Revista Brasileira de Geofísica
RBGf
Brazilian Journal of Geophysics

Revista Brasileira de Geofísica (2011) 29(3): 497-509
© 2011 Sociedade Brasileira de Geofísica
ISSN 0102-261X
www.scielo.br/rbg

COMPARAÇÃO ENTRE AS TÉCNICAS MULTIVARIADAS MAF E PCA APLICADAS NA CLASSIFICAÇÃO DE ELETROFÁCIES

Rodrigo Duarte Drummond[1] e Alexandre Campane Vidal[2]

Recebido em 8 abril, 2010 / Aceito em 12 agosto, 2011
Received on April 8, 2010 / Accepted on August 12, 2011

ABSTRACT. A vast amount of data is obtained during the development of a petroleum field. Seismic data, well logs, core and production data, all contribute to a better reservoir characterization and modeling. Several methods of multivariate data analysis can be used to support its interpretation, helping in important tasks as the identification of lithological facies. The most used and widely known of those methods is Principal Component Analysis (PCA) which intends to reduce data dimension while keeping as much as possible of their variance. Data dimension reduction can also be performed with the method of Maximum Autocorrelation Factors (MAF) which seeks to keep the spatial autocorrelation in data. In this work both methods were applied to data from well logs of the Namorado field, testing their performances in the classification of electrofacies. Following data dimension reduction, supervised classification methods known as k-nearest neighbors (k-NN) and weighted k-nearest neighbors (wk-NN) were applied, and the results obtained were compared by cross-validation. MAF showed to be more efficient than PCA in reducing data dimension, while keeping relevant information. The wk-NN performed a little better in classifying electrofacies than the usual k-NN. According to these results, the combination of MAF and wk-NN can be a valuable tool for classifying the facies of uncored wells from their logs.

Keywords: electrofacies, MAF, PCA, k-NN.

RESUMO. Uma grande quantidade de dados é obtida durante o desenvolvimento de um campo de petróleo. Dados sísmicos, perfis de poços, dados de testemunho e de produção, todos contribuem para uma melhor caracterização e modelagem dos reservatórios. Vários métodos de análise multivariada de dados podem ser usados na interpretação destes dados, auxiliando em tarefas importantes como a identificação de fácies litológicas. O mais conhecido e largamente utilizado dentre estes métodos é a Análise de Componentes Principais (PCA, *Principal Component Analysis*), a qual busca reduzir a dimensão dos dados mantendo o máximo possível de sua variância. A redução da dimensão dos dados também pode ser realizada com o método de Fatores de Máxima Autocorrelação (MAF, *Maximum Autocorrelation Factors*), o qual procura manter o máximo possível da autocorrelação espacial presente nos dados. Neste trabalho, ambos os métodos foram aplicados a dados de perfis de poços do Campo de Namorado, testando seu desempenho na classificação de eletrofácies. Após a redução na dimensão dos dados, os métodos de classificação supervisionada conhecidos como k-vizinhos mais próximos (k-NN, *k-nearest neighbors*) e k-vizinhos mais próximos ponderados (wk-NN, *weighted k-nearest neighbors*) foram aplicados aos dados, e os resultados obtidos foram comparados por validação cruzada. MAF mostrou ser mais eficiente que PCA para reduzir a dimensão dos dados mantendo a informação mais relevante e wk-NN teve desempenho um pouco melhor na classificação de eletrofácies em relação ao k-NN usual. De acordo com esses resultados, a combinação de MAF e wk-NN pode ser uma ferramenta valiosa para a classificação de fácies em poços não testemunhados a partir de seus perfis.

Palavras-chave: eletrofácies, MAF, PCA, k-NN.

[1]UNICAMP, Centro de Estudos do Petróleo, Rua 06 de Agosto, 50, Caixa Postal 6052, Cidade Universitária "Zeferino Vaz", 13083-970 Campinas, SP, Brasil. Tel.: (19) 3521-4659 – E-mail: rdrummond@ige.unicamp.br

[2]UNICAMP, Instituto de Geociências, Departamento de Geologia e Recursos Naturais, Rua Pandiá Calógeras, 51, Caixa Postal 6152, 13083-970 Campinas, SP, Brasil. Tel.: (19) 3521-5198 – E-mail: vidal@ige.unicamp.br

INTRODUÇÃO

Ao longo das várias etapas de trabalho para o desenvolvimento de um campo de petróleo, um grande volume de dados é gerado. São obtidos dados sísmicos, de produção, de poços e de testemunhos. A correta interpretação desses dados depende de um tratamento matemático adequado, capaz de extrair as informações relevantes. Devido à elevada dimensão usualmente encontrada, com diversas variáveis sendo medidas em paralelo, a área da estatística conhecida como análise multivariada fornece as ferramentas adequadas para o tratamento desses dados. De forma mais específica, tais ferramentas são de grande valia na definição da litologia ao longo do poço, tarefa de fundamental importância, pois a determinação incorreta pode acarretar na propagação do erro às etapas subsequentes de geração do modelo geológico.

Os estudos relacionados à caracterização de eletrofácies por meio de perfis de poços e de dados geológicos diretos são alvos de grande número de trabalhos (Serra, 1986a, b; Doveton, 1994; Rider, 2000). Os métodos usualmente aplicados incluem a utilização de métodos estatísticos multivariados (Sancevero et al., 2008), os quais permitem a operação com grande diversidade de variáveis, como dados de perfis geofísicos de poços, associadas aos dados de testemunho. Na literatura destacam-se contribuições importantes na classificação de fácies obtidas com a utilização de ferramentas estatísticas de análise multivariada, tais como a análise discriminante (Bucheb, 1991; Avseth et al., 2001; Flexa et al., 2004; Tang et al., 2004; Li & Anderson-Sprecher, 2006).

Além da dificuldade em associar diferentes perfis de poços para a definição de eletrofácies, outro problema é relacionado à representatividade dos dados de testemunhos, que são provenientes apenas de alguns intervalos de profundidade, enquanto os perfis são corridos ao longo de todos os poços. Estes últimos são, portanto, mais indicados para a aplicação de técnicas multivariadas que permitam realizar inferências através da redução da dimensão dos dados e da sua classificação, utilizando os dados de testemunho como fonte de informação *a priori* sobre as fácies encontradas.

O objetivo principal desse trabalho é a utilização de uma técnica recente, aplicada a dados geológicos, denominada por Fatores de Máxima Autocorrelação (MAF – *Maximum Autocorrelation Factors*; Switzer & Green, 1984), que considera a estrutura espacial dos dados. Para a análise dos resultados foi realizada a comparação com os dados gerados pelo mesmo procedimento de redução de dimensão pelo método de Análise de Componentes Principais (PCA; Jolliffe, 2004), provavelmente uma das técnicas de análise multivariada mais antiga e conhecida. Além disso, os métodos de classificação supervisionada conhecidos como k-vizinhos mais próximos e k-vizinhos mais próximos ponderados foram utilizados para avaliar a eficiência da redução de dimensão obtida por cada uma das técnicas. Por fim, através de validação cruzada, obteve-se uma comparação tanto das técnicas de redução de dimensão quanto dos métodos classificatórios como estratégias de análise de dados de poço e testemunho.

Neste trabalho, a análise estatística foi aplicada aos dados de perfis de poços do Campo de Namorado da Bacia de Campos, cuja base de dados é disponibilizada pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP).

METODOLOGIA

Revisão teórica

Análise de componentes principais

Conhecida como PCA, esta é uma das mais antigas técnicas de análise de dados multivariados, tendo sido proposta por Pearson (1901) e posteriormente desenvolvida por Hotelling (1933). A ideia central do método é reduzir a complexidade de um conjunto de dados, constituído de diversas variáveis inter-relacionadas, através da redução de sua dimensão, mantendo o máximo possível da informação. Isto é feito através da transformação linear para um novo conjunto de variáveis, os componentes principais, os quais são não correlacionados e ordenados de forma que os primeiros retenham a maior parte da variância presente nas variáveis originais (Jolliffe, 2004).

