Computers & Geosciences 57 (2013) 175–182

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Facies recognition using a smoothing process through Fast Independent Component Analysis and Discrete Cosine Transform

Alexandre Cruz Sanchetta [a,*], Emilson Pereira Leite [b], Bruno César Zanardo Honório [b]

[a] Rua Mendeleiev, s/n, Cidade Universitária "Zeferino Vaz", Barão Geraldo, Campinas, São Paulo 13083-970, Brasil
[b] Rua João Pandiá Calógeras, 51, Cidade Universitária "Zeferino Vaz", Barão Geraldo, Campinas, São Paulo 13983-970, Brasil

ARTICLE INFO

Article history:
Received 6 June 2012
Received in revised form
14 March 2013
Accepted 25 March 2013
Available online 3 April 2013

Keywords:
Discrete Cosine Transform
Independent Component Analysis
Automatic classification
Reservoir characterization

ABSTRACT

We propose a preprocessing methodology for well-log geophysical data based on Fast Independent Component Analysis (FastICA) and Discrete Cosine Transform (DCT), in order to improve the success rate of the K-NN automatic classifier. The K-NN have been commonly applied to facies recognition in well-log geophysical data for hydrocarbon reservoir modeling and characterization.

The preprocess was made in two different levels. In the first level, a FastICA based dimenstion reduction was applied, maintaining much of the information, and its results were classified; In second level, FastICA and DCT were applied in smoothing level, where the data points are modified, so individual points have their distance reduced, keeping just the primordial information. The results were compared to identify the best classification cases. We have applied the proposed methodology to well-log data from a petroleum field of Campos Basin, Brazil. Sonic, gamma-ray, density, neutron porosity and deep induction logs were preprocessed with FastICA and DCT, and the product was classified with K-NN. The success rates in recognition were calculated by appling the method to log intervals where core data were available. The results were compared to those of automatic recognition of the original well-log data set with and without the removal of high frequency noise. We conclude that the application of the proposed methodology significantly improves the success rate of facies recognition by K-NN.

1. Introduction

Well-log data have been used in many areas of geological and geophysical data analysis, such as in reservoir characterization where models of subsurface properties that take into account details about rock physics and the fluids contained in the rocks are constructed (Avseth et al., 2005; Coconi-Morales et al., 2010; Doyen, 2007; Dubrule, 1994). Another example is the use of well-log data to predict seismic parameters related to Amplitude vs. Offset (AVO) data such as Vp, s, Poisson's ratio (σ), among others, that aid in the comprehension of reservoirs (Rutheford and Willians, 1989). Well-log data has also been widely used in structural and stratigraphic mapping.

In order to connect the well-log data with other geological or geophysical information, it is important to correlate them with lithofacies described from core samples. However, such direct information is often not available to the entire length of the wells, mainly because of financial restrictions. Therefore, pattern recognition methods must be applied for prediction or classification.

To mention a few examples of application of pattern recognition methods in well-log data, Grana et al. (2012) had constructed a complete statistical workflow for obtaining petrophysical properties at the well location and the corresponding facies classification; Messina and Langer (2011) have implemented unsupervised algorithms based on self-organizing maps and cluster analysis to analyze and to interpret volcanic tremor data; Turlapaty et al. (2010) proposed a method based on wavelet-based feature extraction and one-class support vector machines to analyze satellite remote sensing data applied to soil moisture and vegetation mapping; and Rosati and Cardarelli (1997) applied texture features based on gray tone spatial dependence matrices to classify patterns observed on magnetic anomaly maps.

In fact, all types of data can be separated into several subsets, where each data element present in a random subset contains some information in common with the other elements in that subset. In other words, it is possible to classify all elements based on some common characteristics identified in the data set. This is the basic principle of automatic classification (MacQueen, 1967). The several approaches to automatic classification can be divided into two major groups: supervised methods and unsupervised methods (Duda and Hart, 1973; Mitchell, 1997; Schuerman, 1996).

* Correspondence to: Rua Doutor Geraldo de Campos Freire, 567, Barão Geraldo, Campinas, São Paulo 13083-480, Brasil. Tel.: +55 19 3287 6618 (residence), mobile: +55 19 8811 9176.
E-mail addresses: alexandr@dep.fem.unicamp.br (A.C. Sanchetta), emilson@ige.unicamp.br (E.P. Leite), brunohonorio@gmail.com (B.C.Z. Honório).

0098-3004/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2013.03.021

Supervised methods are based on the knowledge of classification labels (targets), i.e., it is known that an input sample corresponds to a certain label (e.g. K-Nearest Neighbor, Artificial Neural Networks). Unsupervised methods cluster the samples into subsets through statistical and other mathematical approaches (e.g. K-Means, Self-Organizing Maps). Supervised methods need a label to classify, in our case, the core data. The application of supervised methods are robust, once we can assume that heterogeneous environments are periodic for mathematical simplicity; and computations of local problems that should be done on a sufficiently large representative elementary volume R.E.V. instead of a single cell, because random flows have different results than periodic flows (Amaziane et al., 2006). In this work, we have employed the K-Nearest Neighbor (K-NN) method (Toussaint, 2005). K-NN is a pattern recognition method that classifies elements in a data set based on the spatial distribution of a training set.

Many parameters affect the quality of the prediction: choice of variables; quality of the measured data; and the associated uncertainty are just a few examples. For spatial classifiers such as K-NN, the perfect choice would be a set of orthogonal variables that do not carry any kind of redundant or misplaced information. However, in practice we often notice high correlations among different well-log profiles. For example, the well NA04, in Namorado Field, sonic log (DT) and neutron porosity log (NPHI) have cross-correlations higher than 0.8 and DT and bulk density log (RHOB) have cross-correlation higher than 0.7 (Carrasquilla and Leite, 2009). Those high correlations mean that the information is redundant, which may jeopardize the performance of the K-NN. To minimize this problem and to prepare the data for automatic classification we have applied a multivariate analysis, namely Independent Component Analysis (ICA) (Comon, 1994). ICA is a blind source separation method that transforms an input signal into a new set of independent signals. Statistically, independence means that the occurrence of a signal does not have any relation with the occurrence of another signal contained in the data set (Russell and Norvig, 2002). The independence of the signals provides an orthogonal set of new variables such that they can be understood as individual events. With the aim of reducing the computation effort, an iterative algorithm based on the Newton's method was developed by Hyvärinen (1999), namely The Fast Independent Component Analysis (FastICA). Independent variables are necessarily orthogonal, so the FastICA method provides orthogonality in the classification. In order to smooth data, we have also applied the Discrete Cosine Transform (DCT) (Rao and Yip, 1990).

