc, A.Y., 1995. The Cretaceous source rocks in the Zagros foothills of Iran: an example of a large size intra-cratonic basin. Revue de l'Institut Français du Pétrole 50, 527–753.

Brock, G., Pihur, V., Datta, S., and Datta, S., 2008. clValid: An R Package for Cluster Validation, Journal of Statistical Software, Vol. 25, Iss. 4, 2008, pp. 1-20.

Camps-Valls, G. and Bruzzone, L. eds., 2009. Kernel methods for remote sensing data analysis (Vol. 2). New York: Wiley.

Carpentier, B., Huc A.Y., and Besserau, G., 1989. Wireline logging and source rocks: estimation of organic carbon contents bye the CARBOLOG method. Rev. Inst. Fr. Pet. 44, 669-719.

Chiu, S., 1994. Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy Systems 2, 267–278.

Davies, D.L., Bouldin, D.W., 1979. "A Cluster Separation Measure." IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224-227.

Dellenbach, J., Espitalie, J., Lebreton, F., 1983. Source Rock Logging: Transactions of 8th European SPWLA Symposium, paper D

Edgell, H.S., 1996. Salt tectonism in the Persian Gulf Basin. Geological Society, London, Special Publications, 100, 129-151.

Espitalie´, J., Madec, M., Tissot, B., Menning, J.J., Leplat, P., 1977. Source rock characterization method for petroleum exploration .Proceedings 9th Annual Offshore Technology Conference 3,439–448.

Fertle, H., 1988. Total organic carbon content determined from well logs: SPE Formation Evaluation 15612, pp. 407– 419.

Fertle, H., and Rieke, H., 1980. Gamma-ray spectral evaluation techniques identify fractured shale reservoirs and source rock characteristics: Journal of Petroleum Technology, v. 31, pp. 2053-2062.

Guillaume, S., 2001. Designing Fuzzy Inference Systems from Data: An Interpretability-Oriented Review. IEEE Transections on Fuzzy Systems 9(3), 426-443

Herron, S.L., 1988. Source rock evaluation using geochemical information from wireline logs and cores (abs): AAPG Bulletin, v. 72, 1007.

Hsu, C.W., Chang, C.C. and Lin, C.J., 2003. A practical guide to support vector classification.

Huang, Z., Williamson, M.A., 1996. Artificial neural network modeling as an aid to source rock characterization. Marine and Petroleum Geology 13 (2), 227–290.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. "Data Clustering: A Review." ACM Computing Survey, 31(3), 264-323.

James, G.S., Wynd, J.G., 1965. Stratigraphic nomenclature of Iranian Oil Consortium Agreement area. Am. Assoc. Pet. Geol., 49(12), P. 2182-2245.

Kadkhodaie, A., Rezaee, M.R., Moallemi, S.A., 2006. A fuzzy logic approach for the estimation of permeability and rock types from conventional well log data: an example from the Kangan reservoir in Iran Offshore Gas Field, Iran. Journal of Geophysics and Engineering, UK, Vol. 3, pp. 356-369.

Kadkhodaie, A., Takahashi Monteiro, S., Ramos, F., Hatherly, P., 2010. Rock Recognition from MWD Data: A Comparative Study of Boosting, Neural Networks and Fuzzy Logic. IEEE Transactions on Geosciences and Remote Sensing Letters (GSRL), vol. 7, No. 4, 680-684.

Kadkhodaie-Ilkhchi, A., Rahimpour-Bonab, H., Rezaee, M., 2009. A committee machine with intelligent systems for estimation of total organic carbon content from petrophysical data: An example from Kangan and Dalan reservoirs in South Pars Gas Field, Iran. Computers & Geosciences, 35(3), 459-474.

Kamali, M. R., Allah Mirshady, A., 2004. Total organic carbon content determined from well logs using ΔlogR and neuro fuzzy techniques. Journal of Petroleum Science and Engineering, 45(3), 141-148.

Khedairia, S., Khadir, M.T., 2008. Self-Organizing Map and K-Means for Meteorological Day Type Identification for the Region of Annaba-Algeria. In Computer Information Systems and Industrial Management Applications. CISIM'08. 7th (pp. 91-96). IEEE.

Konyuhov, A.I., Maleki, B., 2006. The Persian Gulf Basin: geological history, sedimentary formations, and petroleum potential. Lithology and Mineral Resources 41, 344–361.

Lafargue, E., Marquis, F., Pillot, D., Bernard, M., Beauducel, G., Antonas, R., Burwood, R., 1997. Rock-Eval 6: a new generation of Rock-Eval pyrolyser for a wider use in petroleum exploration/production and in soil contamination studies. AAPG International Conference and Exhibition; Abstracts. American Association of Petroleum Geologists, Tulsa 81; 8: 1393.

Langford, F.F., Blanc-Valleron, M.M., 1990. Interpreting Rock-Eval pyrolysis data using graphs of pyrolizable hydrocarbons vs. total organic carbon. American Association of Petroleum Geologists, Bulletin 6 (74), 799–804.

