AUTOMATIC CLASSIFICATION OF LITHOFACIES USING FAST INDEPENDENT COMPONENT ANALYSIS

Alexandre Cruz Sanchetta¹, Emilson Pereira Leite², Bruno César Zanardo Honório¹ and Alexandre Campane Vidal²

ABSTRACT. The problem of automatic classification of facies was addressed using the Fast Independent Component Analysis (FastICA) of a data set of geophysical well logs of the Namorado Field, Campos Basin, Brazil, followed by a $k$-nearest neighbor ($k$-NN) classification. The goal of an automatic classification of facies is to produce spatial models of facies that assist the geological characterization of petroleum reservoirs. The FastICA technique provides a new data set that has the most stable and less Gaussian distribution possible. The $k$-NN classifies this new data set according to its characteristics. The previous application of FastICA improves the accuracy of the $k$-NN automatic classification and it also provides better results in comparison with the automatic classification by means of the Principal Component Analysis (PCA).

Keywords: automatic classification, geophysical well logs, Independent Component Analysis.

RESUMO. O problema da classificação automática de facies foi abordado através da Análise de Componentes Independentes Rápida (FastICA – Fast Independent Component Analysis) de um conjunto de dados de perfis geofísicos de poços do Campo de Namorado, Bacia de Campos, seguida de classificação por $k$ vizinhos mais próximos ($k$-NN – $k$-nearest neighbor). A classificação automática de facies é utilizada para gerar modelos de distribuição espacial de facies que auxiliam a caracterização geológica dos reservatórios de petróleo. A técnica FastICA encontra um novo conjunto de dados com distribuição mais estável e menos gaussiana possível e o $k$-NN classifica esse novo conjunto de acordo com suas características. A aplicação prévia da FastICA melhora a porcentagem de acerto da classificação automática pelo $k$-NN, fornecendo melhores resultados quando comparada com a classificação automática por Análise de Componentes Principais (PCA – Principal Component Analysis).

Palavras-chave: classificação automática, perfis geofísicos de poços, Análise de Componentes Independentes.

¹Universidade Estadual de Campinas, Reitoria V, Barão Geraldo, 13083-872 Campinas, SP, Brazil – E-mails: alexandr@dep.fem.unicamp.br; brunohonorio@gmail.com
²Universidade Estadual de Campinas, Prédio do DGRN, Rua João Pandiá Calógeras, 51, Barão Geraldo, P.O. Box: 6152, 13983-970 Campinas, SP, Brazil. Phone: +55(19) 3521-4659; Fax: +55(19) 3289-1562 – E-mails: emilson@ige.unicamp.br; vidal@ige.unicamp.br
INTRODUCTION

The automatic interpretation of well log data is supported by multivariate analysis because there are many factors like data complexity, processing errors, and statistical limitations of the multivariate analysis techniques. Many statistical multivariate methods like Principal Component Analysis (PCA) or Discriminant Analysis (DA) have been applied to analyse multiple geophysical well logs to produce lithofacies models (Flexa et al., 2004; Rosa et al., 2008; Sancevero et al., 2008). These methods together with the knowledge of petrophysical proprieties are often used in flow simulators of a petroleum field. Such task is complex because the measurements express the change of the physical proprieties which can not always be directly associated with lithological variations. The purpose of this work is to apply a two-fold methodology to automatic classification of lithotypes using geophysical well logs as input data. In this two-fold methodology, the Fast Independent Component Analysis (FastICA) is applied first, followed by the k-nearest neighbor (k-NN) pattern recognition method.

In general, the classification of raw well log data using the k-NN is not sufficiently efficient and/or efficacious because of inherent redundancies in the data. This problem may be overcome by rearranging the information of the original data set into a new linearly independent (LI) set which still contemplates the original information, but remove such redundancies. A simple strategy to obtain a new LI data set is the well known process of dimension reduction. For example, in the PCA a new data set with reduced dimension can be created by selecting only the autovectors corresponding to the highest autovalues. However, the focus of automatic classification is not to select a particular new data set, but to obtain the highest prediction accuracy.