In our analysis, we have used well-log data from Campos Basin, which is located in the southeastern portion of Brazil, along the north coast of the State of Rio de Janeiro. Campos Basin covers an area of 100.000 km^2 up to a water depth of 3000 m. Particularly, all wells are located in the Namorado Field, which is located on the north-central hydrocarbon accumulations zone of the Campos Basin, about 80 km offshore under water depths ranging from 140 m to 250 m (Vidal et al., 2007).

2. Fast Independent Component Analysis

Fast Independent Component Analysis (FastICA) is an optimized process of Independent Component Analysis (ICA, Comon, 1994). ICA is a method for finding underlying factors or components from multivariate data that are related to Blind Signal Separation (Hyvärinen et al., 2001). The usual ICA method have some problems such as maximization of contrast functions, computational cost and/or selection of the learning rate, while FastICA was proposed to remedy these problems and to converge to the result faster than ICA (Hyvärinen, 1999).

The FastICA process is based on a fixed-point algorithm, which is well-known for its speed. The requirement of such speed is justified by the large number of data points that are used in a single step of the algorithm (block model). The algorithm has some improvements when compared to common fixed-point algorithms (Burden and Faires, 1985). Besides, the algorithm is parallel, computationally simple and requires little computational memory, similarly to some neural algorithms (Hyvärinen et al., 2001).

The fixed-point algorithm is based on the Newton's method (Hyvärinen, 1999), which within the framework of the common ICA reads as

$$\mathbf{s}^{+} = \mathbf{s} - \frac{f(x)}{f'(x)}, \tag{1}$$

where $\mathbf{s}$ is a random non-gaussian initial guess vector which in turn will iteratively converge to $\mathbf{s}^{+}$ through some objective function $f(x)$. Commonly, objective fuction are filled with negentropy equations. Negentropy, the denial of entropy, is a measure that is maximized when the distributions are as far as possible from a Gaussian distribution (Comon, 1994).

In the case of FastICA, negentropy is used considering the stationarity conditions of Kuhn–Tucker

$$\nabla f(x^*) + \sum_{i=1}^{m} \mu_i \nabla \mathrm{g_i}(x^*) + \sum_{j=1}^{l} \lambda_j \nabla \mathrm{h_j}(x^*) = 0, \tag{2}$$

where x^* is a local minimum, μ_i and λ_j are constants called KKT multipliers, $\mathrm{g_i}$ are the inequality constraint functions, $\mathrm{h_j}$ are the equality constraint functions and $\nabla f(x^*)$, $\nabla \mathrm{g_i}(x^*)$ and $\nabla \mathrm{h_j}(x^*)$ are their gradients, respectively. These function are optimized obtained by (Hyvärinen, 1999)

$$E\{\mathbf{x}g(\mathbf{s}^T\mathbf{x})\} - \beta\mathbf{s} = 0, \tag{3}$$

where $\beta = E\{\mathbf{s}_0^T\mathbf{x}g(\mathbf{s}_0^T\mathbf{x})\}$ with s_0 being the optimal value from the process.

Newton's method requires the derivative of negentropy. Actually, because $\mathbf{s}$ is a vector, it is necessary to calculate the Jacobian of the negentropy. Denoting the negentropy equation as F, its Jacobian is

$$JF(\mathbf{s}) = E\{\mathbf{x}\mathbf{x}^T g'(\mathbf{s}^T\mathbf{x})\} - \beta\mathbf{I}. \tag{4}$$

In statistical analysis, it is common to apply a whitening process, with eigenvalue decomposition, in the vectors (Franklin, 1968). Therefore $E\{\mathbf{x}\mathbf{x}^T\} \approx \mathbf{I}$ and we can rewrite $JF(\mathbf{s})$ as

$$JF(\mathbf{s}) = E\{g'(\mathbf{s}^T\mathbf{x})\} - \beta\mathbf{I}. \tag{5}$$

Approximating β for values of $\mathbf{s}$, instead of $\mathbf{s}_0$, we can rewrite the fixed-point method in its complete form as

$$\mathbf{s}^{+} = \mathbf{s} - \frac{[E\{\mathbf{x}g(\mathbf{s}^T\mathbf{x})\} - \beta\mathbf{s}]}{[E\{g'(\mathbf{s}^T\mathbf{x})\} - \beta]}, \tag{6}$$

where $\mathbf{s}^* = \mathbf{s}^{+}/\|\mathbf{s}^{+}\|$. Normalizing the equation by $\beta - E\{g'(\mathbf{s}^T\mathbf{x})\}$, the fixed-point algorithm equation is finally defined as

$$\mathbf{s}^{+} = E\{\mathbf{x}g(\mathbf{s}^T\mathbf{x})\} - E\{g'(\mathbf{s}^T\mathbf{x})\}\mathbf{s}. \tag{7}$$

The above solution searches for one independent component. In order to find all possible independent components, the method simply needs to be applied as much as necessary, just needing decorrelate the independent components. To perform decorrelation process, it is necessary to calculate $n+1$ independent components, because we subtract each one projection from the previous components founded. A deflationary scheme based on Gram–Schmidt decorrelation is applied to these subtractions

(Farina and Studer, 1984)

$$\mathbf{s}_{n+1} = \mathbf{s}_{n+1} - \sum_{i=1}^{n} \mathbf{s}_{n+1}^{T}\mathbf{s}_i\mathbf{s}_i. \quad (8)$$

Considering that the data was preprocessed with whitening, it is sufficient to renormalize the solution, obtaining $\mathbf{s}^{*}_{n+1} = \mathbf{s}_{n+1}/\|\mathbf{s}_{n+1}\|$.