MacQueen, J.B., 1967. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press, pp. 281–297

Magoon, L.B., Dow, W.G., 1994. The petroleum system-from source to trap. AAPG Memoir 6. Tulsa, American Association of Petroleum Geologist, pp. 25-49.

Matlab user's Guide 2012. Neural Network, Fuzzy Logic and Direct Search toolboxes, Matlab CD-ROM, by the Mathworks, Inc.

Meissner, F.F., 1978. Petroleum geology of Bakken Formation, Williston basin, North Dakota and Montana. The economic geology of the Williston basin: Montana geological society, Williston Basin Symposium, pp. 207–227.

Mendelson, J.D., Toksoz, M.N., 1985, Source rock characterization using multivariate analysis of log data, Transactions of the Society of Professional Well Log Analysts 26th Annual Logging Symposium, Paper UU, pp. 1-21.

Meyer, B.L., Nederlof, M. H., 1984. Identification of source rocks on wireline logs by density/resistivity and sonic transit time/resistivity cross plots. American Association of Petroleum Geologists, Bulletin 68, 121–129.

Morris, P., 1977. Basement structure as suggested by aeromagnetic surveys in S W Iran. Second Geological Symposium of Iran, Tehran: Iranian Petroleum Institute.

Motiei, H., 1995. Petroleum Geology of Zagros. Publication of the Geological Survey of Iran, 589 p. (in Persian)

Nguyen, N., Caruana, R., 2007. “Consensus Clusterings." In Proceedings of IEEE International Conference on Data Mining, pp. 607-612. IEEE Computer Society, Washington, DC.

Nixon, R.P., 1973. Oil source beds in Cretaceous Mowry shale of north western interior United States. AAPG

Olson, D.L. and Delen, D., 2008. Advanced data mining techniques. Springer Science & Business Media. Bull. 52, 136–161.

Passey, O.R., Moretti, F.U., Stroud, J.D. 1990. A practical modal for organic richness from porosity and resistivity logs. American Association of Petroleum Geologists Bulletin 74, 1777–1794.

Peters, K.E., Cassa, M.R., 1994. Applied source rock geochemistry. The Petroleum System From Source to Trap (L.B. Magoon and W. G. Dow, eds.), American Association of Petroleum Geologists, Tulsa, OK, pp. 93–117.

Petrović, S., 2006. A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In Proceedings of the 11th Nordic Workshop of Secure IT Systems, pp. 53-64.

Rousseeuw, P. J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Schmoker, J.W., 1979. Determination of Organic Content of Appalachian Devonian Shales from Formation Density Logs. AAPG Bulletin, Vol. 63, pp. 1504-1537.

Schmoker, J.W., 1981. Determination of Organic-matter content of Appalachian Devonian shales from gamma-ray logs: AAPG Bulletin, v. 56, p. 1285-1298.

Schmoker, J.W., Hester, T.C., 1983, Organic carbon in Bakken Formation, United States portion of Williston Basin. American Association of Petroleum Geologists Bulletin 67, pp. 2165–2174.

Schmoker, J.W., Hester, T.C., 1989. Oil generation inferred from formation resistivity- Bakken Formation, Williston basin, North Dakota. Transactions of the Thirtieth SPWLA Annual Logging Symposium, paper H.

Sebtosheikh, M.A. and Salehi, A., 2015. Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance in a heterogeneous carbonate reservoir. Journal of Petroleum Science and Engineering, 134, pp.143-149

Smola, A.J., Schollköpf, B., 1998. A tutorial on support vector regression, NeuroCOLT2 Technical Report Series NC2-TR-1998-030, ESPRIT work- ing group on Neural and Computational Learning Theory, NeuroCOLT2.

Suykens, J.A.K., and Vandewalle, J., 1999. Least Squares Support Vector Machine Classifiers, Neural Processing Letters, 9: 29 3-300.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines, World Scientific Publishing Co., Singapore.

Takagi, T., Sugeno, M., 1985. Identification of systems and its application to modeling and control. IEEE Transaction on Systems, Man and Cybernetics 15, 116–132.

Vapnik, V., Chervonenkis, A., 1964. A note on class of perceptron, Automation and Remote Control, 24.

Vapnik, V. and Lerner, A., 1963, Pattern recognition using generalized portrait method, Automation and Remote Control, 24.

Zadeh, L.A., 1965. Fuzzy Sets. Information and Control. 8: 338-353.

Figure captions

Fig. 1. Illustration of the study area in the Persian Gulf shown by hatched area.

Fig. 2. Geochemical log based on Rock-Eval pyrolysis for Kazhdumi formation.

Fig. 3. Cross-plots showing relationship between measured TOC content and DT (a), GR (b), LLD (c), LLS (d), NPHI (e), RHOB (f), URAN (g), POTA (h), and THOR (i).

Fig. 4. Measured TOC on top of the logs for showing log correlations.

Fig. 5. Schematic illustration of SVM for separation hyper-plane in two-dimensional case (hyper-plane is a red)

Fig. 6. Comparison of FL and Δ log R predicted TOC (wt.%), Crossplot showing correlation coefficient between and FL and Δ log R predicted TOC with measured TOC values.

Fig. 7. Finding the number of clusters using the Knee Method.