Like the aforementioned multivariate methods, the Independent Component Analysis (ICA) is classified as a Blind Signal Separation method (Murata et al., 2001) and does not need a priori any information neither about the statistical distribution nor about the spatial coordinates of the data. Because the existence of independent input signals is assumed based on the particular physical laws of independent events that are involved, this method can in principle be applied to several kinds of geological, physical or geophysical data (Casey, 2001). Different from other methods like PCA, ICA searches for non-Gaussian component estimations, that is, the components that are as disparate as possible from the Gaussian distribution (Comon, 1994). Most of the multivariate analyses find new Gaussian components and this fact distinguishes ICA from the other methods. Being so, ICA provides more reliable results in situations where the Gaussian components do not conform to the expected result. Then, the search for non-Gaussianity is one of the most important aspects of the method. In terms of computing, there are many different approximations that can be used to find the least Gaussian components possible.

The ICA, as it is presented in its first versions (Comon, 1994), has a high computational cost, because it is based on successive matrix multiplications, requiring a great computational effort for its use. For this reason, many researches were motivated to develop optimized ways of finding the independent components (Cardoso & Souloumiac, 1993; Hyvärinen, 1999; Marchini et al., 2009). In the literature, the FastICA method stands out for being a hundred times faster than the ICA method (Hyvärinen, 1999).

The main advantages of FastICA when compared to traditional ICA methods are:

(i) the convergence is cubic, which means a very fast convergence, a fact that has been confirmed by simulations and experiments based on real data (e.g. Giannakopoulos et al., 1999);

(ii) there are no step size parameters to choose, meaning that the algorithm is easy to use;

(iii) it does not require previous estimation of a probability distribution function;

(iv) the independent components can be estimated one by one, decreasing the computational effort of the method in cases where only some of the independent components are of interest.

In this work, the independent components found using the FastICA method were used as training data for the pattern recognition method known as k-nearest neighbor (k-NN). The k-NN is a classifier algorithm based on the structural distance between test and training sets (Cover & Hart, 1967). Given a test set with many samples, each one is spatially analyzed and a label (class) is assigned according to the maximum number of neighbors in the chosen neighborhood. For this particular application, the classes were defined using core samples extracted from wells of the Namorado Field in Campos Basin, Brazil, avoiding approximations or calibration problems between lithology and log samples for each log. A quantitative analysis of the proposed methodology was conducted by a series of tests varying the values of the FastICA and k-NN parameters, using well log input data extracted from depth intervals where core samples are available. The final product of each test is a success rate plot, where the performance of the classifier can be observed according to the variation of the parameter value. These results can be used for comparison with those obtained using other statistical multivariate methods, or to support geological interpretation in regions where reservoir properties have to be estimated from seismic data.
METHODOLOGY

The methodology consists in using the independent components extracted by the FastICA as a new training and test data set for the k-NN classification method. This new data set has a maximum number of dimensions equal to the number of input variables, i.e., geophysical well logs. For example, if the sonic (DT) and density (RHOB) logs of a well are used, two independent components are found. The set of values of the independent components at a particular depth is called sample.

The k-NN is a well known and simple automatic classification method (Dasarathy, 1991; Sancevero et al., 2008). The k-NN method is easy to implement and allows spatial sample calculations, justifying its use in this work. However, as a supervised classification method, it is necessary to use at least some of the core data as target values in the training process. The training and test set (Cover & Hart, 1967). Given a test set with many samples, each sample will be spatially analyzed and a label (class) will be assigned that is equal to the most commonly found label among the nearest neighbors.

FastICA

The ICA approach has some disadvantages, such as problems with the contrast function maximization, computational cost and the choice of the learning rate. To solve those limitations, the FastICA approach was proposed.

FastICA is based on a fixed-point algorithm and the speed of the method, compared with ICA, is improved by the large number of data points (samples) used in a single step of the algorithm (Hyvärinen & Oja, 1997). In addition to this characteristic, FastICA works similarly to a neural algorithm, because both of them allow parallel computing, i.e., the data processing can be carried out by multiple operations in each step. Moreover, the FastICA is computationally simple and requires small space in memory (Hyvärinen, 1999).