3. Discrete Cosine Transform

Discrete Cosine Transform (DCT) is a method that allows expressing a function in terms of a weighted sum of cosines oscillating at different frequencies. It is closely related to the Discrete Fourier Transform (DFT; Oppenheim et al., 2009).

For a sequence of N complex signals $x_0, x_1, \cdots, x_{N-1}$, the sequence of their respective DFTs (Blinn, 1993) $\mathcal{F}_0, \mathcal{F}_1, \cdots, \mathcal{F}_{N-1}$, are

$$\mathcal{F}_k(\omega) = \sum_{n=0}^{N-1} x_n \left[\cos\left(-\frac{2\pi}{N}kn\right) + i\sin\left(-\frac{2\pi}{N}kn\right)\right]. \quad (9)$$

with $k = 0, 1, 2\cdots, N-1$.

In particular, it is possible to work only with the real part of the DFT, which is governed by the cosine function, or with the imaginary part, governed by the sine. This choice leads to two other transforms that are similar and related to the DFT: the Discrete Cosine Transform (DCT) and the Discrete Sine Transform (DST). The use of the DCT is focused on data compression or noise suppression (Rao and Yip, 1990), while the DST is mainly applied to solve Partial Differential Equations (Martucci, 1994).

Because of its significant data and energy compression, the most widely used procedure for the calculation of the DCT in signal and image processing is the DCT-II method (Rao and Yip, 1990), which is summarized in the equation

$$\mathrm{DCT-II}_k(\omega) = \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right). \quad (10)$$

Multidimensional DCT applications rigorously follow the same definitions of the one-dimensional method and are simply represented by a separate product of all cosines along each dimension. For example, if the data is in the form of a matrix, the DCT process implementation has to transform the data into two dimensions by applying the equation $\mathrm{DCT}_{k_1,k_2}(\omega): \mathbb{R}^2 \to \mathbb{R}^2$,

$$\mathrm{DCT}_{k_1,k_2}(\omega) = \sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1} x_{n_1,n_2} \cos\left(\frac{\pi}{N_1}\left(n_1+\frac{1}{2}\right)k_1\right)\cos\left(\frac{\pi}{N_2}\left(n_2+\frac{1}{2}\right)k_2\right), \quad (11)$$

where k_1, n_1 e N_1 refer to the first dimension and k_2, n_2 e N_2 are associated with the second dimension.

These multiplications can be extrapolated to the m-dimensional DCT calculations:

$$\mathrm{DCT}_{k_1,k_2,\cdots,k_m}(\omega): \mathbb{R}^m \to \mathbb{R}^m,$$

$$\mathrm{DCT}_{k_1,k_2,\cdots,k_m}(\omega) = \sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1}\cdots\sum_{n_m=0}^{N_m-1} x_{n_1,n_2,\cdots,n_m} \prod_{j=1}^{m} \cos\left(\frac{\pi}{N_j}\left(n_j+\frac{1}{2}\right)k_j\right) \quad (12)$$

Well-log profiles can be interpreted as a matrix $A_{m\times n}$, where m is the number of samples and n is the number of profiles. The application of the DCT on this matrix results in a new matrix where most of the information is allocated in its first terms (Fig. 1), for $n \geq 1$.

The amplitudes of the first terms of the DCT matrix are much higher than the last terms. These last terms do not carry important amplitudes and are set to zero, in a process alike a threshold (Battiato et al., 2001). It is important to notice that these last terms are very close to zero, but these small values still can influence Unsupervised Learning Machines such as Principal Component Analysis (Abdi and Williams, 2010), ICA or some Artificial Neural Networks such as Self-Organizing Maps (Liu et al., 2006).

After setting these last terms to zero, the Inverse Discrete Cosine Transform (IDCT) can be applied to bring back the data to the original domain.

The $\mathrm{IDCT}_{n_1,n_2}(t): \mathbb{R}^2 \to \mathbb{R}^2$, is given by

$$\mathrm{IDCT}_{n_1,n_2}(t) = \sum_{k_1=0}^{N_1-1}\sum_{k_2=0}^{N_2-1} \alpha(k_1)\alpha(k_2)x_{k_1,k_2}\cos\left(\frac{\pi}{N_1}\left(n_1+\frac{1}{2}\right)k_1\right)\cos\left(\frac{\pi}{N_2}\left(n_2+\frac{1}{2}\right)k_2\right), \quad (13)$$

where

$$\alpha(k_m) = \begin{cases} \sqrt{\frac{1}{N}}\ , \ se\ k=0 \\ \sqrt{\frac{2}{N}}\ , se\ k = 1,2,\cdots,N-1 \end{cases} \quad (14)$$

4. K-Nearest Neighbors

The K-Nearest Neighbor (K-NN) classifier is a decision rule to classify samples, classifying them by their nearest and previously classified samples (Cover and Hart, 1967). The K-NN is very simple, highly efficient and effective, but the memory requirement and the computation complexity are its disadvantages (Bhatia et al., 2010).