Fatores de máxima autocorrelação

Os fatores de máxima autocorrelação, conhecidos como MAF, foram propostos por Switzer & Green (1984) para a análise de dados que apresentam estrutura espacial. A ideia do método é isolar o sinal e o ruído presentes nos dados a partir da premissa que o sinal possui uma correlação espacial maior que a do ruído. Com isso pode-se reduzir a dimensão dos dados mantendo o máximo possível dos sinais de interesse.

A análise de MAF requer o conhecimento prévio ou uma estimativa tanto da matriz de variância-covariância dos dados como da matriz de variância-covariância da diferença entre os dados originais e uma versão dos mesmos deslocada espacialmente. A técnica pode ser então formulada como um problema de análise de correlação canônica (Hotelling, 1936). Para isso, considera-se que os dados observados encontram-se organizados em

uma matriz $\mathbf{X}$, onde cada linha corresponde a uma realização de uma variável aleatória multivariada $X = [x_k]$ (onde $k = 1 \ldots p$). A estimativa da matriz de variância-covariância, Σ, pode ser obtida diretamente dos dados, como é usual em aplicações de PCA. Para obter a estimativa da variância-covariância das diferenças entre os dados e sua versão deslocada, é preciso determinar um deslocamento espacial Δ. A princípio, será adotado que este deslocamento esteja definido e ao final deste item será abordado o problema de determiná-lo. Será considerada também a matriz $\mathbf{X}_\Delta$ dos dados deslocados espacialmente, isto é, suas colunas $x_{\Delta j}$ são determinadas por

$$x_{\Delta j}(k) = x_j(k + \Delta)\,,$$

para todo ponto k tal que $x_j(k + \Delta)$ está definido (x_j pode ser visto aqui como a j-ésima coluna da matriz $\mathbf{X}$ ou como a variável aleatória correspondente a essa coluna). Assumindo "p" combinações lineares das variáveis originais, $y_1, \ldots, y_p$ com $y_i = a_i^t X$, $i = 1 \ldots p$, e definindo seus valores deslocados $y_{\Delta j}$ de maneira análoga aos $x_{\Delta j}$, isto é,

$$y_{\Delta i}(k) = y_i(k + \Delta)\,,$$

estas combinações serão os fatores de máxima autocorrelação se suas correlações com os deslocamentos $y_{\Delta i}(k)$ satisfizerem as seguintes propriedades:

$$corr(y_1, y_{\Delta 1}) = \max_a\ corr(a^t X, a^t X_\Delta)$$

$$corr(y_1, y_{\Delta i}) = \max_a\ corr(a^t X, a^t X_\Delta)$$

$$\text{restrito a } corr(y_i, y_j) = 0 \text{ para } j < i$$

$$corr(y_p, y_{\Delta p}) = \min_a\ corr(a^t X, a^t X_\Delta)$$

$$\text{com } corr(y_p, y_j) = 0 \text{ para } j < p\,.$$

Para obter os fatores de máxima autocorrelação é preciso adotar algumas definições: sejam

$$\Gamma(\Delta) = Cov\{X, X_\Delta\} \text{ e } \Sigma_\Delta = Cov\{X - X_\Delta\}\,,$$

respectivamente a matriz de covariâncias entre as variáveis originais e seus deslocamentos e a matriz de variância-covariância da diferença entre os dados originais e os deslocamentos. Supondo então que as variáveis apresentem estacionaridade de segunda ordem, tem-se que $\Gamma(-\Delta) = \Gamma(\Delta)^t$. Além disso, com alguma manipulação algébrica das matrizes, obtém-se que

$$\Sigma_\Delta = 2\Sigma - \Gamma(\Delta) - \Gamma(-\Delta)\,.$$

Com esses resultados, pode-se agora obter a covariância entre uma combinação linear das variáveis originais e seu deslocamento:

$$\begin{aligned} Cov\{a_1^t X, a_1^t X_\Delta\} &= a_1^t \Gamma(\Delta) a_1 \\ &= a_1^t \Gamma^t(\Delta) a_1 \\ &= a_1^t \Gamma(-\Delta) a_1 \\ &= \frac{1}{2} a_1^t (\Gamma(\Delta) + \Gamma(-\Delta)) a_1 \\ &= \frac{1}{2} a_1^t (2\Sigma - \Sigma_\Delta) a_1 \\ &= a_1^t \left(\Sigma - \frac{1}{2}\Sigma_\Delta\right) a_1\,. \end{aligned}$$

A suposição de estacionaridade de segunda ordem garante que

$$Var(a_1^t X_\Delta) = Var(a_1^t X) = a_1^t \Sigma a_1,$$

com isso obtém-se a correlação:

$$\begin{aligned} Corr\{a_1^t X, a_1^t X_\Delta\} &= \frac{Cov\{a_1^t X, a_1^t X_\Delta\}}{\sqrt{Var(a_1^t X) \times Var(a_1^t X_\Delta)}} \\ &= \frac{a_1^t \left(\Sigma - \frac{1}{2}\Sigma_\Delta\right) a_1}{a_1^t \Sigma a_1} \\ &= \frac{a_1^t \Sigma a_1 - \frac{1}{2} a_1^t \Sigma_\Delta a_1}{a_1^t \Sigma a_1} \\ &= 1 - \frac{1}{2}\frac{a_1^t \Sigma_\Delta a_1}{a_1^t \Sigma a_1}\,. \end{aligned}$$

Para maximizar a correlação acima, é preciso então minimizar o coeficiente de Rayleigh (Larsen, 2002):

$$R(a_1) = \frac{a_1^t \Sigma_\Delta a_1}{a_1^t \Sigma a_1}\,.$$

Ao observar que o coeficiente de Rayleigh é invariante para multiplicações por escalares, isto é, $R(\lambda a_1) = R(a_1)$, onde λ é uma constante, e supondo que a matriz Σ é não degenerada ($\det(\Sigma) > 0$), podemos restringir o problema de minimização aos vetores a_1 tais que $a_1^t \Sigma a_1 = 1$. Assim precisamos encontrar:

$$\min_{a_1} \frac{a_1^t \Sigma_\Delta a_1}{a_1^t \Sigma a_1}\,,\ \text{restrito a } a_1^t \Sigma a_1 = 1\,.$$

Aplicando o método de multiplicadores de Lagrange ao problema de otimização acima, obtemos que a_1 é um autovetor

de $\Sigma^{-1}\Sigma_\Delta$ associado ao menor autovalor, λ_1. Tomando então $a_1 = e_1$, autovetor unitário associado ao autovalor $\lambda_1 (e_1^t e_1 = 1)$, obtemos o primeiro fator de máxima autocorrelação como $y_1 = e_1^t X$ e suas observações são dadas por $\mathbf{X}e_1$.

O método de determinação dos MAF segue então recursivamente, encontrando fatores que possuem máxima autocorrelação e são não correlacionados aos fatores anteriormente determinados. O segundo MAF é dado por $y_2 = {e_2}^t X$, onde e_2 é o autovetor de $\Sigma^{-1}\Sigma_\Delta$ associado ao segundo menor autovalor, λ_2, e assim sucessivamente: o i-ésimo MAF é dado por $y_i = e_i^t X$, onde e_i é o autovetor de $\Sigma^{-1}\Sigma_\Delta$ associado ao autovalor λ_i, com $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$, e suas observações são dadas por $\mathbf{X}e_i$. A autocorrelação do i-ésimo MAF é dada por

$$Corr\left\{e_i^t X, e_i^t X_\Delta\right\} = 1 - \frac{1}{2}R(e_i) = 1 - \frac{1}{2}\lambda_i \, .$$

Finalmente, para aplicar a técnica, é preciso determinar o deslocamento Δ. Switzer & Green (1984), no trabalho original em que o método de MAF foi proposto, sugerem que para imagens bidimensionais a matriz Σ_Δ seja estimada separadamente para um deslocamento unitário vertical e um deslocamento unitário horizontal, e depois a média dessas duas matrizes seja adotada como Σ_Δ no restante do método. Generalizando essa ideia, em três dimensões seria possível estimar Σ_Δ separadamente para um pequeno deslocamento em cada dimensão e depois tomar a média dessas três matrizes como sendo Σ_Δ nas etapas subsequentes do método.

