Predição de Perfis Geofísicos de Poços por Classificador KNN

Frederico Silva de Azevedo Ribeiro[1], José Agnelo Soares[2], Luiz Landau[1]
[1]COPPE/UFRJ, [2]Universidade Federal de Campina Grande (UFCG)

Abstract

There is a large diversity of well log curves but, in practice, its availability is always limited. Thus, sometimes a desired log curve is not available for a given application. However, well logs are in some level interdependent, so, it is possible to use some well logs to estimate another missing curve. In this work the multivariate statistical method KNN (K-nearest neighbors) is used to estimate a basic suite of well logs (GR, RHOB, NPHI, DT and ILD) from Namorado Field, offshore Brazil. KNN is adopted due to its simple implementation, low computer cost and high resolution. A training data set was composed by random choice of 30% of standardized well log curves from 12 vertical wells. The remaining data was used for well log prediction. All wells have the complete suite of logs, but in order to check the performance of prediction, each time one log curve was removed from data base, a synthetic one was estimated and compared to the original curve, furnishing an average prediction error. KNN was a suitable method for synthetic curve estimation of the majority of well logs, with estimated curves well correlated to real curves in terms of curve shape, well log values and resolution level. Nevertheless, the same prediction performance was not achieved for all log curves. For ILD curve the general prediction error was 186.9%, an unacceptable high value, meanwhile the prediction error was clearly satisfactory for GR (13.2%), NPHI (12.6%), DT (4.1%) and RHOB (1.5%).

Introdução

A perfilagem de poços no Brasil teve grande impulso nas décadas de 70 e 80 em função do aumento na atividade exploratória no país e de avanços tecnológicos, como por exemplo, o emprego de unidades de aquisição informatizadas e de novas ferramentas. A disponibilidade de perfis, no entanto, quase sempre é limitada, sendo essa restrição dada por razões de contingência financeira, disponibilidade de sondas ou dificuldades operacionais. Dessa forma, é relativamente corriqueira a situação em que, para uma dada aplicação, se deseja conhecer uma curva de perfil que não foi registrada. Felizmente, as propriedades petrofísicas representadas nos perfis de poços não são inteiramente independentes, mas guardam relações, nem sempre lineares, de dependência entre si. Graças a essa interdependência, é possível estimar, com variados níveis de precisão, um dado perfil faltante a partir dos demais perfis disponíveis em um poço.

Diversos métodos numéricos podem ser utilizados para a estimativa de perfis faltantes em poços. Basicamente existem três famílias principais de métodos: estatística multivariada, rede neural e lógica fuzzy. Este trabalho apresenta resultados obtidos pela aplicação de um método de estimativa baseado em estatística multivariada: o método KNN (*K-nearest neighbors*, ou o método dos K vizinhos mais próximos). Uma grandeza desconhecida pode ser estimada por análise estatística multivariada em um processo que envolve pelo menos duas fases: treinamento e predição. Na fase de treinamento a regra de classificação é estabelecida e na fase de predição essa regra de classificação é aplicada aos dados para os quais se deseja estimar a variável desconhecida. Quanto ao operador de classificação na fase de treinamento os métodos da estatística multivariada podem ser agrupados em métodos de classificação supervisionada e métodos de classificação não-supervisionada. O primeiro grupo diz respeito aos métodos em que, durante a fase de treinamento, a resposta esperada é conhecida. Neste grupo estão incluídas as regras discriminantes, a correlação canônica e o método KNN. No segundo grupo o treinamento é realizado apenas com os dados de entrada, sem o conhecimento da resposta desejada. Neste caso há apenas uma classificação baseada na separação de grupos de amostras com propriedades similares, como é feito no caso do método da análise de agrupamento.

Os métodos de estatística multivariada podem ainda ser classificados em métodos paramétricos ou métodos não-paramétricos. Os primeiros exigem que todas as variáveis de análise apresentem distribuições estatísticas parametrizadas, por exemplo, que todos os perfis apresentem distribuição gaussiana. Já os métodos não-paramétricos não fazem essa exigência. No primeiro grupo estão as regras discriminantes e no segundo grupo está o método KNN. A regra discriminante passo-a-passo seleciona, entre os vários perfis disponíveis, aqueles que mais contribuem para a predição de um dado perfil faltante, criando um ranking de perfis com contribuição decrescente, e exclui do ranking os demais perfis, segundo um dado critério de aceitação.

