AUTOMATIC CLASSIFICATION OF LITHOFACIES
USING FAST INDEPENDENT COMPONENT ANALYSIS

Alexandre Cruz Sanchetta[1], Emilson Pereira Leite[2],
Bruno César Zanardo Honório[1] and Alexandre Campane Vidal[2]

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 fácies 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 fácies é utilizada para gerar modelos de distribuição espacial de fácies 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.

[1]Universidade Estadual de Campinas, Reitoria V, Barão Geraldo, 13083-872 Campinas, SP, Brazil – E-mails: alexandr@dep.fem.unicamp.br; brunohonorio@gmail.com

[2]Universidade 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 orginal 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 process defines labels for each sample that is used to classify the rest of the samples that have no associated core data. The result of each sample classification is an output label (class). The output label is compared to the known core data, allowing the calculation of the number of correct predictions. The number of correct predictions divided by the total number of samples is the success rate of the method.

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 $\mathbf{w}^+$ is iteratively updated by

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

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

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

In the FastICA, the negentropy (F) contrast function

$$F = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{w} = \mathbf{0} \qquad (2)$$

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

$$JF(\mathbf{w}) = E\{\mathbf{x}\mathbf{x}^T g'(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{I}. \qquad (3)$$

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 $E\{\mathbf{x}\mathbf{x}^T\} \approx \mathbf{I}$ and Eq. (3) can be rewritten as $JF(\mathbf{w}) = E\{g'(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{I}$. Approximating β to $\mathbf{w}$ values, instead of $\mathbf{w}_0$, the complete fixed-point algorithm can be written as

$$\mathbf{w}^+ = \mathbf{w} - \frac{[E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{w}]}{[E\{g'(\mathbf{w}^T\mathbf{x})\} - \beta]}. \qquad (4)$$

After rationalizing Eq. (4), we obtain

$$\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}. \qquad (5)$$

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 independents. To accomplish the linear independence, a deflation scheme based on the Gram-Schmidt-like decorrelation (Zhang & Chan, 2006) is applied, resulting in

$$\mathbf{w}_{n+1} = \mathbf{w}_{n+1} - \sum_{i=1}^{n} \mathbf{w}_{n+1}^T \mathbf{w}_i \mathbf{w_i}. \qquad (6)$$

Considering the effect of data whitening, Eq. (6) is normalized: $\mathbf{w}_{n+1}^* = \mathbf{w}_{n+1}^+ / ||\mathbf{w}_{n+1}^+||$.

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 intricated. 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. The number of available samples in the data of all seven wells is 4732 samples. There are 1950 samples that can be directly associated to the available core data. The samples with core data are used in the validation tests.

TRAINNING AND VALIDATION

Five independent components were calculated by applying FastICA to the five well logs, where each component has 1950 samples. Therefore, the data matrix composed of the independent components (IC) has dimensions 1950×5. However, it is important to notice that, as the classification is carried out by means of statistical relationships based on the spatial distance among the samples, the number of independent components (NC) for the training and test will vary based on the particular available data (e.g. Efron & Tibshirani, 1995). For this reason, NC was a variable from 1 to 5 in four of our tests.

Having fixed a particular NC, the samples were divided into two sets to allow validation of the classifications: a training set (TR) and a testing set (TS). The TR contains the samples associated to the available core data and was used to predict the labels of each sample in TS. There are many ways to separate the data in TR and TS, and the interpreter must choose the way this division will be carried out (e.g. Leite & Souza Filho, 2009).

Two types of data division were applied: (Type 1) after assigning an index $i = 1, 2, \ldots, 1950$ to each sample, TR was composed by even indexed samples and TS by odd indexed samples. In this case, both TR and TS have 975 samples; and (Type 2) a random division into two groups containing 975 samples each. When Type 2 TR/TS division was used, each test consisted of a hundred times runs of the algorithm, each one corresponding to a different random TR/TS distribution of samples. In this case, only the average results are presented.

Two types of labels were used: (Type A) labels based on 29 lithotypes identified in the core samples of the wells in the Namorado Oil Field; and (Type B) labels based on groups of characteristics of these lithotypes, dividing them into reservoir, interlaminated and non reservoir rocks. The reservoir label corresponds to the facies related to thicker packages of sandstones with high values of porosity and permeability; the interlaminated label refers 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 was applied. Results show the maximum success rate obtained when all independent components and small values for k are used. Tests with fewer components show that the maximum success rate is observed for large values of k and do not have the same effiency (Fig. 1).