Vale observar que os fatores de máxima autocorrelação são invariantes a transformações lineares aplicadas aos dados (Switzer & Green, 1984). Esta importante propriedade significa que se $Z = \mathbf{B}X$, onde $\mathbf{B}$ é qualquer matriz não singular $p \times p$, os MAF de Z serão os mesmos de X. Particularmente, este resultado permite aplicar aos dados o pré-processamento conhecido como clareamento (*whitening* ou *sphering*: após subtrair cada variável de sua média, os dados são projetados nos componentes principais obtidos com a PCA e cada um destes componentes é dividido por seu desvio-padrão), antes de obter os MAF. Esse procedimento é interessante porque, após o clareamento, a matriz de variância-covariância dos dados torna-se a matriz identidade ($\Sigma_Z = I_{pxp}$), o que facilita o cálculo dos MAF, pois estes passam a ser fornecidos simplesmente pelos autovetores da matriz Σ_Δ.

A técnica dos MAF foi desenvolvida para dados com característica espaciais e apresentada originalmente aplicada a dados de imagens aéreas multicanais (Switzer & Green, 1984), porém apesar disso o método ainda é pouco conhecido e suas aplicações à geofísica são incipientes. Desbarats & Dimitrakopoulos (2000) aplicaram-no na simulação da distribuição regionalizada do tamanho de poros. Os autores concluíram que o método é uma ferramenta poderosa para a análise e simulação de variáveis corregionalizadas, fornecendo um método de redução de dimensão mais eficiente que o PCA. Aplicações dos MAF também têm sido desenvolvidas em outras áreas da ciência. Larsen (2002) propôs uma extensão deste conceito à análise fatorial modo Q, chamada de Q-MAF e aplicada ao estudo de formatos de ossos humanos.

A técnica de MAF ainda se encontra em desenvolvimento e constitui uma área fértil de pesquisa. Vargas-Guzmán & Dimitrakopoulos (2003) avaliaram as propriedades do método e suas limitações, concluindo que este é equivalente à análise de discriminante da estrutura aninhada dos dados. Os autores observaram ainda que a eficiência do método é reduzida quando os dados apresentam características mais complexas, como três ou mais estruturas aninhadas ou anisotropia, portanto sua aplicação nestes casos requer novos estudos. Bailey e Krzanowski (Bailey & Krzanowski, 2000; Krzanowski & Bailey, 2007) também revisaram os MAF e outros métodos de análise de fatores espaciais, propondo novas extensões dos mesmos e aplicando-as a dados geoquímicos. Bandarian et al. (2008) propuseram uma versão dos MAF, chamada de MAF direto (DMAF), que concluíram ser mais eficiente em modelos lineares com duas estruturas de corregionalização.

Classificação supervisionada (k-NN e k-NN ponderado)

Após a redução de dimensão, é usual automatizar a identificação de padrões por meio de algoritmos de classificação ou aprendizagem de máquina. Estes algoritmos podem se basear em informações disponíveis sobre parte dos dados ou em características esperadas nos mesmos, neste caso sendo chamados de algoritmos supervisionados.

A técnica de k-vizinhos mais próximos (k-NN, Fix & Hodges Jr., 1951; Cover & Hart, 1967) é o método mais simples e intuitivo de classificação. Consiste em observar, para cada ponto "w" a ser classificado, os "k" pontos do conjunto de treino que se encontram mais próximos a w, chamados de vizinhos. A classe mais frequente entre os vizinhos é então associada ao ponto w. O valor de k é usualmente pequeno, para evitar que pontos mais distantes tenham influência sobre o ponto w. Além disso, se k for muito grande a classe mais presente em todo o conjunto de treino acaba por dominar as classificações de todos os pontos. Por outro lado, valores maiores de k reduzem o efeito do ruído

na classificação, isto é, pontos do conjunto de treino classificados erroneamente interferem menos nas classificações. A escolha de k apresenta-se então como um ponto delicado do método. Uma extensão do algoritmo, que visa aumentar sua eficiência e contornar o problema da escolha do valor de k, é o método de k-NN ponderado (Hechenbichler & Schliep, 2004).

A técnica de k-NN ponderado é uma alternativa ao k-NN, na qual a determinação da classe de cada ponto não leva em conta apenas as classes dos k vizinhos mais próximos dentre os pontos do conjunto de treino, mas também a distância de cada um destes vizinhos ao ponto em questão. Para isso, após obter as distâncias dos vizinhos ao ponto, uma função de peso é aplicada a essas distâncias para determinar o quanto a classe associada a cada vizinho vai ser determinante na classificação do ponto. Por fim, os pesos dos vizinhos pertencentes a cada classe são somados e associa-se ao ponto a classe onde essa soma é maior. Com isso, os vizinhos mais próximos são mais determinantes na classificação, e ainda que k seja grande a influência de vizinhos mais distantes na mesma será reduzida. Dentre as funções de peso utilizadas, podemos citar a função inversa, a linear e a função gaussiana.

Conjunto de dados

O conjunto de dados utilizado no trabalho é proveniente do Campo de Namorado. Descoberto em 1975, o Campo de Namorado está localizado na parte central norte da zona de acumulação de hidrocarbonetos da Bacia de Campos, a 80 km da costa, inserido na seção de calcilutitos, margas e folhelhos da Formação Macaé, membros Outeiro e Quissamã. O reservatório está posicionado em estrutura alongada de direção NW-SE, associado a depósitos em canais e em lobos, intercalados por sedimentos hemipelágicos.

As informações obtidas da descrição sequencial de testemunho evidenciam a complexidade geológica desse reservatório, que é composto pela predominância de arenitos e folhelhos e, secundariamente, por conglomerados, brechas, siltitos e margas. De acordo com as descrições de testemunho, são observadas, no total, 21 fácies litológicas.

A rocha reservatório é constituída por arenitos arcosianos de granulometria fina a grossa, denominados de arenitos turbidíticos de Namorado (Johann, 1997). Duas principais fácies de arenitos constituem a rocha reservatório, que apresenta espessuras métricas nos testemunhos. A fácies de maior ocorrência nos testemunhos corresponde a fácies arenito médio maciço, arcoseano e bem selecionado. A segunda fácies são arenitos grossos amalgamados, com gradação da fração areia grossa-conglomerática na base para fração grossa no topo.

Nos últimos anos, vários foram os trabalhos que objetivaram o estudo do Campo de Namorado (Johann, 1997; Souza Jr., 1997; Barboza, 2005). Atualmente os dados referentes ao Campo de Namorado são disponibilizados pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Os perfis de poços disponibilizados incluem a densidade da formação (RHOB), a porosidade neutrão (NPHI), o perfil de raios gama (GR) e o sônico (DT).

Neste trabalho foram utilizados dados de sete poços para os quais, além dos perfis mencionados, também foram disponibilizadas análises sequenciais de testemunhos. Esses poços são identificados pelos códigos: NA01, NA02, NA04, NA07, NA11A, RJS42 e RJS234. As fácies identificadas nos testemunhos encontram-se na Tabela 1.

Processamento de dados

As análises realizadas neste trabalho foram feitas no ambiente estatístico R (*R Development Core Team*, 2009), um *software* gratuito voltado para a computação estatística. As técnicas de PCA e MAF foram programadas usando funções do pacote básico disponibilizado com o *software*.