Neste trabalho foram utilizados dados de uma suíte básica de perfis do Campo Escola de Namorado, cedida pela ANP. Ela é composta pelas curvas Raios Gama (GR), Indução Profunda (ILD), Porosidade de Nêutrons (NPHI), Densidade (RHOB) e Tempo de Trânsito (DT). Foram utilizados dados de 12 poços, segundo o critério de adotar apenas poços verticais (ou com pequeno desvio) e que possuam a suíte completa de perfis. Maiores detalhes sobre os dados utilizados podem ser acessados em Ribeiro (2008).

O método KNN

O método KNN consiste em atribuir a uma dada amostra o rótulo (ou valor) que aparece mais vezes entre os rótulos (ou valores) dos seus K vizinhos mais próximos, segundo a métrica adotada, no espaço n-dimensional. No caso da predição de perfis faltantes a dimensão n é igual ao número de perfis disponíveis e utilizados na entrada do modelo de análise multivariada. A métrica é dada por uma medida de similaridade no espaço n-dimensional. No caso deste trabalho a métrica utilizada é a distância euclidiana que, essencialmente, é uma medida do comprimento de um segmento de reta, no espaço n-dimensional, desenhado entre duas amostras (Hair *et al*., 2005):

$$d(1,2) = [(GR_1 - GR_2)^2 + ... + (RHOB_1 - RHOB_2)^2]^{1/2} \quad (1)$$

A Figura 1 apresenta um exemplo ilustrativo do mecanismo de classificação KNN com duas dimensões (n = 2: GR e DT), três rótulos de saída (os valores possíveis de RHOB: 2.0, 2.5 e 3.0) e duas amostras onde o valor de RHOB é desconhecido: 1 e 2. Deseja-se classificar estas duas amostras através dos 7 vizinhos mais próximos (K=7). Analisando o rótulo predominante entre os 7 vizinhos mais próximos, à amostra desconhecida 1 será imputado o valor de RHOB = 3.0, pois na base de treinamento há quatro vizinhos mais próximos com esse valor, e à amostra desconhecida 2 será imputado o valor de RHOB = 2.5, pois há cinco amostras mais próximas com esse valor de RHOB no espaço bidimensional GR-DT. O mecanismo ilustrado na Figura 1 pode ser extrapolado para o espaço n-dimensional, onde a dimensão espacial reflete o número de variáveis, ou perfis, utilizados como dados de entrada para a predição do perfil faltante.

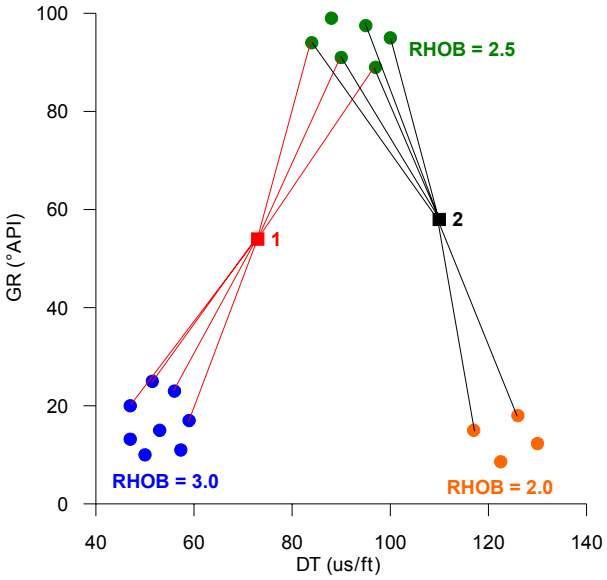

Figura 1 – Exemplo ilustrativo do mecanismo de classificação do método KNN.

KNN é um classificador que possui apenas um parâmetro livre (o número K) que é controlado pelo usuário com o objetivo de obter uma melhor classificação. Como pode ser visualizado da Figura 1, o resultado da classificação depende do valor adotado para o parâmetro K. Determinar o valor ideal para K não é uma tarefa trivial. No entanto alguns princípios gerais devem ser observados: a) ele deve ser preferencialmente um número ímpar, pois ao adotar um número par para K o procedimento de classificação KNN pode identificar um mesmo número de rótulos entre os K vizinhos mais próximos, o que resulta na não classificação da amostra; b) K deve ser suficientemente grande para garantir um nível aceitável de acertos na fase de predição; e, c) K deve ser suficientemente pequeno para garantir a resolução desejada.