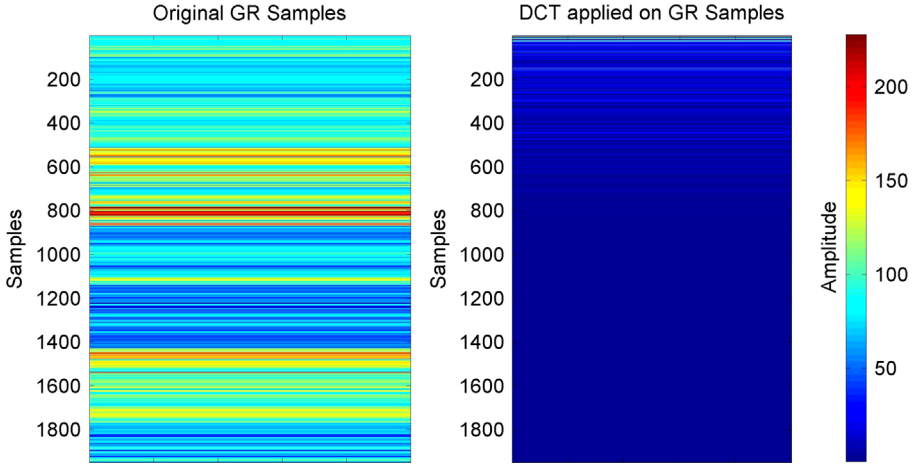

Fig. 1. Compression power of DCT in comparison with original data samples.

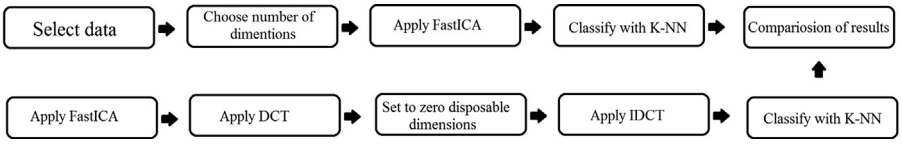

Fig. 2. Workflow methods.

The K-NN inputs are a training set and a test set. The training set is composed of n pairs (x_1,θ_i); $(x_2,\theta_i);\cdots;(x_n,\theta_i)$, where x_j represent the values of some variable in a metric space and θ_i represents the different classes of labels for each pair. Each one of θ_i is contained in the set $\{1, 2,\cdots,M\}$, where M denotes the total number of different labels (Cover and Hart,1967). The test set with n' elements x_1', x_2', $\cdots,x_{n'}'$, and the objective is to classify each of its elements, obtaining n new pairs (x'_1,θ_i); $(x'_2,\theta_i);\cdots;(x'_{n'},\theta_i)$. At this point, the value of K must be chosen, which represents the number of nearest neighbors to be considered. This choice will affect the number of samples taken into account for labeling our test set.

We can summarize K-NN in some few steps:

1. organize the training set with the values of its variables x_j and their respective labels θ_i. Organize the testing set with samples with variable values x_j' and no labels;
2. choose a value for K;
3. iteratively, for each x_j', repeat the following process:
   a. use a metric function d to detect the closest sample from x_j' in the training set, e.g., find $x_J=\min(d(x_j',x_j))$;
   b. remove x_J from the training set and repeat this process until you have removed K samples from the training set;
   c. analyze the labels θ_i from the removed samples and assign the more frequent label to x_j'.
4. construct the n' pairs $(x_1',\theta_i);(x_2',\theta_i);\cdots;(x_{n'}',\theta_i)$, based on the above procedure.

For our particular research, we have chosen the Euclidian Distance $d_E(x_J,x_j')=\sqrt{x_j{}^2-(x_j')^2}$, for simplicity.

In our research, the choice of K was based on the best success rate founded in the preliminary tests, that can be seen in Section 3. The labels θ_i were chosen based on the geological precepts, separating them by lithotypes or for them chances to be a rock reservoir.

5. Methodology application

The general workflow is as follows (Fig. 2): (i) firstly we calculate the success rate of the automatic classification for a FastICA variance/dimension reduction method. The output of this method was classified using K-NN and the success rate was calculated; (ii) for the same original data, we apply FastICA without dimension reduction and DCT as a smoothing method (Simonoff, 1996) and calculate the success rate. The best success rate is achieved when few information from the original data is used (about 1%), resulting in a smooth signal with only the primary signal behavior. This result is consistent with the spatial assumption of K-NN, because the original well-log profiles provide very different amplitude results for the same lithotype, thus disturbing the automatic classification.

All tests were run on Matlab R2010b and the toolboxes FastICA 2.5[1] and Mirt_dctn[2] were used. Well-log data from seven wells of the Namorado Field (Campos Basin), namely NA01, NA02, NA04, NA07, NA011A, RJS234 and RJS42, provided by the Brazilian National Petroleum Agency, were analyzed. For the K-NN method (Toussaint, 2005), 1950 core samples and the respective values of sonic profile (DT), gamma-ray profile (GR), resistivity profile (ILD), density profile (RHOB), neutron porosity profile (NPHI) for each core sample were used. The five logs can generate a maximum of five independent components.

In order to avoid false predictions due to the spatial proximity to the core samples, a random division was applied to create a training set and a test set with half of the data, resulting in two sets with 975 samples. The core samples were labeled based on 21 lithotypes and also as “reservoir rocks”, “possible reservoir rocks” and “non-reservoir rocks”. Each test was performed 50 times and the average of the predicted rates, amplitudes and independent components were calculated. All values represented in the subsequent figures are those averages.

Performing Eq. (8) for all independent components in the data, FastICA can be linearized (Hyvärinen, 1999) in a mixture as

$$\mathbf{s}=\mathbf{W}\mathbf{x}, \tag{15}$$

where $\mathbf{x}$ represents the recorded original data, $\mathbf{W}$ is the un-mixture matrix and $\mathbf{s}$ are the independent components. The existence of the inverse matrix $\mathbf{W}$ is guaranteed by appropriate approximations in the FastICA process (Hyvärinen, 1999).

FastICA variance/dimension reduction is achieved using fewer independent components for signal reconstruction. For example, for dimension $d<5$, we can reconstruct the original data for the 1950 samples by following the equation:

$$\mathbf{s}_{d\times 1950}=\mathrm{W}_{d\times 5}\cdot\ \mathbf{x}_{5\times 1950} \tag{16}$$

and the utput was used in K-NN classifier. This is the research's first step.