A técnica de MAF foi baseada apenas no deslocamento (Δ) na direção vertical, com tamanho igual à menor distância entre as observações dos perfis de poço (0,2 m). Com isso, a técnica buscou os fatores cuja autocorrelação nessa direção é máxima. Devido à grande distância entre os poços, haveria pouco sentido em considerar as direções horizontais no cálculo da autocorrelação espacial.

As classificações de fácies também foram feitas em R. As fácies observadas nos testemunhos foram divididas em três classes: reservatório, possível reservatório e não reservatório, de acordo com a Tabela 1. Estas classes foram definidas conforme o potencial como rocha reservatório, estabelecido com base na descrição dos testemunhos e nos dados petrofísicos.

Os testemunhos foram então utilizados como conjunto de treino para a divisão dos demais pontos dos poços estudados nessas três classes. A medida de distância entre os pontos utilizada nas classificações foi a distância euclidiana, porém vale observar que, como os dados foram submetidos ao pré-processamento de clareamento (no caso do PCA cada componente foi dividido por seu desvio antes da classificação) essa distância é equivalente à distância de Mahalanobis (Mahalanobis, 1936).

As classificações por k-vizinhos mais próximos e k-NN ponderado foram realizadas com funções disponíveis no pacote *kknn*

Tabela 1 – Classificação das litofácies descritas nos testemunhos e as três classes utilizadas neste trabalho.

Fácies	Classificação
Arenito Grosso Amalgamado	Reservatório
Arenito Médio Gradado ou Maciço	
Arenito Médio Fino Laminado	Possíveis reservatórios
Arenito/Folhelho Interestratificado	
Arenito/Folhelho Finamente Interestratificado	
Conglomerados Residuais	
Interlaminado Lamoso Deformado	Não reservatórios
Conglomerados e Brechas Carbonáticas	
Diamictito Arenoso Lamoso	
Arenito Médio Cimentado	
Siltito Argiloso Estratificado	
Interlaminado Siltito Argiloso e Marga	
Interlaminado Arenoso Bioturbado	
Interlaminado de Siltito e Folhelho, Deformado, Bioturbado	
Marga Bioturbada	
Ritmito	
Folhelho Siltico com Níveis de Marga Bioturbada	
Arenito Cimentado, com Feições de Escorregamento	
Siltito Argiloso/Arenito Deformado	
Arenito Médio/Fino Laminado Cimentado	
Interestratificado Siltito/Folhelho Intensamente Bioturbados	

(Schliep & Hechenbichler, 2009) do R. As classificações foram aplicadas separadamente aos resultados obtidos com cada um dos métodos de redução de dimensão, utilizando um, dois ou três dos componentes encontrados por cada método. Em cada caso, o método de k-NN foi testado utilizando de 3 a 14 vizinhos mais próximos. O método de k-NN ponderado foi testado com esses mesmos valores de k, porém com as funções de peso triangular, inversa ou gaussiana. A classificação obtida em cada um desses cenários foi avaliada por validação cruzada, isto é, cada um dos testemunhos foi separado dos demais e estes foram usados como conjunto de treino para repetir a classificação, obtendo assim uma classe para o testemunho isolado e comparando-a com a sua classe original. Esse procedimento foi repetido para todos os testemunhos e as porcentagens de acerto da validação foram computadas para cada cenário.

RESULTADOS E DISCUSSÃO

Os resultados obtidos com cada uma das técnicas de redução de dimensão podem ser observados na Figura 1, onde os dados foram projetados nos dois primeiros componentes obtidos com cada técnica. Ao observar a figura é possível constatar que o MAF separa os pontos correspondentes a reservatórios e a possíveis reservatórios melhor que o PCA e apresenta menor dispersão em relação à classe reservatório.

Outra forma visual de avaliar os resultados de cada uma das técnicas de redução de dimensão é comparar o primeiro componente encontrado por cada uma delas com os testemunhos dos poços dispostos em profundidade (Figs. 2 e 3). Esses gráficos auxiliam na avaliação do quanto os componentes identificados pelas técnicas são eficientes em separar as fácies reservatório das não reservatório. Observando-os fica claro que o componente que melhor diferencia os reservatórios é o identificado pelo MAF (Fig. 3): nos poços NA04 e NA011A, por exemplo, valores menores desse componente identificam reservatórios e possíveis reservatórios.

A análise dos valores das variâncias dos componentes encontrados pela técnica de PCA (Fig. 4B) mostra que os dois primeiros componentes concentram 94,52% da variância total dos dados, enquanto os três primeiros componentes representam 99,99% da mesma. Além disso, o decréscimo nos valores das variâncias (Fig. 4A) é aproximadamente constante até o terceiro componente, após o qual se torna significativamente menor. De acordo com a estratégia proposta por Cattell (1966) para

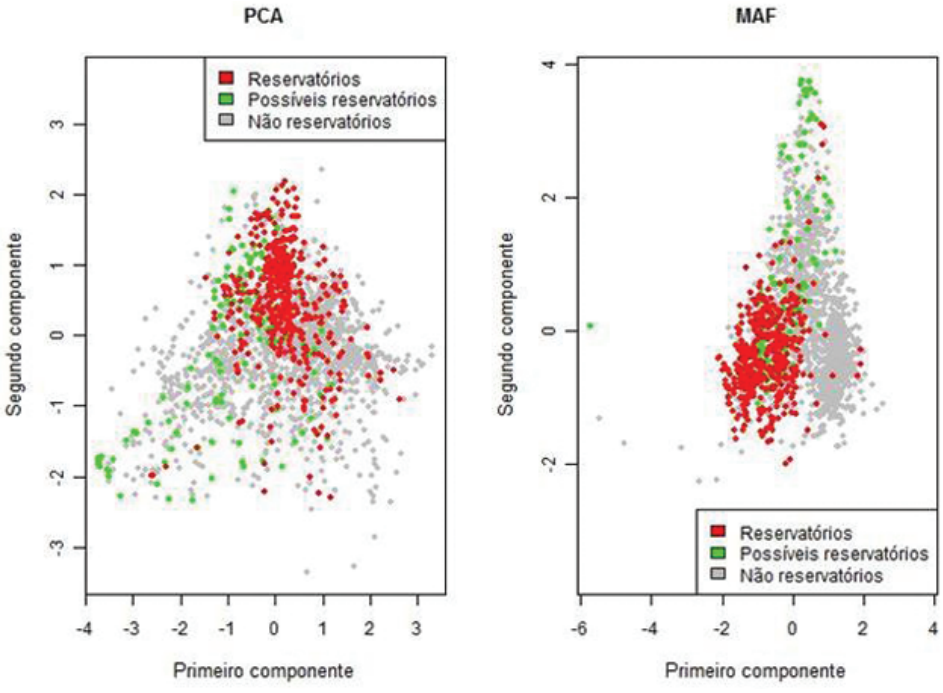

Figura 1 – Dispersão obtida com a projeção dos dados nos dois primeiros componentes encontrados por cada método.