The employed fixed-point algorithm is based on Newton’s method (Macleod, 1984). Each independent component \( w^+ \) is iteratively updated by

\[
 w^+ = w - \frac{f(x)}{f'(x)}, \quad (1)
\]

where \( w \) is a random initial guess of a non-Gaussian component that iteratively converges to the actual independent component \( w^+ \), by maximizations of some contrast function \( f(x) \) (Comon,

1994). In the traditional ICA context, the contrast functions can be: Skew : \( f(x) = x^2 \), Pow3 : \( f(x) = x^3 \), Gauss : \( f(x) = \exp \left( -\frac{x^2}{2} \right) \) and Tanh : \( f(x) = \tanh(x) \).

In the FastICA, the negentropy (\( F \)) contrast function

\[
 F = E\{xg(w^T x)\} - \beta w = 0 \quad (2)
\]

is used, because it allows the Kuhn-Tucker stationarity condition (Kuhn & Tucker, 1951). \( \beta = E\{w^T x g(w^T x)\} \) and \( w_0 \) is the optimum value of the process. The Jacobian of the Negentropy (\( F \)) can be written as

\[
 JF(w) = E\{xx^T g'(w^T x)\} - \beta I. \quad (3)
\]

As it is usual in multivariate statistical methods, a statistical centralization of the data is applied, known as whitening (Hyvärinen et al., 2001). If the data are whitened, then \( E\{xx^T\} \approx I \) and Eq. (3) can be rewritten as \( JF(w) = E\{g'(w^T x)\} - \beta I \). Approximating \( \beta \) to \( w \) values, instead of \( w_0 \), the complete fixed-point algorithm can be written as

\[
 w^+ = w - \frac{E\{xg'(w^T x)\}}{E\{g'(w^T x)\}} \beta I. \quad (4)
\]

After rationalizing Eq. (4), we obtain

\[
 w^+ = E\{xg'(w^T x)\} - E\{g'(w^T x)\} w. \quad (5)
\]

Equation (5) finds one independent component at a time. If the purpose is to find all \( n \) possible independent components of a particular data set, then it is necessary to calculate \( n + 1 \) different independent components, and for each single component, decorrelate the other \( n \) components, resulting in components that are linearly independent. To accomplish the linear independence, a deflation scheme based on the Gram-Schmidt-like decorrelation (Zhang & Chan, 2006) is applied, resulting in

\[
 w_{n+1} = w_{n+1} - \sum_{i=1}^{n} w_{n+1}^T w_i w_i. \quad (6)
\]

Considering the effect of data whitening, Eq. (6) is normalized:

\[
 w_{n+1} = w_{n+1} / ||w_{n+1}||. \quad (7)
\]

k-Nearest Neighbors (k-NN)

The k-Nearest Neighbors (k-NN) is a supervised classifier algorithm based on the structure and distance between samples of a test and a training set (Cover & Hart, 1967). Given a test set with many samples, each sample will be spatially analyzed and a label (class) will be assigned that is equal to the most commonly found label among the nearest neighbors.
The k-NN is a type of instance-based learning, because even with the establishment of the whole training set, it does not perform an explicit generalization beyond the training set. Instead, it simply compares new samples with the training samples. Nevertheless, k-NN is one of the simplest classification algorithms and it is recommended in cases where statistical distributions of the samples are unknown or in cases where the initial parameter choices are intricate. For example, classifiers based on Artificial Neural Networks (ANN) demand the choice of some crucial parameters to obtain a better performance, such as numbers of neurons, learning rate, number of iterations, etc. These informations are in general selected according to the statistical characteristics of the data (França, 2009). Unsupervised classifiers like the k-means demand the choice of several parameters (MacQueen, 1967). In the k-NN, sample distances are calculated using Euclidian metric and the only initial parameter that needs to be set is k, which is the number of neighbors to be analyzed for each sample (Toussaint, 2005).

DATA

Well log data from the Namorado Field in Campos Basin were selected to evaluate the proposed methodology. The data selection was motivated by the immediate availability of these data from the Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), and for being considered a school field (Barboza, 2005). The Namorado Field is located in the north central part of the hydrocarbon accumulation zone of the Campos Basin and it was discovered in 1975 (Vidal et al., 2007). The log data used were: Sonic (DT), Gamma Ray (GR), Resistivity (ILD), Density (RHOB), and Neutronic Porosity (NPHI), extracted from seven wells: NA01, NA02, NA04, NA07, NA011A, RJS234, and RJS42. The sampling frequency of the log data is 0.2 m depth. Core data were available only through some specific intervals of the wells. Core data were labeled as non-reservoir, sandstones and conglomerates; and the remaining lithotypes refer to sandstones intercalated with shales, fine grained sandstones and conglomerates; and the remaining lithotypes were labelled as non-reservoir.