Fig. 8. Correlation between well logs and FL predicted TOC (wt.%) and results from the cluster analysis (TOC zones).

Fig. 9. Finding the number of neurons in the ANN model.

Fig. 10. Display of results from SVM and ANN.

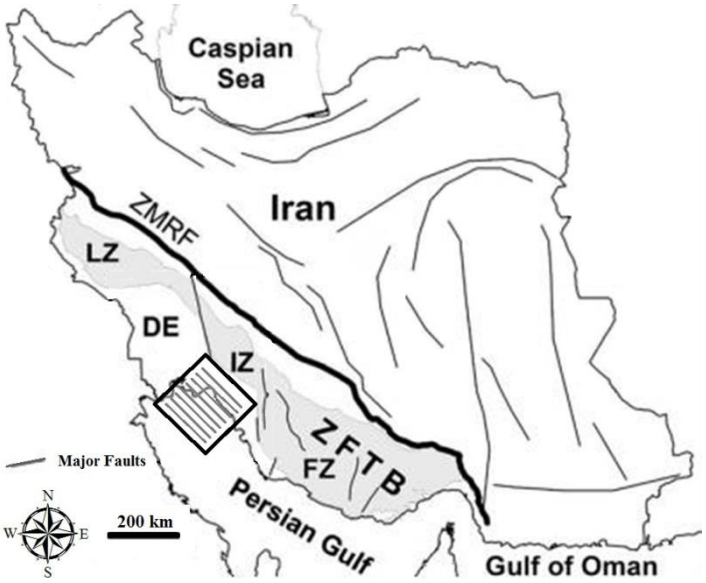

Caspian
Sea
Iran
ZMRF
LZ
DE
IZ
ZFTB
FZ
Major Faults
N
W
E
S
200 km
Persian Gulf
Gulf of Oman

Fig. 1

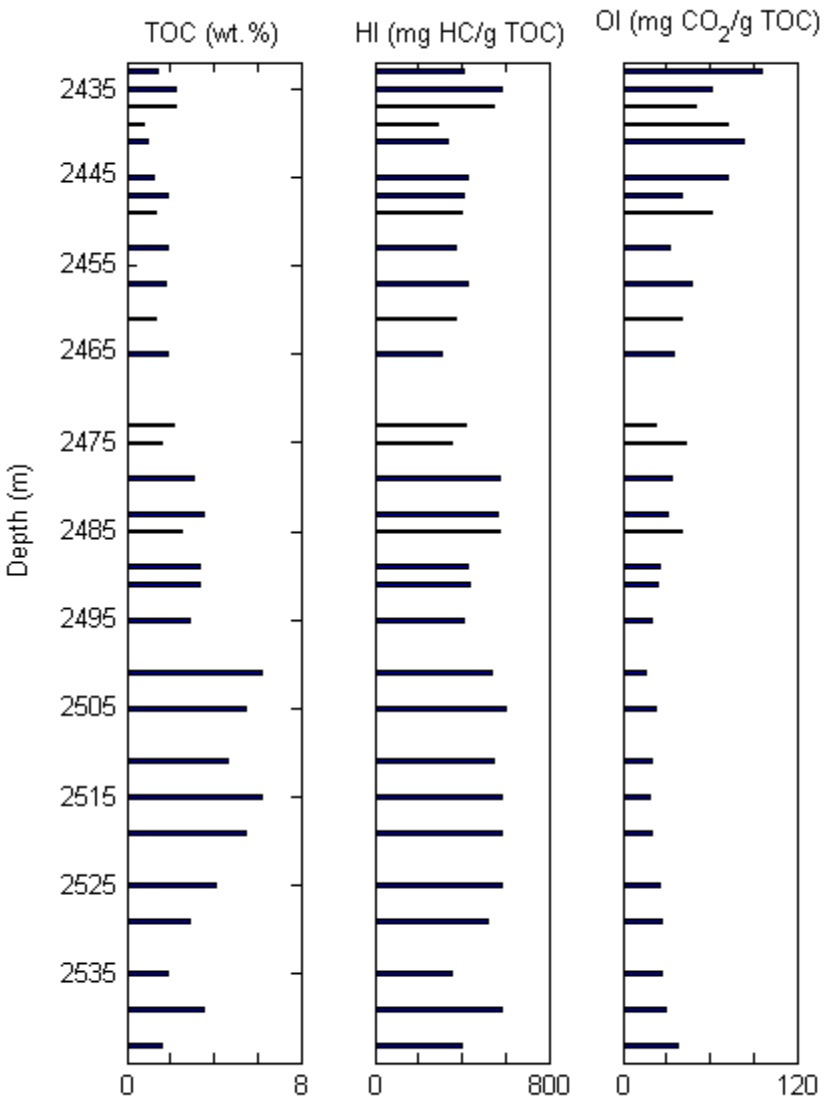

TOC (wt.%)
HI (mg HC/g TOC)
OI (mg CO2/g TOC)
Depth (m)
2435
2445
2455
2465
2475
2485
2495
2505
2515
2525
2535
0
8
0
800
0
120