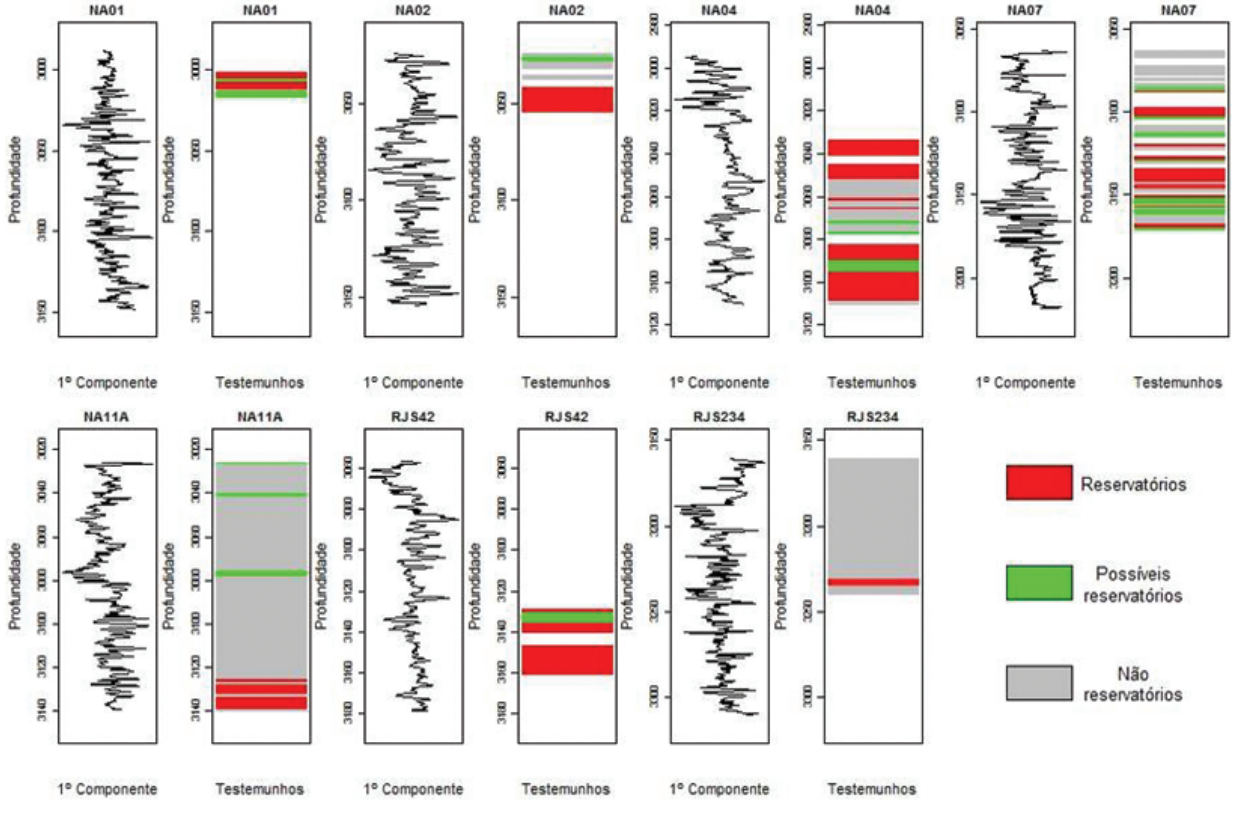

Figura 2 – A cada poço, o perfil do primeiro componente obtido com PCA em comparação com os testemunhos.

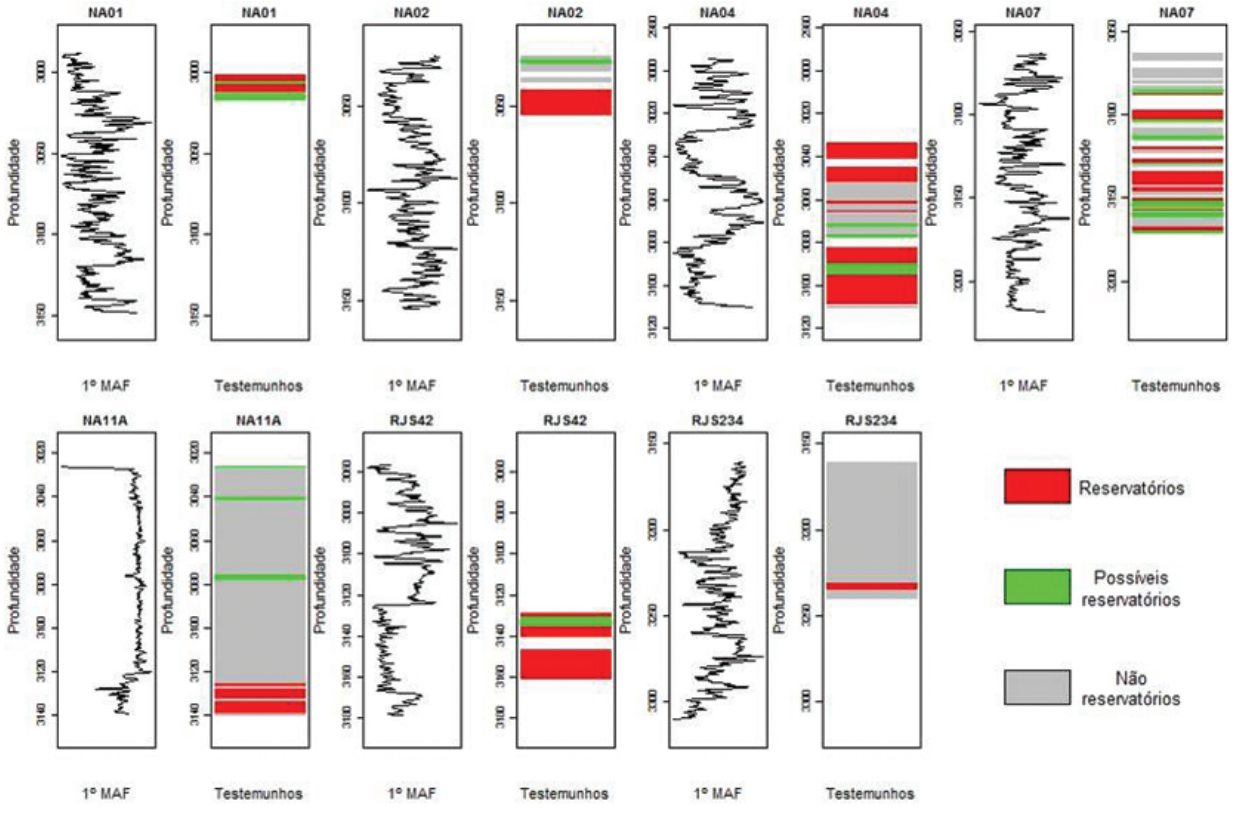

Figura 3 – A cada poço, o perfil do primeiro componente obtido com MAF em comparação com os testemunhos.

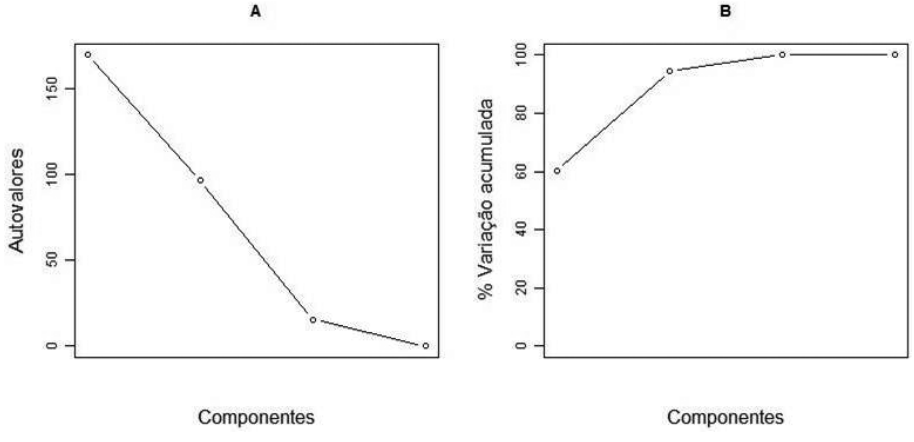

Figura 4 – Variâncias de cada componente encontrado por PCA (A) e porcentagem da variância total representada pelos primeiros componentes determinados pela técnica (B).

determinar o número de componentes principais a serem selecionados, esses dois resultados levam a concluir que os três primeiros componentes são os mais indicados para prosseguir com as análises. Ainda assim, neste trabalho testamos os resultados de classificações baseadas em um, dois ou três dos primeiros componentes encontrados por cada técnica de redução de dimensão.