The second test batch was carried out using Type 1 TR/TS divison 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 two previous test batches can already be used to evaluate the methodology. Those results are not related to a possi-
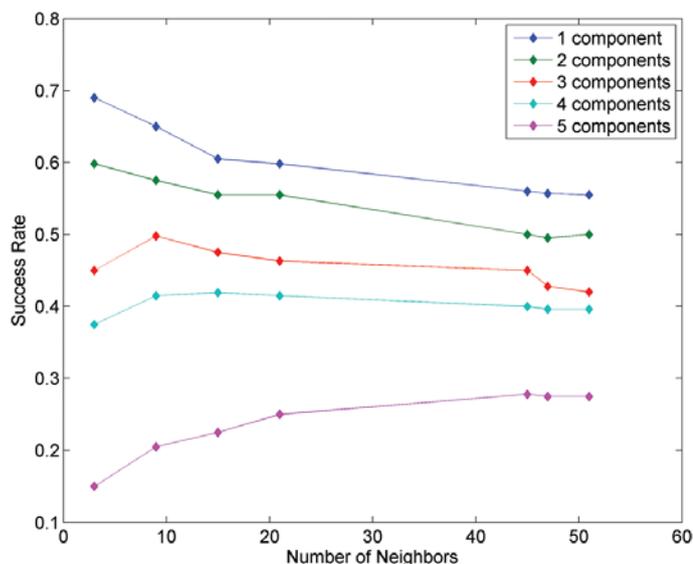
Figure 1 – First test batch: Even/Odd test, Classification of lithofacies.
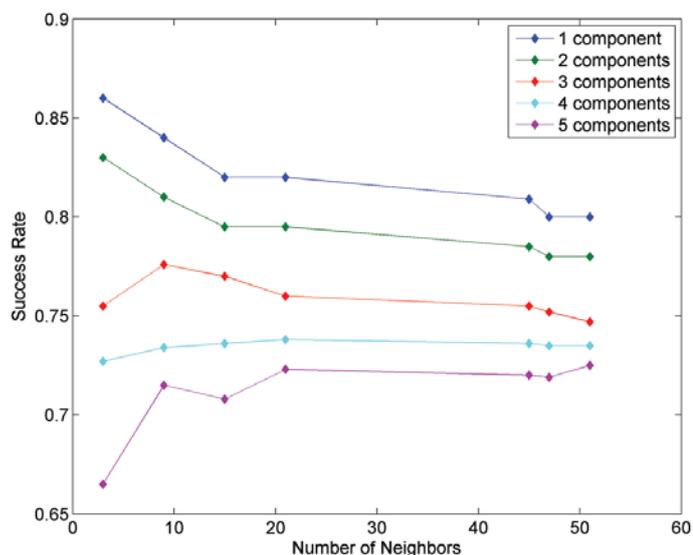

Figure 2 – Second test batch: Even/Odd test, Classification of characteristics.

ble spatial correlation among the samples, and the efficiency of the classification does not come from that particular TR/TS division. However, to include randomness in the creation of TR and TS, a third test batch using Type 2 TR/TS division was employed. Type A labelling was used.

As seen in the two previous tests, the number of independent components directly influences the success rates (Fig. 3). In comparison with the first test batch, there is a small decrease in the average of correct predictions. However, a direct comparison is not possible, because the first batch was conducted using a fixed TR/TS separation and resulted in the same success rate for all

tests, while the third test batch, with a variable TR/TS separation, produced different success rates for each classification attempt.

In the fourth test batch, Type 2 TR/TS divison and Type B labelling was used. Similarly to the previous test batch, the average success rate decreases in comparison to the rates of the second batch, but both have a similar behavior (Fig. 4). It is clear that the results strongly depend on how the data are split into TR and TS and also on the labelling of the samples.