Fig. 2

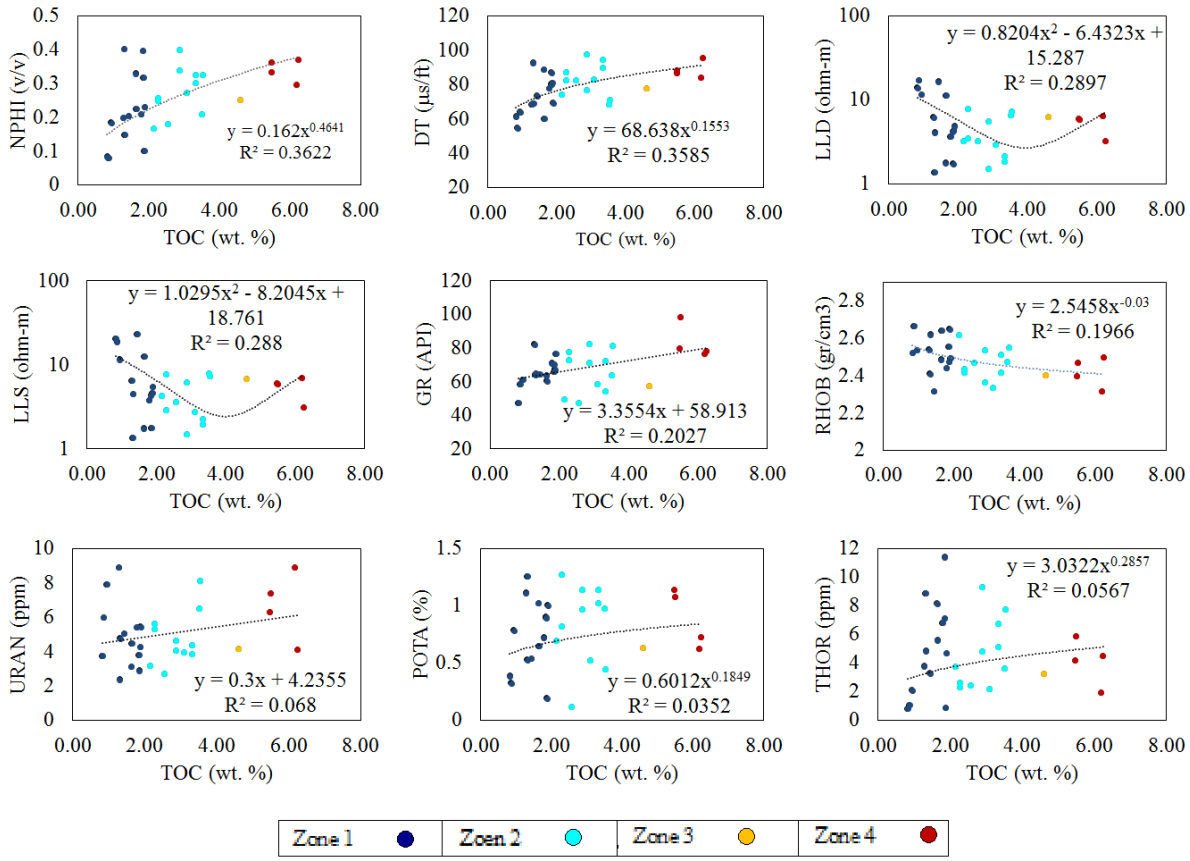

NPHI (v/v)
y = 0.162x^0.4641
R² = 0.3622
TOC (wt. %)
DT (µs/ft)
y = 68.638x^0.1553
R² = 0.3585
TOC (wt. %)
LLD (ohm-m)
y = 0.8204x² - 6.4323x + 15.287
R² = 0.2897
TOC (wt. %)
LLS (ohm-m)
y = 1.0295x² - 8.2045x + 18.761
R² = 0.288
TOC (wt. %)
GR (API)
y = 3.3554x + 58.913
R² = 0.2027
TOC (wt. %)
RHOB (gr/cm3)
y = 2.5458x^-0.03
R² = 0.1966
TOC (wt. %)
URAN (ppm)
y = 0.3x + 4.2355
R² = 0.068
TOC (wt. %)
POTA (%)
y = 0.6012x^0.1849
R² = 0.0352
TOC (wt. %)
THOR (ppm)
y = 3.0322x^0.2857
R² = 0.0567
TOC (wt. %)
Zone 1
Zoen 2
Zone 3
Zone 4