RESULTS

In the first test batch, TR and TS was defined using Type 1 division and Type A labelling. As seen in the first test batch, the classification using all components provides the higher success rate. Compared with the first batch, this one presents a higher success rate for all cases (Fig. 2). This is expected because the possible labels were reduced from 29 to only three.

The second test batch was carried out using Type 1 TR/TS division and Type B labelling. As seen in the first test batch, the classification using all components provides the higher success rate. Compared with the first batch, this one presents a higher success rate for all cases (Fig. 2). This is expected because the possible labels were reduced from 29 to only three.
The four test batches show that the FastICA provides maximum success rates between 60% and 70% for Type A labelling, and between 83% and 86% for Type B labelling.
COMPARISON BETWEEN FastICA AND PCA

In order to compare the use of ICA and PCA to generate the input data to the $k$-NN classifier, two more test batches were conducted. A fifth test batch was applied using Type 1 TR/TS division, Type A labelling and five FastICA components. The results are shown in Figure 5. The sixth and last test batch was conducted using Type 1 TR/TS division, and Type B and five FastICA components. The results are found in Figure 6. In all situations, the FastICA was more efficacious than the PCA.

CONCLUSIONS

The results of this work show that the use of independent components calculated by FastICA as input to the $k$-NN classifier provides high accurate lithofacies (reservoir; non reservoir; interlaminated) predictions, particularly when training and test samples are randomly divided and all independent components are used. In such cases, the success rates are above 80%. However, when all lithotypes has to be classified, the success rate is reduced to less than 70%. In comparison to the use of PCA, FastICA may

Figure 5 – PCA × ICA Comparative: Even/Odd Test, Classification of Lithofacies, All Components.

Figure 6 – PCA × ICA Comparative: Even/Odd Test, Classification of characteristics, All Components.

improve the $k$-NN success rate from 77% to 86% when few neighbors and all independent components are used.

ACKNOWLEDGEMENTS
The authors would like to thank Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP and PETROBRAS for the availability of the data from Namorado Field.

REFERENCES


NOTES ABOUT THE AUTHORS

Alexandre Cruz Sanchetta is a Ph.D. student in Petroleum Engineering at Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas, since August 2010. Received a master's degree in Sciences and Petroleum Engineering at the same institution (2010) and a degree in Mathematics from the same university (2007). Developed a Scientific Initiation project in Lie Groups and Differential Equations financed by PIBIC/CNPq, from August 2006 to December 2007. Research interests are multivariate statistical analysis of well log and geophysical data.

Emílson Pereira Leite received a Ph.D. degree in Geophysics from the Universidade de São Paulo (2005) and was a postdoctoral fellow at the Instituto de Geociências, Universidade Estadual de Campinas, where he is now a professor. Research focuses on the integration of geophysical and geologic data to the exploration of Earth's natural resources. Areas of research include gravity and seismic inversion, seismic attribute analysis, geologic interpretation of radiometric and magnetic anomalies, and application of neural networks and geostatistical methods to the prediction of rock properties.

Bruno César Zanardo Honório received a B.Sc. in Physics (2008) and an M.Sc. in Science and Petroleum Engineering (2011), both from the Universidade de Campinas – UNICAMP, where is pursuing a Ph.D. in Geosciences. Works as a researcher at the Center for Oil Studies, CEPETRO/UNICAMP. Research interests are signal processing, time-frequency analysis, statistical methods, noise attenuation, and geophysical attributes.

Alexandre Campano Vital received a B.Sc. degree in Geology from the Universidade de São Paulo (1993), a M.Sc. degree in the Universidade Estadual de Campinas (1997), and a Ph.D. in Regional Geology from Universidade Estadual Paulista – UNESP (2002). Completed postdoctoral studies in the Departamento de Geologia Aplicada, UNESP (2003). Has experience in geology with an emphasis on reservoir geology.