Analisando os resultados da validação cruzada das classificações baseadas em um, dois ou três dos primeiros componentes identificados por cada técnica de redução de dimensão (Figs. 5 a 7) nota-se que o MAF apresentou resultados melhores em todos os casos. Fica claro também que quanto mais componentes são utilizados nas classificações, mais acuradas estas se tornam, o que seria de se esperar uma vez que uma porcentagem maior

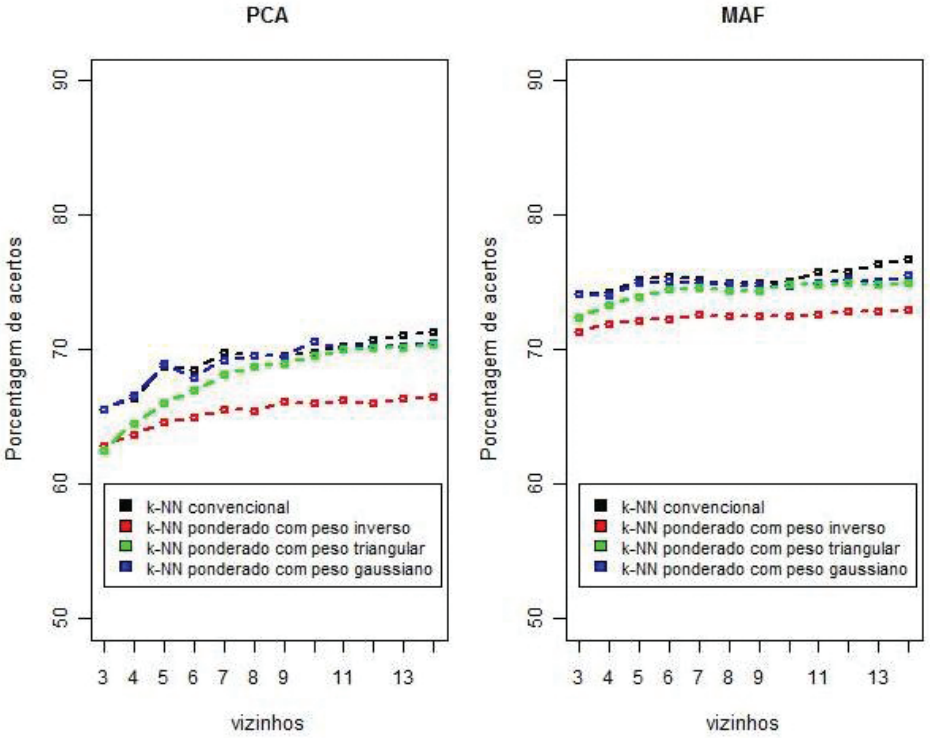

Figura 5 – Resultados da validação cruzada das classificações obtidas com k-NN e k-NN ponderado, baseadas no primeiro componente encontrado por cada técnica de redução de dimensão.

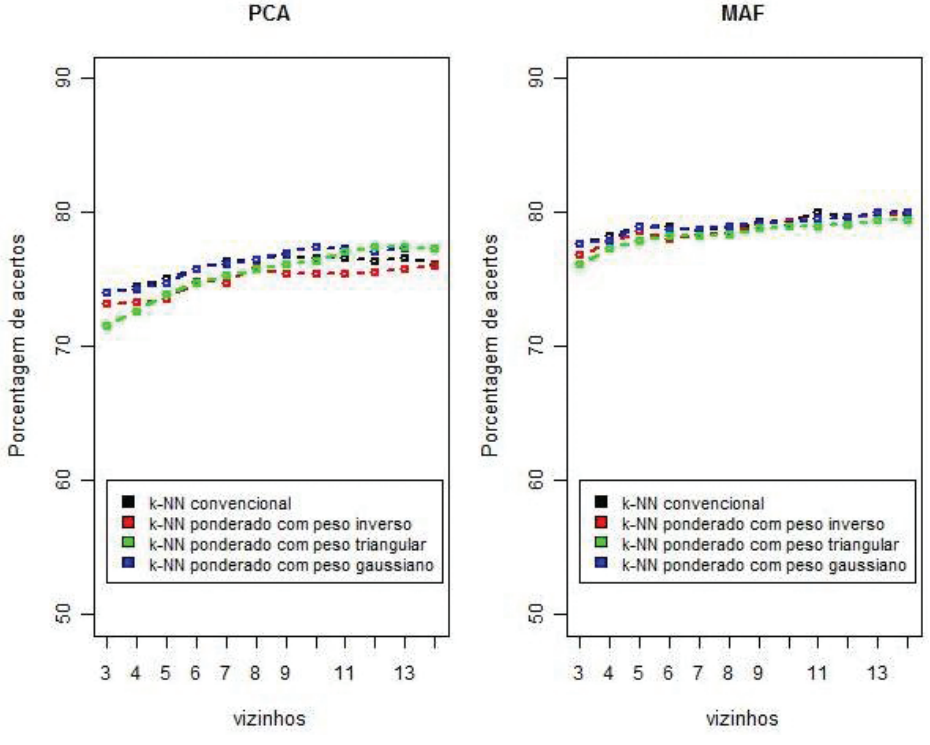

Figura 6 – Resultados da validação cruzada das classificações obtidas com k-NN e k-NN ponderado, baseadas nos dois primeiros componentes encontrados por cada técnica de redução de dimensão.

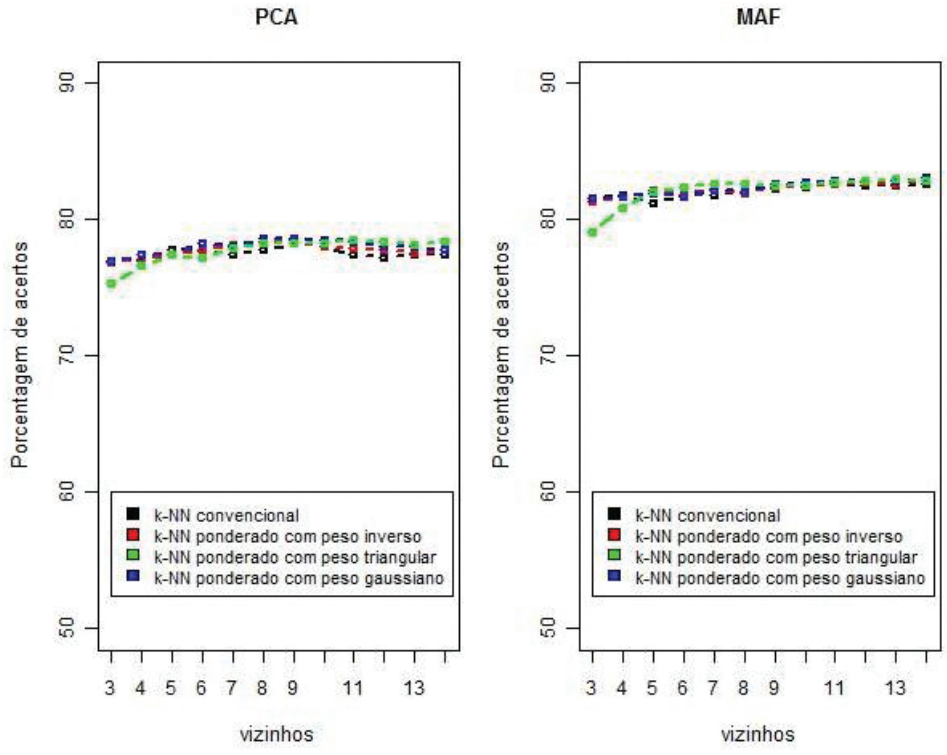

Figura 7 – Resultados da validação cruzada das classificações obtidas com k-NN e k-NN ponderado, baseadas nos três primeiros componentes encontrados por cada técnica de redução de dimensão.

da variância é mantida nos componentes utilizados (isso implica, porém, num custo computacional maior das classificações). Os melhores resultados são obtidos utilizando-se três componentes encontrados por cada técnica, o que está de acordo com as previsões provenientes da análise da Figura 4.