Fig. 3

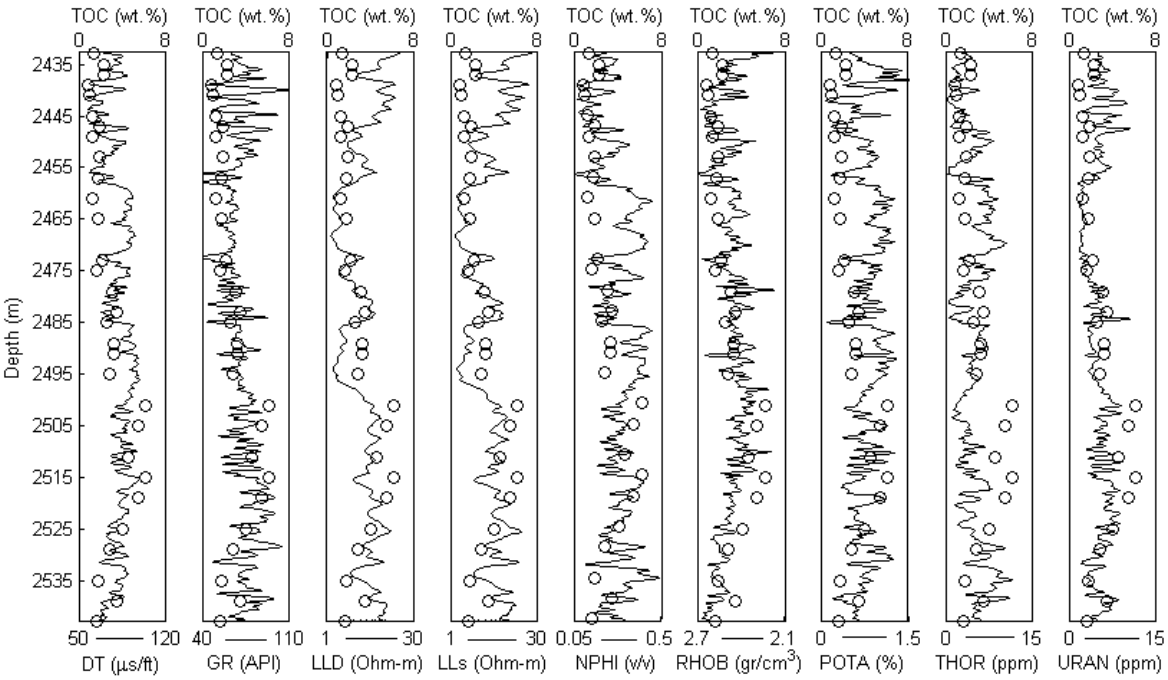

TOC (wt.%)
0
8
Depth (m)
2435
2445
2455
2465
2475
2485
2495
2505
2515
2525
2535
50 — 120
DT (µs/ft)
40 — 110
GR (API)
1 — 30
LLD (Ohm-m)
1 — 30
LLs (Ohm-m)
0.05 — 0.5
NPHI (v/v)
2.7 — 2.1
RHOB (gr/cm3)
0 — 1.5
POTA (%)
0 — 15
THOR (ppm)
0 — 15
URAN (ppm)