The second step, and the main objective, is uniting FastICA and DCT to prepare data to be classify. In this case, no dimension reduction was applied in FastICA methods, generating five independent components in a matrix form $\mathbf{s}'_{1950\times 5}$. Applying DCT in the matrix will generate a matrix $\mathbf{s_{dct}}'_{1950\times 5}$ already with data compression, in other others, a few dimensions $\mathbf{s'_{dct}}(i,j)$ have considerable amplitudes, mainly in the upper right block of the matrix. The other dimensions have amplitudes proportionally disposable, and in some cases, tending to zero.

Utilizing an empirical threshold, we set to zero the disposable dimensions and apply the IDCT to get back to FastICA's domain, resulting in the smoothing matrix $\mathbf{s_{sm}}$. Depending in the threshold choice, the data reconstruction is assumed as a smoothing process. It is important to note that DCT and thresholding are working similar as filters, but with the advantage of not filtering an entire frequency, because the dimensions of $\mathbf{s'_{dct}}$ are not separated in frequencies; and through DCT's matrix is possible to see the dimensions which have lower influence, a priori, any other type of method is needed to choose the dimensions need to be set to zero.

6. Automatic classification using K-Nearest Neighbors

The results of the first step evaluation are shown in Table 1 for the complete lithotype classification.

[1] http://bsp.teithe.gr/members/downloads/fastica/FastICA_2.5.zip.
[2] http://www.mathworks.com/matlabcentral/fileexchange/24050-multidimensional-discrete-cosine-transform-dct/content/mirt_dctn.m.

The highest rate of correct prediction in the automatic classification occurred in the signal reconstruction from four components. Using these four components and the corresponding matrix of inversion, we can reconstruct our signals in the data's original domain. Some part of the information is lost and the output signals are similar to the output of common noise removal methods. In Fig. 3, we separate three evidences of similarity of FastICA dimension reduction method and an expected output of noise removal methods. The first evidence in Fig. 3(b) shows some artifacts created in the noise reduction method. The second evidence indicated by Fig. 3 (c) refers to the sufficiently accurate reconstruction of the original signal with a small amplitude difference. In Fig. 3(d), the third evidence identifies a great amount of amplitude been removed in the signal reconstruction, resulting in a quite different amplitude from the original data. This disparity between the signals shows the reassembly of the original signal, not only with the removal of high frequencies, but also with modifications at various frequency levels.

The variance/dimension reduction by FastICA proved to be simple and efficient, but not effective in increasing the correct prediction rate of the automatic classification. In general, the simply variance/dimension reduction of the original data does not improve the automatic classification significantly, also because rocks of the same lithotype do not have a unique correspondent well-log value, which will be the perfect case for spatial classifiers. For instance, Table 2 shows some values extracted from the GR profile corresponding to the Coarse Sandstone Amalgamated in the core.

These very distinct values shown in Table 2 hinder the classification used by K-NN and other spatial automatic classifiers, because these methods take into account the specific coordinate information within an n-dimensional space. So variance/dimension reduction is not sufficient to improve the classification method.

Based on this assumption, all dimensions of the data were gradually set to zero in DCT domain in order to analyze its behavior while information is removed. This procedure was applied in four different types of tests: (a) classification of all lithotypes using only one component, (b) classification of all lithotypes using five components, (c) classification of the reservoir characteristics using only one component and (d) classification of the reservoir characteristics using five components. Fig. 4 shows the behavior of automatic classification according to the percentage of information used.

Table 1
Correct prediction rate of automatic classification according to the number of independent components.

1 Components (%)	2 Components (%)	3 Components (%)	4 Components (%)	5 Components (%)
14.6	30.7	53.4	61.2	59.2

Table 2
Examples of GR values for one of the lithologies.

Coarse sandstone amalgamated	
Deep (m)	**GR (API)**
3096.6	54.517
3127.2	45.504
3151.6	71.287
3158.2	102,473
3167.6	48.625

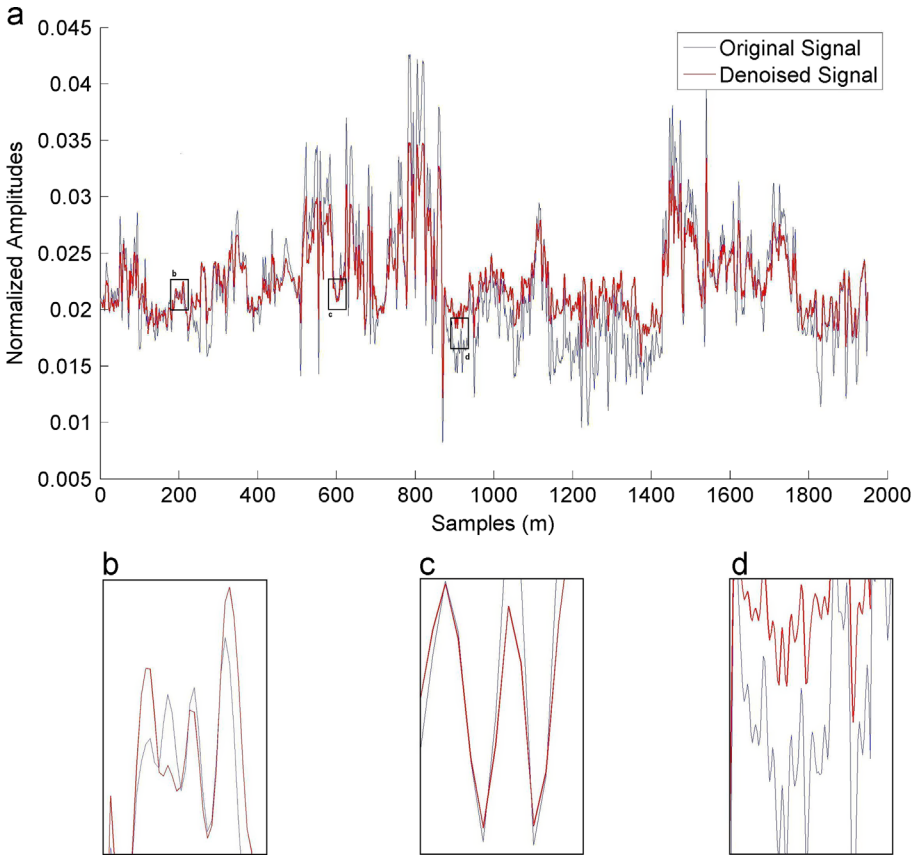

Fig. 3. Original and denoised signal via FastICA: (a) complete signal and (b–d) rectangular windows highlighting variance/dimension reduction.