Comparando os métodos de classificação, observa-se que o k-NN e o k-NN ponderado apresentam resultados semelhantes, porém o k-NN ponderado com função de peso triangular apresenta ligeira vantagem quando 12 ou mais vizinhos são utilizados. Em todas as versões do k-NN, os resultados melhoram quando mais vizinhos são utilizados, porém essa melhora atinge um patamar a partir de 12 vizinhos. Isso leva a concluir que é preferível utilizar 12 vizinhos que valores superiores de k, pois o custo computacional envolvido nas classificações aumenta com k.

De acordo com os resultados da validação cruzada das classificações, o MAF mostrou-se a técnica de redução de dimensão mais eficiente. Além disso, para ambas as técnicas, a classificação por k-NN ponderado com função de peso triangular utilizando 12 vizinhos, baseada nos três primeiros componentes, apresentou os melhores resultados. Esse método de classificação, combinado com MAF e PCA, obteve respectivamente 82,8% e 78,4% de acerto na validação cruzada. Utilizamos então esse método de classificação, baseado nos três primeiros componentes encontrados por cada uma das técnicas de redução de dimensão, para classificar todos os pontos dos poços estudados (utilizando todos os testemunhos como conjunto de treino). Os resultados podem ser vistos nas Figuras 8 e 9.

CONCLUSÃO

As técnicas de redução de dimensão, aliadas a métodos de classificação, fornecem ferramentas poderosas para a identificação de fácies a partir de perfis de poços. Isto as torna de grande interesse econômico, uma vez que a correta identificação de fácies é de vital importância na modelagem geológica de reservatórios. Dentre as estratégias de análise testadas neste trabalho, a aplicação de MAF para reduzir a dimensão dos dados de quatro (RHOB, NPHI, GR e DT) para três, aliada à classificação por k-NN ponderado com função de peso triangular utilizando 12 vizinhos, foi a que obteve melhores resultados, atingindo 82,8% de acerto na validação cruzada. O melhor desempenho do MAF, em relação à PCA, deve-se ao fato deste método levar em consideração a estrutura espacial dos dados, característica fundamental dos perfis de poços estudados e que é ignorada pela PCA. O nível de precisão obtida na

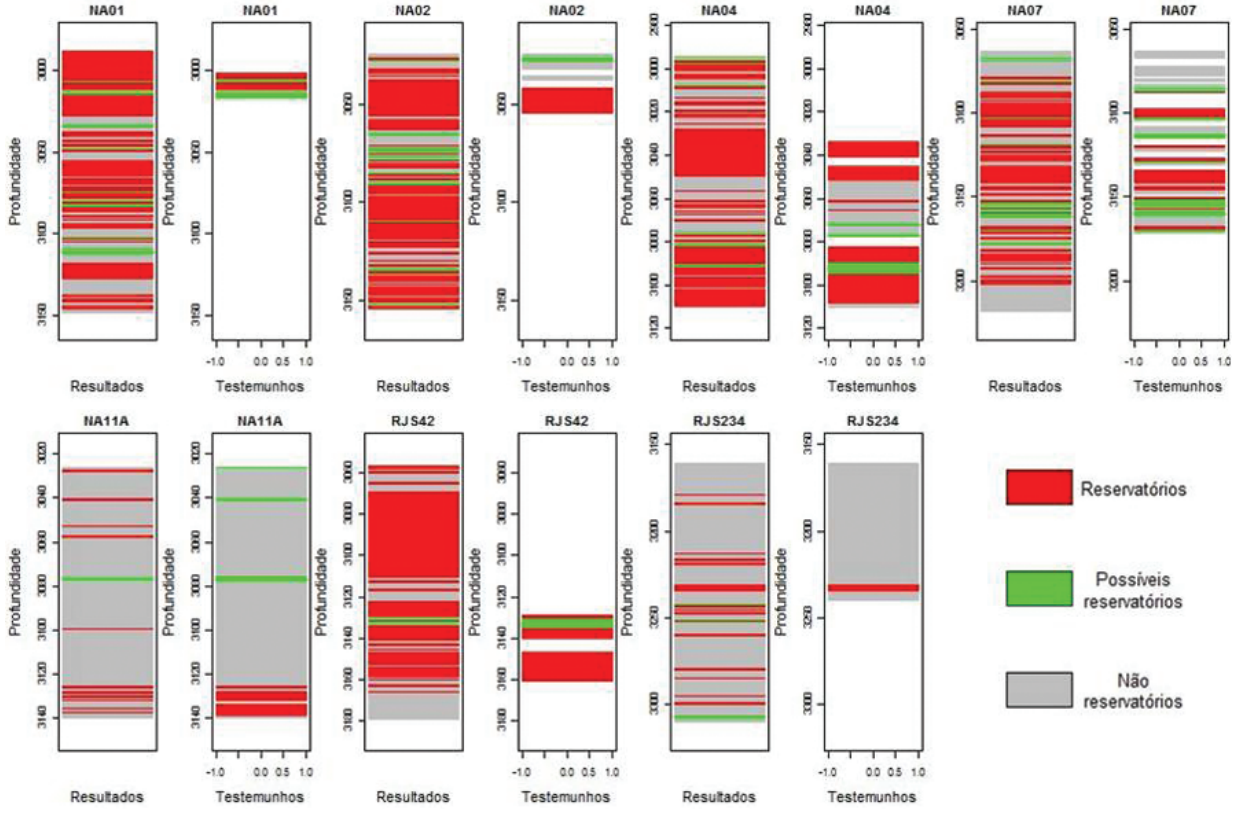

Figura 8 – Classificação das fácies ao longo dos poços, obtida com k-NN ponderado por função triangular ($k = 12$), baseada nos três primeiros componentes encontrados por PCA.

classificação mostra o quanto estas técnicas podem ser valiosas na interpretação de dados geológicos diretos e de perfis de poços.

Os resultados obtidos neste trabalho levam a supor que as técnicas de análise estudadas também terão bom desempenho quando aplicadas a outros tipos de dados geofísicos, como dados sísmicos. Porém, quando não houver informações anteriores sobre as fácies representadas, como é usual em dados sísmicos, os métodos de classificação supervisionados terão que ser substituídos por métodos não supervisionados.

REFERÊNCIAS

AVSETH P, MUKERJI T, JORSTAD A, MAVKO G & VEGGELAND T. 2001. Seismic reservoir mapping from 3-D AVO in a North Sea turbidite system. Geophysics, 66(4): 1157–1176.

BAILEY TC & KRZANOWSKI WJ. 2000. Extensions to Spatial Factor Methods with an Illustration in Geochemistry. Math. Geol., 32(6): 657–682.

BANDARIAN EM, BLOOM LM & MUELLER UA. 2008. Direct minimum/maximum autocorrelation factors within the framework of a two structure linear model of coregionalisation. Comput. Geosci., 34: 190–200.

BARBOZA EG. 2005. Análise estratigráfica do Campo de Namorado (Bacia de Campos) com base na interpretação sísmica tridimensional. Tese de doutorado em Geociências, Universidade Federal do Rio Grande do Sul, 167 p.

BUCHEB JA. 1991. Aplicação de tratamento estatístico multivariante em dados de perfis de poços da Bacia de Sergipe, Alagoas. Tese de Mestrado, CG/UFPA, Belém, 136 p.

CATTELL RB. 1966. The scree test for the number of factors. Multiv. Behav. Res., 1: 245–276.

COVER TM & HART PE. 1967. Nearest Neighbor pattern classification. IEEE Trans. Inf. Theory, IT-13(1): 21–27.

DESBARATS AJ & DIMITRAKOPOULOS R. 2000. Geostatistical simulation of regionalized pore-size distributions using Min/Max Autocorrelation Factors. Math. Geol., 32(8): 919–942.

DOVETON JH. 1994. Geologic log analysis using computer methods. Am. Assoc. Petroleum Geologists, Computer Applications in Geology, n. 2, 169 p.