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.
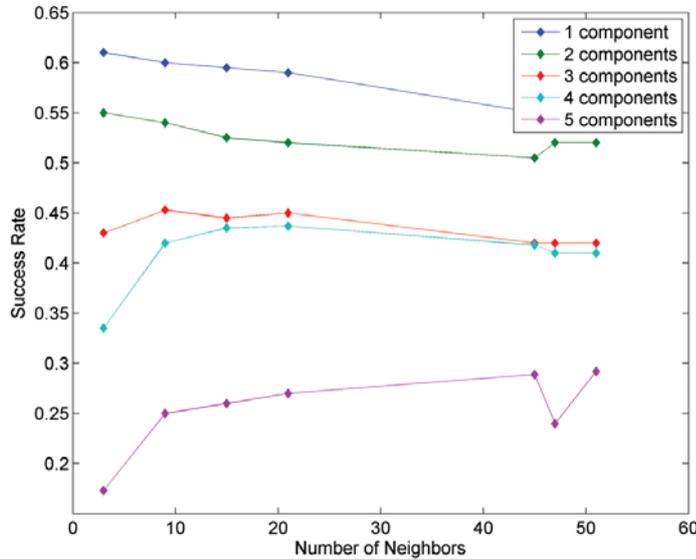
Figure 3 – Third test batch: Random test, Classification of lithofacies.
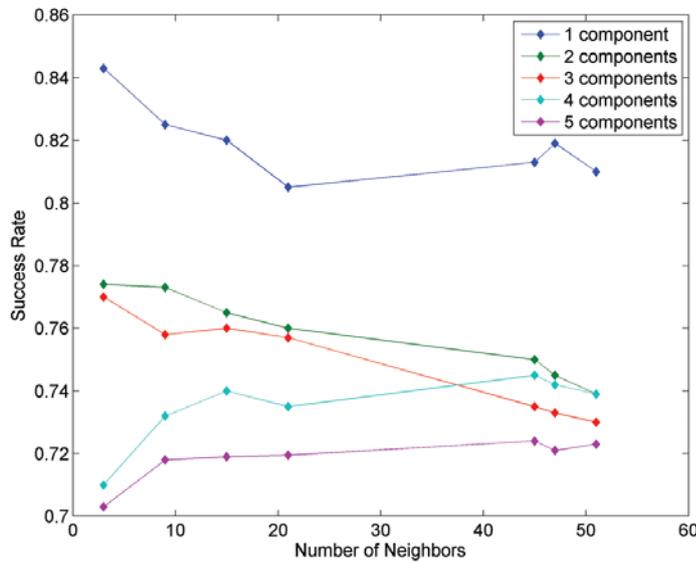

Figure 4 – Fourth test batch: Random test, Classification of characteristics.

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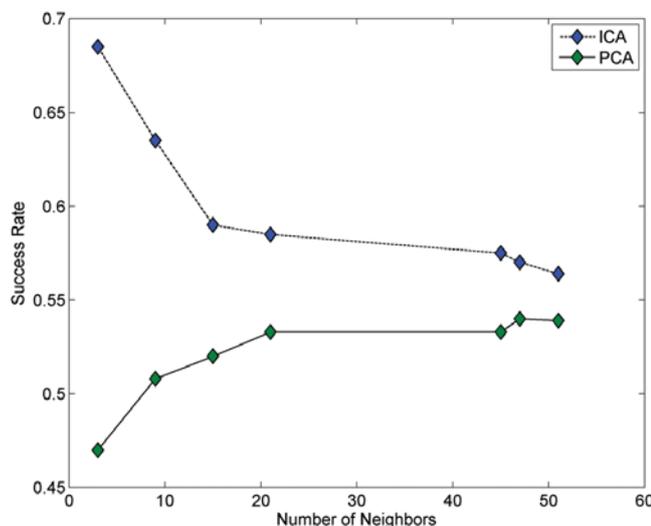
Figure 5 – PCA × ICA Comparative: Even/Odd Test, Classification of lithofacies, All Components.
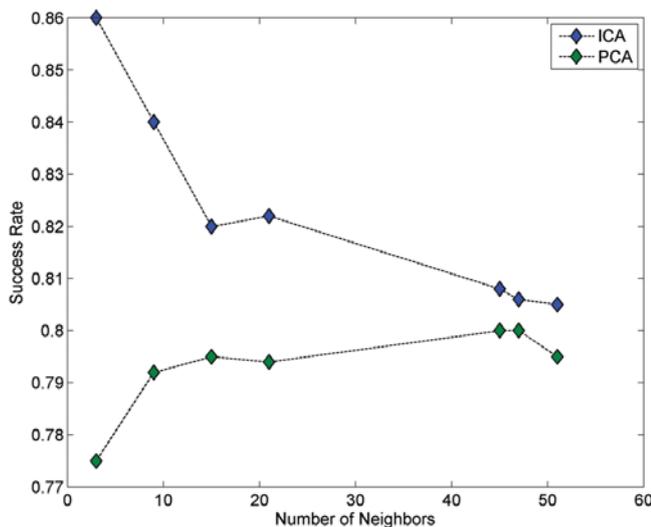

Figure 6 – PCA × ICA Comparative: Even/Odd Test, Classification of characteristics, All Components.