Fig. 4

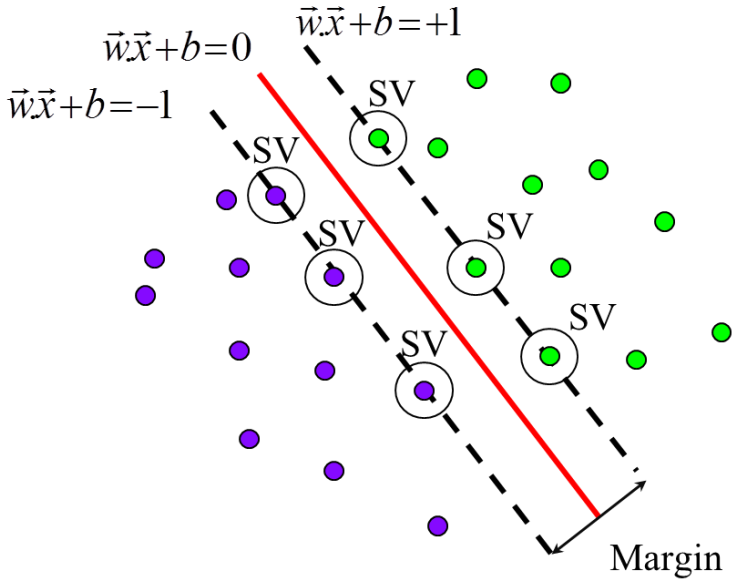
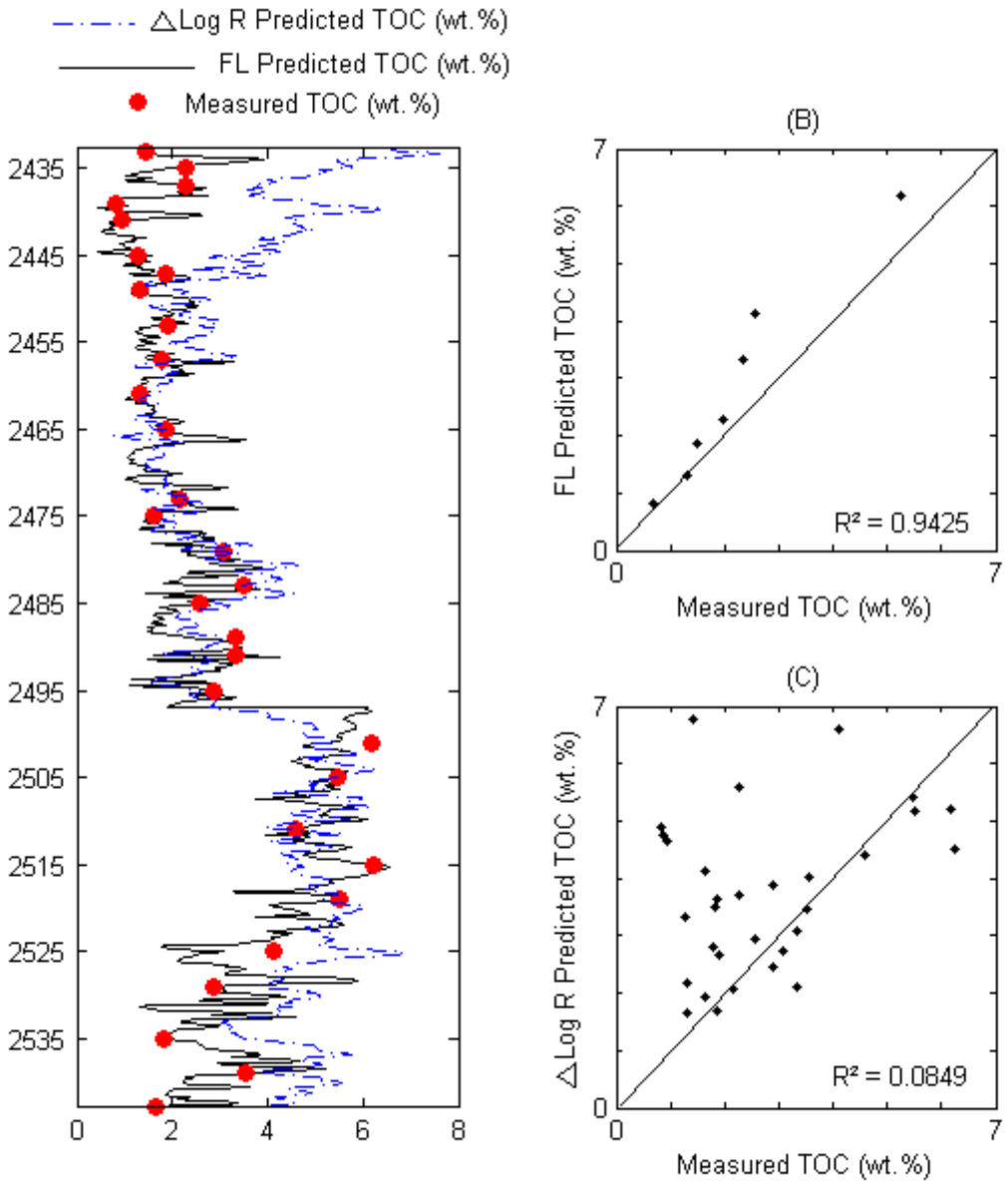


Fig. 5


△Log R Predicted TOC (wt.%)
FL Predicted TOC (wt.%)
Measured TOC (wt.%)
2435
2445
2455
2465
2475
2485
2495
2505
2515
2525
2535
0
2
4
6
8
(B)
FL Predicted TOC (wt.%)
R² = 0.9425
Measured TOC (wt.%)
(C)
△Log R Predicted TOC (wt.%)
R² = 0.0849
Measured TOC (wt.%)