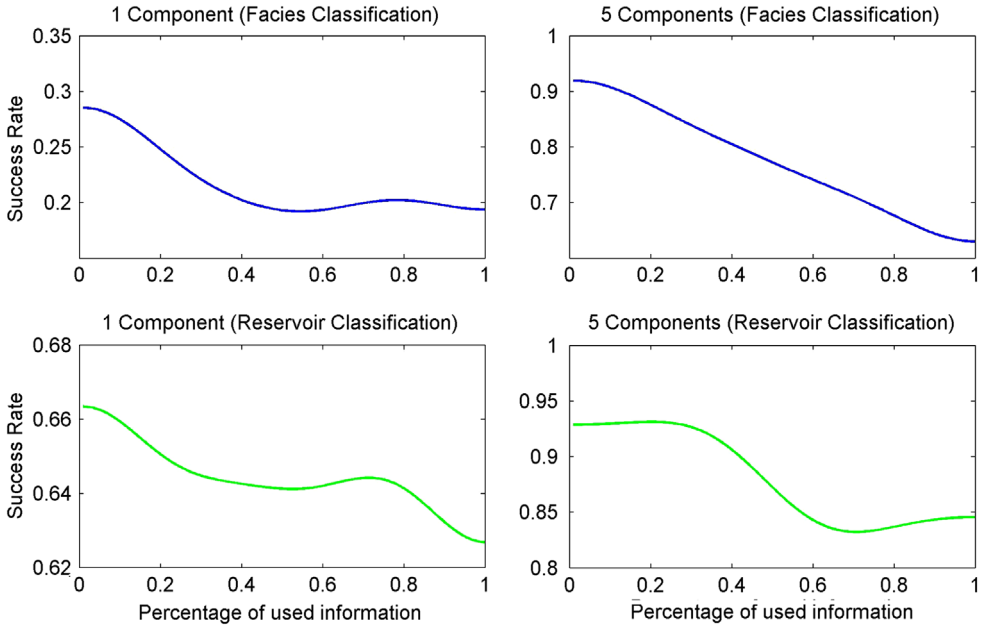

Fig. 4. Correct prediction rate of automatic classification vs. dimensions retained.

The highest correct prediction rate was observed exactly at the lower percentage of information used (approximately only 1%). The same amount of information maximizes the correct prediction rate in the four tests, so this was chosen as a parameter for tests involving well-log data and DCT. To validate these procedures, we have applied a *t*-test between the best DCT results in each one of the tests, compared with the results seen in the test without using DCT. Two-tailed *t*-tests were made with unpaired samples and the results can be seen in Table 3. The closer the results are to unity, the higher are our chances to accept the hypothesis that it is better to use DCT. Based on a threshold for the *p*-value equal to 0.05, the *t*-test results show that the use of DCT has statistically significant values.

Thus, the data were transformed to the DCT domain, then the independent components are found by FastICA and dimensions of these components are canceled, after which an inverse DCT is applied to return to the original data domain. This sequence was applied based on the hypothesis that a better spatial arrangement of the data would be generated, therefore having a better correct prediction rate in the automatic classification.

Notice that another hypothesis can be assumed: that specific information such as noise reduction has not much influence in the classification via K-NN. To analyze the impact of such premise, the four applications that are shown in Fig. 4 were used. The automatic classification was tested for three cases: (I) when FastICA is applied before the DCT; (II) when FastICA is applied during the DCT; and (III) when the FastICA is applied after the DCT. Table 4 contains the success rate for each application (a, b, c and d) for the three situations(I, II, III).

Table 4 shows that the order of applications has a small impact on the correct prediction rate of the automatic classification, but case (III) works slightly better. However, the most important changes in the correct predictions, as seen in Fig. 4, are not related to the order of application of the methods and also are not related only to noise reduction; therefore, this improvement can only occur due to the behavior of the signals processed with the DCT. To confirm this hypothesis, a new test was carried out under the same conditions of the previous applications (a, b, c, d) shown in Fig. 4 and Table 4. This was done in order to compare the rate of correct classification with four different approaches: (I) automatic classification with no processing method; (II) automatic classification with FastICA; (III) automatic classification with DCT; and (IV) automatic classification with DCT and FastICA. The results of these tests are shown in Fig. 5.

Table 3
T-test results between classification outputs before and after DCT application.

	Aplication (a)	Aplication (b)	Aplication (c)	Aplication (d)
t	15.725	50.3392	5.53807	13.81263
***P*-value**	0	0	0.00003	0
Power student	1	1	0.9985	1

Table 4
Maximum correct prediction percentage relative to FastICA calculation order.

	Aplication (a,%)	Aplication (b,%)	Aplication (c,%)	Aplication (d,%)
Before	72.10	89.44	84.00	93.64
During	71.90	88.62	84.82	92.72
After	67.49	89.44	83.49	93.33

By inspecting Fig. 5, we observe that the union of DCT and FastICA methods provides the best prediction rate and the most consistent results, but most of the automatic classification accuracy is derived from the signal processing with DCT and not with FastICA. This statement is coherent because Multivariate Analysis (e.g. FastICA, ICA, and PCA) generally improves the interpretation of spatial data, but the signals have an equivalent reconstruction from original data, thus its isolated application for spatial classifiers is not the ideal choice for automatic classification.

Fig. 6 illustrates the effect of the DCT application on a Gamma-Ray (GR) signal into four levels of used information (70%, 50%, 20%, 1%).