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

ACKNOWLEDGEMENTS

REFERENCES

BARBOZA EG. 2005. Análise Estratigráfica do Campo de Namorado com base na interpretação Sísmica Tridimensional. Doctorate Thesis – UFRGS, Brazil. 230 pp.

CARDOSO JF & SOULOUMIAC A. 1993. Blind beamforming for non Gaussian signals. IEE Proceedings-F, 140(6): 362–370.

CASEY MA. 2001. Method for extracting features from a mixture of signals. Mitsubishi Electric Research Laboratories, Inc., Cambridge, MA, U.S. Patent n. 6,321,200.

COMON P. 1994. Independent Component Analysis: a new concept? Signal Processing, 36(3): 287–314.

COVER TM & HART PE. 1967. Nearest neighbor pattern classification. In: IEEE Transactions on Information Theory, 13(1): 21–27.

DASARATHY BV. 1991. Nearest Neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos, Calif. Washington. pp. 211–215.

FLEXA RT, ANDRADE A & CARRASQUILLA A. 2004. Identificação de litotipos nos perfis de poço do Campo de Namorado (Bacia de Campos, Brasil) e do Lago Maracaibo (Venezuela) aplicando estatística multivariada. Revista Brasileira de Geociências, 34(4): 571–578.

FRANÇA MM. 2009. Análise do uso da terra no município de Viçosa-MG mediado por classificações supervisionadas com Redes Neurais Artificiais e Maxver. Revista Brasileira de Geografia Física, 2(3): 92–101.

GIANNAKOPOULOS X, KARHUNEN J & OJA E. 1999. A Comparison of neural ICA algorithms using Real-world data. In: International Joint Conference on Neural Networks, Washington, DC, 2: 888–893.

HYVÄRINEN A. 1999. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. In: IEEE Transactions on Neural Networks, 10(3): 626–634.

HYVÄRINEN A & OJA E. 1997. A Fast fixed-point algorithm for Independent Component Analysis. Neural Computation, 9(7): 1483–1492.

HYVÄRINEN A, KARHUNEN J & OJA E. 2001. Independent Component Analysis. Wiley and Sons, New York, 504 pp.

KUHN HW & TUCKER AW. 1951. Nonlinear programming. In: Proceedings of Berkeley Symposium, 2., Berkeley: University of California Press, pp. 481–492.

LEITE EP & SOUZA FILHO CR. 2009. TEXTNN – A MATLAB program for textural classification using neural networks. Computers & Geosciences, 35(10): 2084–2094.

MACLEOD AJ. 1984. A generalization of Newton-Raphson. Int. J. Math. Educ. Sci. Technol., 15(1): 117–120.

MacQUEEN JB. 1967. Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of Berkeley Symposium on Mathematical Statistics and Probability, 5., Berkeley, University of California Press, 1: 281–297.

MARCHINI JL, HEATON C & RIPLEY BD. 2009. FastICA Algorithms to perform ICA and Projection Pursuit. R package version 1.1-11.

MURATA N, IKEDA S & ZIEHE A. 2001. An approach to blind source separation based on temporal structure of speech signals. Neurocomputing, 41: 1–24.

ROSA H, SUSLICK SB, VIDAL AC & SAKAI GK. 2008. Caracterização de eletrofácies por meio de ferramentas estatísticas multivariadas. Rem: Rev. Esc. Minas, 61(4): 415–422.

SANCEVERO SS, REMACRE AZ, VIDAL AC & PORTUGAL RS. 2008. Aplicação de técnicas de estatística multivariada na definição da litologia a partir de perfis geofísicos de poços. Revista Brasileira de Geociências, 38(1): 61–74.

TOUSSAINT GT. 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.

VIDAL AC, SANCEVERO SS, REMACRE AZ & COSTANZO CP. 2007. Modelagem Geoestatística 3D da Impedância Acústica para a Caracterização do Campo de Namorado. Brazilian Journal of Geophysics, 25(3): 295–305.

ZHANG K & CHAN L. 2006. Dimension reduction as a deflation method in ICA. Signal Processing Letters, IEEE, 13(1): 45–48.

NOTES ABOUT THE AUTHORS

Alexandre Cruz Sanchetta is 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 multivariated statistical analysis of well log and geophysical data.

Emilson 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 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 Campane Vidal received a B.Sc. degree in Geology from the Universidade de São Paulo (1993), a M.Sc. degree from 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.