FIX E & HODGES Jr JL. 1951. Discriminatory analysis, nonparametric discrimination: Consistency properties. Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine.

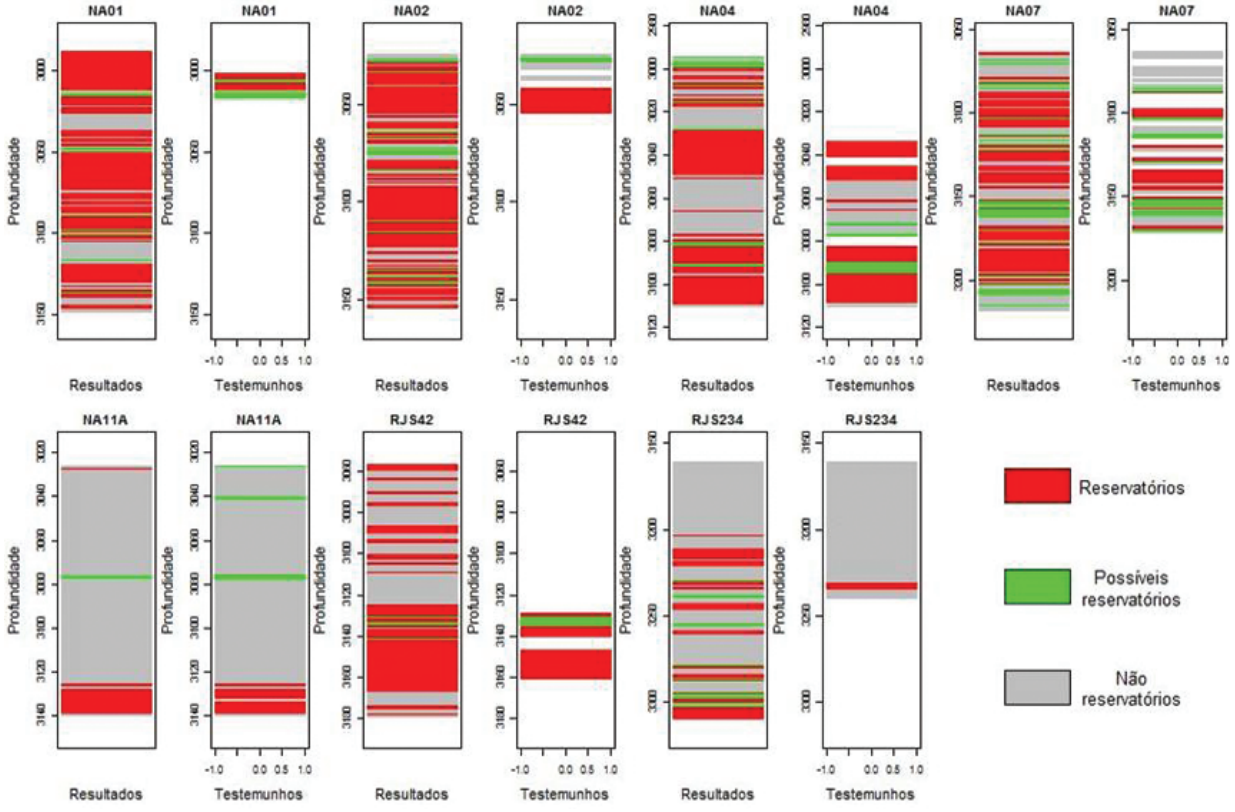

Figura 9 – Classificação das fácies ao longo dos poços, obtida com k-NN ponderado por função triangular ($k = 12$), baseada nos três primeiros componentes encontrados por MAF.

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.

HECHENBICHLER K & SCHLIEP K. 2004. Weighted k-Nearest-Neighbor techniques and ordinal classification. Collaborative Research Center 386, Discussion Paper 399, University of Munich, 16 p.

HOTELLING H. 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol., 24: 417–441, 498–520.

HOTELLING H. 1936. Relations between two sets of variables. Biometrika, 28: 321–377.

JOHANN PRS. 1997. Inversion Sismostratigraphique et Simulations Stochastiques en 3D: Réservoir Turbiditique, Offshore du Brésil. Intégration Géologique, Géophysique et Géostatistique. Thèse de Doctorat. Université Pierre et Marie Curie, Paris, 352 p.

JOLLIFFE IT. 2004. Principal Component Analysis, Springer Series in Statistics. 2nd edition, Springer, New York, 487 p.

KRZANOWSKI WJ & BAILEY TC. 2007. Extraction of spatial features using factor methods illustrated on stream sediment data. Math. Geol., 39(1): 69–85.

LARSEN R. 2002. Decomposition using maximum autocorrelation factors. J. Chemometr., 16: 427–435.

LI Y & ANDERSON-SPRECHER R. 2006. Facies identification from well logs: A comparison of discriminant analysis and naïve bayes classifier. J. Petrol. Sci. Eng., 53: 149–157.

MAHALANOBIS PC. 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India, 2(1): 49–55.

PEARSON K. 1901. On lines and planes of closest fit to systems of points in space. Phil. Mag., 2: 559–572.

R DEVELOPMENT CORE TEAM. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.R-project.org>. Acesso em: 19 jan. 2010.

RIDER M. 2000. The Geological Interpretation of Well Logs. 2nd edition, Rider-French Consulting Ltd., Sutherland, Scotland, 280 p.

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(Supl.1): 61–74.

SCHLIEP K & HECHENBICHLER K. 2009. kknn: Weighted k-Nearest Neighbors. R package version 1.0-7. Disponível em: <http://CRAN.R-project.org/package=kknn>. Acesso em: 22 jan. 2010.

SERRA O. 1986a. Fundamentals of Well-Log Interpretation – 2. The Interpretation of Logging Data, Developments in Petroleum Science (15B), Elsevier Science Publishers B.V., 684 p.

SERRA O. 1986b. Sedimentary environments from wireline logs. Schlumberger Ltd., New York, 211 p.

SOUZA Jr OG. 1997. Stratigraphie Séquentielle et Modélisation Probabiliste des Réservoirs d'un Cône Sous-marin Profond (Champ de Namorado, Brésil). Intégration dês Données Géologiques. Thèse de Doctorat. Université Paris 6, 128 p.

SWITZER P & GREEN A. 1984. Min/Max autocorrelation factors for multivariate spatial imagery. Technical Report No. 6, Department of Statistics, Stanford University, Stanford, CA, 23 p.

TANG H, WHITE C, ZENG X, GANI M & BHATTACHARYA J. 2004. Comparison of multivariate statistical algorithms for wireline log facies classification. AAPG Annual Meeting, Abstract, 88: 13.

VARGAS-GUZMÁN JA & DIMITRAKOPOULOS R. 2003. Computational properties of min/max autocorrelation factors. Comput. Geosci., 29: 715–723.

NOTAS SOBRE OS AUTORES

Rodrigo Duarte Drummond formou-se em Matemática Aplicada e Computacional pela Universidade Estadual de Campinas em 1995. Nessa mesma universidade, obteve o título de Doutor em Genética e Biologia Molecular, com ênfase em Bioinformática, em 2007. Atualmente trabalha como pesquisador no Centro de Estudos do Petróleo da Universidade Estadual de Campinas. Tem experiência em análise computacional de dados, com ênfase em dados de expressão gênica e dados geofísicos.

Alexandre Campane Vidal é formado em Geologia pela Universidade de São Paulo em 1993, obteve o título de Mestre em Geoengenharia de Reservatórios pela Universidade Estadual de Campinas, em 1997, e de Doutor em Geologia Regional pela UNESP, em 2003. Durante o período de 2002-2003 fez pós-doutorado no Departamento de Geologia Aplicada da Universidade Estadual Paulista. Atualmente é Professor Assistente Doutor do Departamento de Geologia e Recursos Naturais do Instituto de Geociências da Universidade Estadual de Campinas. Tem experiência na área de Geologia, com ênfase em Geologia de Reservatórios.