a

Fig. 6

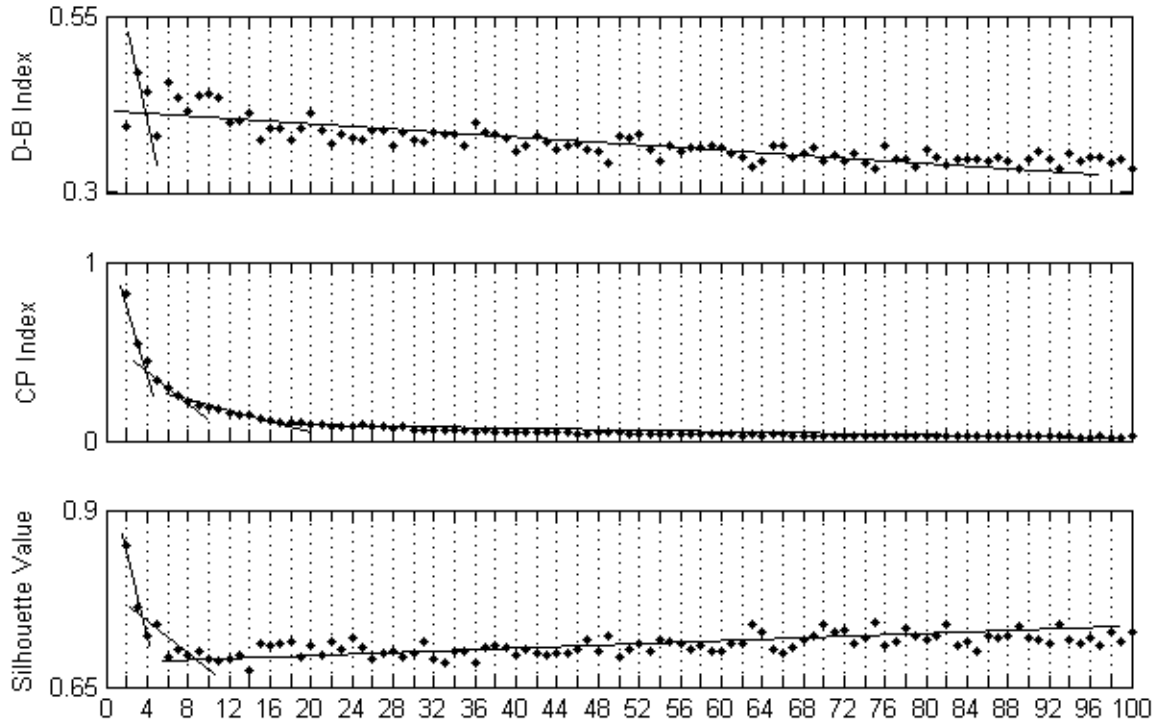

D-B Index
0.55
0.3
CP Index
1
0
Silhouette Value
0.9
0.65
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100

Fig. 7

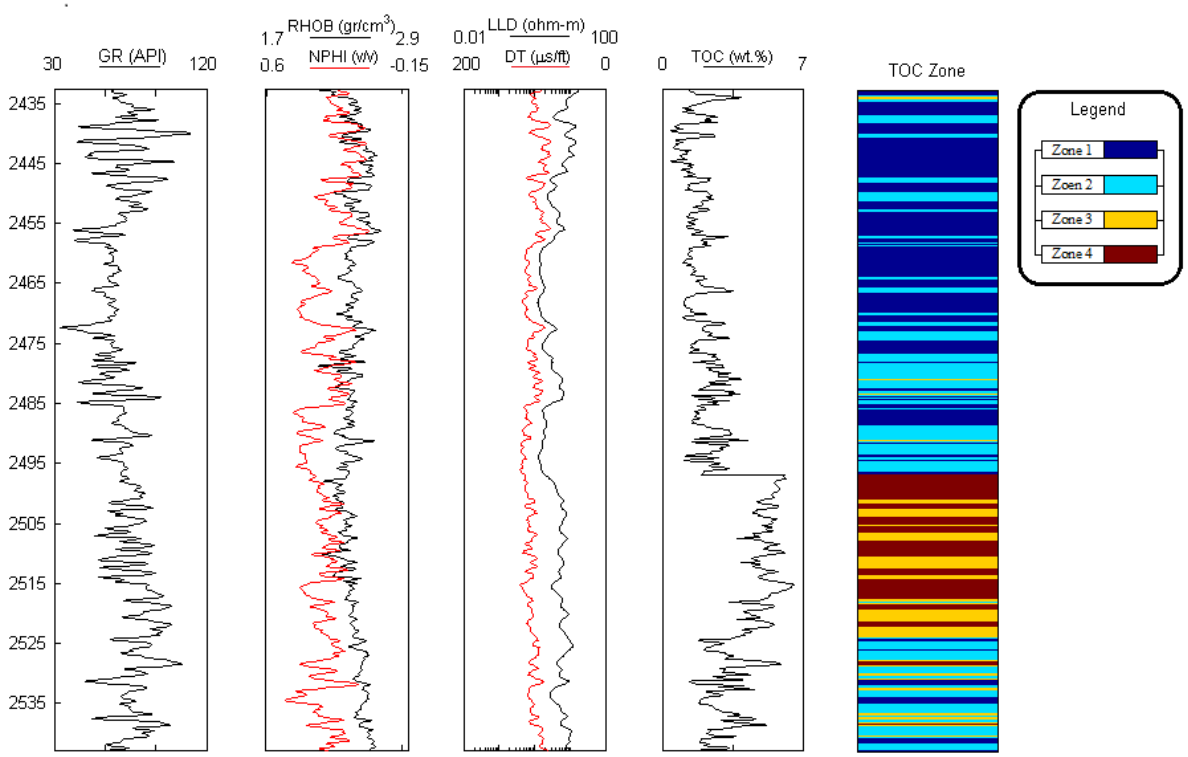

30
GR (API)
120
1.7
RHOB (gr/cm3)
2.9
0.6
NPHI (v/v)
-0.15
0.01
LLD (ohm-m)
100
200
DT (µs/ft)
0
0
TOC (wt.%)
7
TOC Zone
Legend
Zone 1
Zoen 2
Zone 3
Zone 4
2435
2445
2455
2465
2475
2485
2495
2505
2515
2525
2535