Basically, a noise reduction can be seen in the three first levels, but the objective is to maximize the success rate of automatic classification. For this purpose, the best choice is in the 1% parameter for DCT. To maximize the correct prediction rate, it is only necessary to maintain the general signal behavior, meaning that a Smoothing method (Simonoff, 1996) is more appropriate than a variance/dimension reduction method. In Fig. 6(d), the points can be grouped more easily based on GR values than Fig. 6 (a–c). Fig. 7 illustrates better the difference between the original

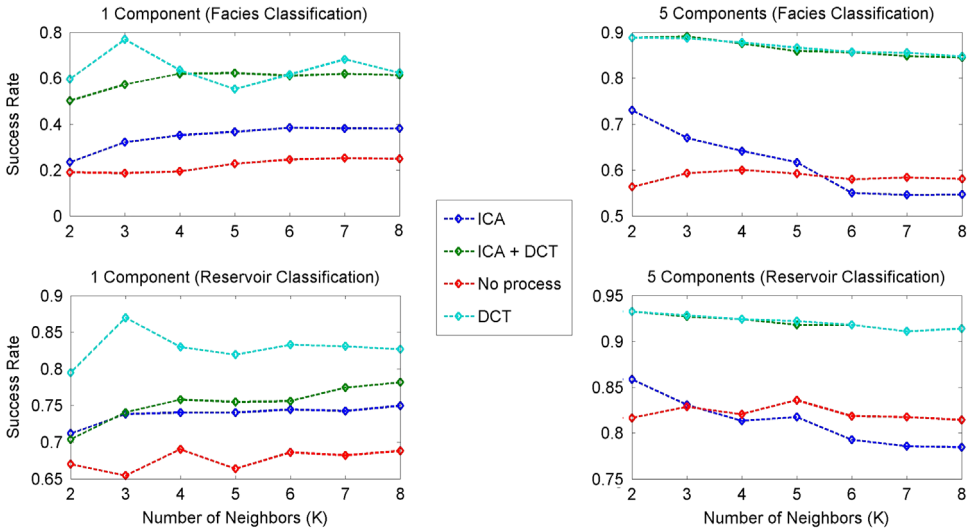

Fig. 5. Automatic classification using four different methods.

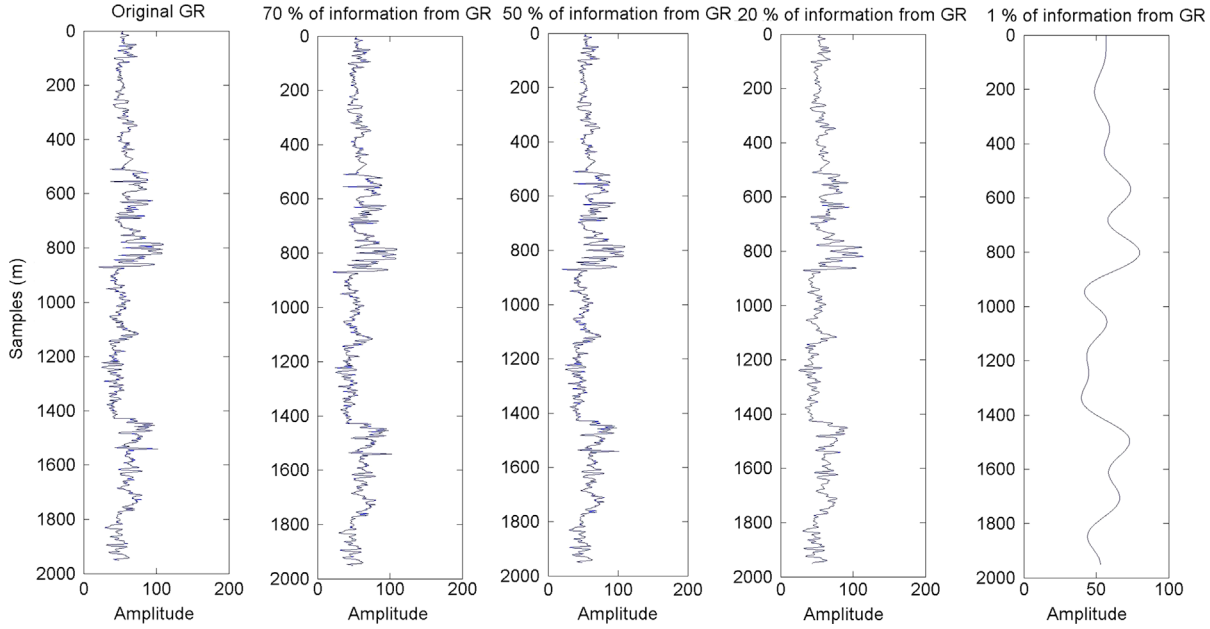

Fig. 6. GR signal reconstruction using different levels of information.

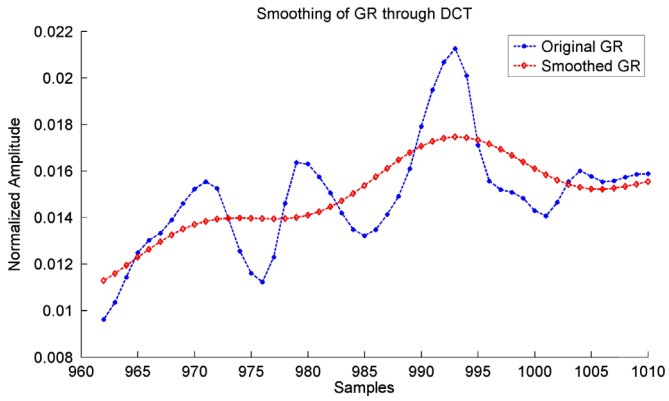

Fig. 7. Smoothing of GR signal of the same lithotypes.

and the smoothed signal, by showing a portion of the GR signal associated to 48 core samples (samples 962–1010) corresponding to Bioturbed Marble.

Although the samples correspond to the same lithotype, the GR responses vary significantly. While the original signal generates a training set with very different values for the same class, the DCT smoothing approximates these points, making the training set more consistent with similar samples. Because of this characteristic, the correct prediction rate of this method is higher than a simple variance/dimension reduction.