Fig. 8

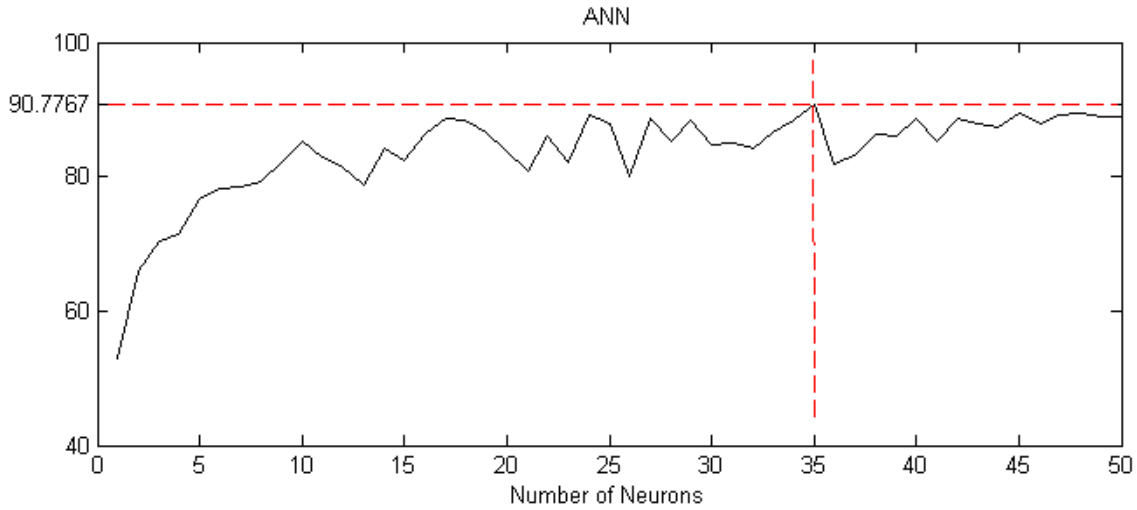

ANN
100
90.7767
80
60
40
0
5
10
15
20
25
30
35
40
45
50
Number of Neurons

Fig. 9

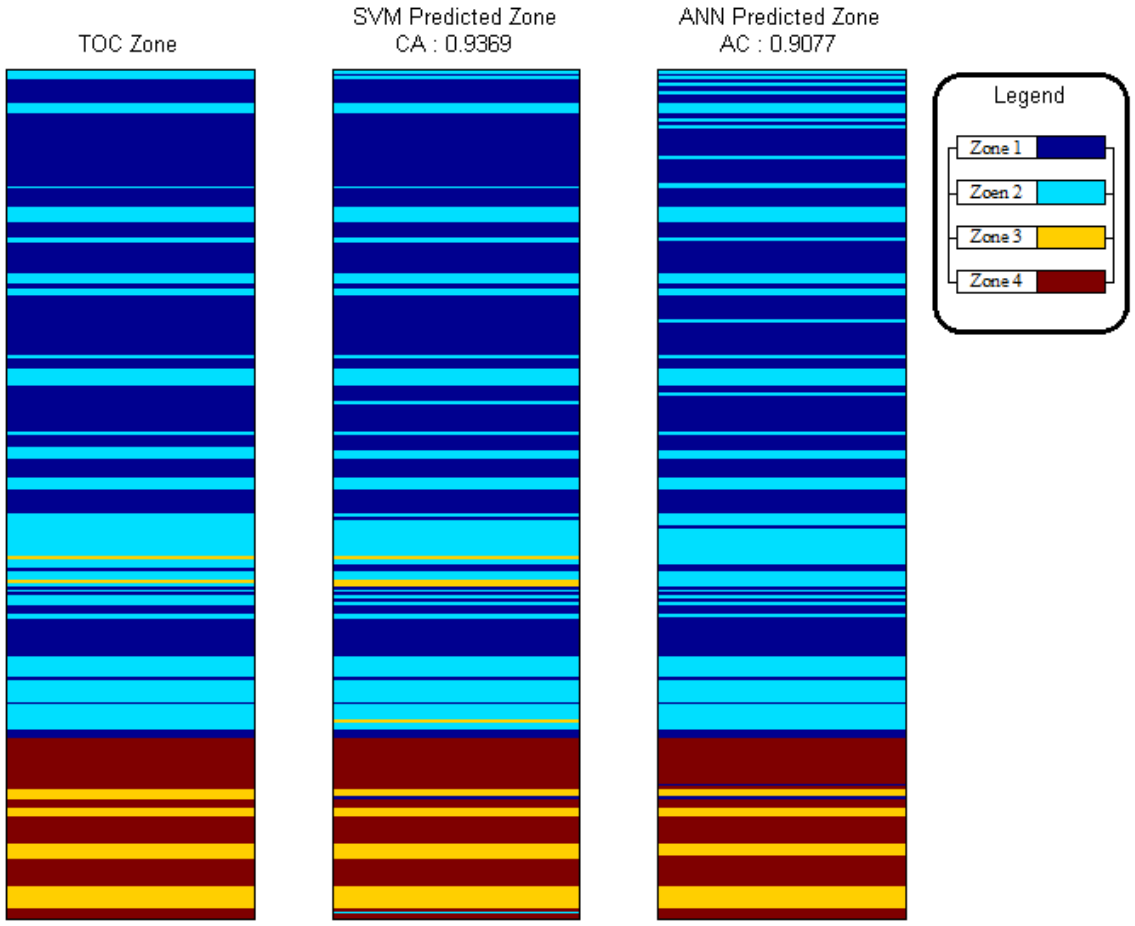

TOC Zone
SVM Predicted Zone
CA : 0.9369
ANN Predicted Zone
AC : 0.9077
Legend
Zone 1
Zoen 2
Zone 3
Zone 4

Fig. 10