7. Conclusions

DCT has a much greater flexibility than FastICA when applied to reduce noise in well-log data. The difference lies in the optimal threshold for the removal of noise. When the FastICA method is applied for variance/dimension reduction, this is carried out by dimensionality reduction. On the other hand, when DCT is applied for noise reduction, the optimal choice is given by an empirical threshold and the best success rate of automatic classification obtained by K-NN controls its performance.

For K-NN, the variance/dimension reduction process improves the automatic classification success rate only in 2%. The K-NN automatic classification accuracy rate was maximized with use of FastICA and DCT, improving the best result in about 8% and the worst result in about 20%. The best results were achieved in a very smoothed signal, possibly because the amplitude difference is smoothed, grouping the points which corresponding to the same rock lithotypes. Further studies can be carried out to check the influence of the signal variograms in the application of this methodology.

References

Abdi, H., Willians, L.J., 2010. Principal Component Analysis. WIREs Comp Stat 2, 433–459 〈http://onlinelibrary.wiley.com/doi/10.1002/wics.101/pdf〉.

Amaziane, B., Bourgeat, A., Jurak, M., 2006. Effective macrodifusion in solute transport through heterogeneous porous media. Multiscale Modeling and Simulation 5, 184–204.

Avseth, P., Mukerji, T., Mavko, G., 2005. Quantitative Seismic Interpretation. Applying Rock Physics to Reduce Interpretation Risk. Cambridge University Press, Cambridge, New York, Melbourne.

Battiato, S., Mancuso, M., Bosco, A., Guarnera, M., 2001. Psychovisual and statistical optimization of quantization tables for DCT compression engines. In: Proceedings of the 11th International Conference on Image Analysis and Processing, ICIAP'01, Palermo, Italy, p. 602.

Bhatia, N., et al., 2010. Survey of nearest neighbor techniques. International Journal of Computer Science and Information Security 2 (8), 302–305.

Blinn, J.F., 1993. What's the deal with the DCT. IEEE Computer Graphics and Applications 13 (4), 78–83.

Burden, R.L., Faires, J.D., 1985. Numerical Analysis, Third Edition Prindle, Weber & Schimidt, Boston.

Carrasquilla, A., Leite, M.V., 2009. Fuzzy logic in the simulation of sonic log using as input combinations of gamma ray, resistivity, porosity and density well logs from Namorado Oilfield. In: Proceedings of the 11th International Congress of the Brazilian Geophysical Society, Salvador, Brazil.

Coconi-Morales, E., Ronquillo-Jarillo, G., Campos-Enríquez, J.O., 2010. Multi-scale analysis of well-logging data in petrophysical and stratigraphic correlation. Geofísica Internacional 49 (2), 55–67.

Comon, P., 1994. Independent component analysis: a new concept?. Signal Processing, 36; 287–314.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13 (1), 21–27.

Doyen, P.M., 2007. Seismic reservoir characterization: an earth modelling perspective. EAGE Publications, Houten, The Netherlands.

Dubrule, O., 1994. Estimating or choosing a geostatistical model. In: Dimitrakopoulos, R. (Ed.), Geostatistics for the Next Century. Kluwer Academic Publishers, Dordcrecht, The Netherlands, pp. 3–14.

Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. Wiley, New-York.

Farina, A., Studer, F.A., 1984. Application of Gram-Schmidt algorithm optimum radar signal processing. IEEE Proceedings Part F 131, 139–145.

Franklin, J.N., 1968. Matrix Theory. Englewood Cliffs: Prentice-Hall. 292 pp.

Grana, D., Pirrone, M., Mukerji, T., 2012. Quantitative log interpretation and uncertainty propagation of petrophysical properties and facies classification from rock-physics modeling and formation evaluation analysis. Geophysics 77, WA45–WA63.

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10 (3), 626–634.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons, Toronto 481 pp.

Liu, Y., Weisberg, R.H., Mooers, C.N.K., 2006. Performance evaluation of the self-organizing map for feature extraction. Journal of Geophysical Research 111, C05018, http://dx.doi.org/10.1029/2005JC003117.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (Eds.), Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp. 281–297.

Martucci, S.A., 1994. Symmetric convolution and the discrete sine and cosine transforms. IEEE Transactions on Signal Processing SP-42, 1038–1051.

Messina, A., Langer, H., 2011. Pattern recognition of volcanic tremor data on Mt. Etna (Italy) with KKAnalysis—a software program for unsupervised classification. Computers and Geosciences 37 (2011), 953–961.

Mitchell, T., 1997. Machine Learning. McGraw-Hill Higher Education, New York 432 pp.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 2009. Discrete-Time Signal Processing, 3th ed. Prentice Hall, NJ 1120 pp.

Rao, K.R., Yip, P., 1990. Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, Boston 512 pp.

Rosati, I., Cardarelli, E., 1997. Statistical pattern recognition technique to enhance anomalies in magnetic surveys. Journal of Applied Geophysics 37 (2), 55–66.

Rutherford, S.R., Willians, R.H., 1989. Amplitude versus offset variations in gas sands. Geophysics 54 (06), 680–688.

Russell, S., Norvig, P., 2002. Artificial Intelligence: A Modern Approach. Prentice Hall, Essex, England 478 pp.

Schuerman, J., 1996. Pattern Classification: A Unified View of Statistical and Neural Approaches. Wiley & Sons, New York 392 pp.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York 368 pp..

Toussaint, G.T., 2005. Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. International Journal of Computational Geometry and Applications 15 (2), 101–150.

Turlapaty, A.C., Anantharaj, V.G, Younan, N.H., 2010. A pattern recognition based approach to consistency analysis of Geophysical datasets. Computers and Geosciences 36, 464–476.

Vidal, A.C., Sancevero, S.S., Remacre, A.Z., Costanzo, C.P., 2007. Modelagem geoestatística 3D da impedância acústica para a caracterização do campo de namorado. Revista Brasileira de Geofísica 25 (3), 295–305.