Quando se dispõe de uma base de dados muito grande, como é, em geral, a base de dados de perfilagem de poços de um campo de petróleo em desenvolvimento, pode-se adotar um valor pequeno para K (como K = 1) sem que se tenha um erro de predição muito grande e, ao mesmo tempo, garantindo uma alta resolução na predição.

Regra discriminante linear passo-a-passo

Uma regra discriminante linear gera uma função linear entre a variável dependente e as variáveis independentes. Para os objetivos deste trabalho a variável dependente é o perfil que se deseja estimar (o perfil faltante) e as variáveis independentes são os perfis conhecidos em um dado poço. Por exemplo, para gerar uma curva sintética do perfil DT, deve-se definir uma função do tipo:

$$DT = c_0 + c_1 \cdot GR + c_2 \cdot RHOB + \ldots + c_n \cdot NPHI \quad (2)$$

Os coeficientes c_0 até c_n são determinados através da solução de um sistema com m equações lineares, onde m é dado pelo número de amostras dos perfis de poços.

A regra discriminante linear passo-a-passo (DLPP) é um método objetivo para selecionar variáveis que maximizam a previsão com o menor número de variáveis empregadas. Na DLPP as variáveis são individualmente avaliadas quanto a sua contribuição na previsão da variável dependente e acrescentadas ao modelo de regressão ou eliminadas do mesmo com base em sua contribuição relativa. Ela permite examinar a contribuição de cada variável independente para o modelo de regressão. Todas as variáveis são testadas no modelo de regressão. Cada variável é considerada para inclusão antes do desenvolvimento da equação. A variável independente com a maior contribuição é acrescentada em um primeiro momento. Variáveis independentes são selecionadas para inclusão ou exclusão, com base em sua contribuição incremental sobre as variáveis já presentes na equação. Ao final, a DLPP fornece um ranking com os perfis que mais contribuem para a estimativa de um dado perfil faltante.

Procedimento e dados utilizados

Embora o Campo de Namorado disponha de uma suíte básica de perfis bastante completa, os dados desse campo foram utilizados para modelar curvas supostamente faltantes nesse campo. Para a predição de cada curva de perfilagem, em cada um dos 12 poços analisados neste trabalho, se retirava a curva original da

suíte de um dado poço e se realizava a predição daquela curva como se ela não existisse, através de análise multivariada. Esse procedimento permitiu analisar o nível de acerto da predição pela posterior comparação entre a curva sintética gerada e a curva real registrada no poço. A técnica adotada para a modelagem dos perfis faltantes foi a KNN devido a sua simplicidade de implementação, seu baixo custo computacional, e sua alta resolução.

Para o trabalho de modelagem supervisionada necessita-se dispor de dois conjuntos de dados: um completo, para treinamento, no qual se dispõe de todas as curvas, e outro incompleto, com perfis faltantes, para a predição. A fim de gerar esses dois conjuntos de dados, procedeu-se a uma amostragem aleatória dos dados. De cada poço foi extraído aleatoriamente um conjunto de dados composto por 30% dos registros de perfis. Os 70% restantes de cada poço foram usados como dados para predição. A concatenação das parcelas com 30% dos dados de todos os poços deu origem ao arquivo com dados de treinamento. Os dados de predição foram processados de forma individual por poço.

A adoção de um conjunto de treinamento de tamanho equivalente a 30% dos dados disponíveis foi uma escolha arbitrária, mas baseada no seguinte princípio: o tamanho da base de treinamento deve ser suficientemente grande para representar todo o intervalo de variação nos valores dos perfis, mas, também suficientemente pequena para não viciar as estimativas, ou seja, se usássemos 100% dos dados na fase de treinamento, as respostas seriam ótimas, mas o teste seria inválido no sentido que se estaria tentando estimar valores já totalmente conhecidos pelo próprio procedimento de estimação.

A predição dos perfis sintéticos foi realizada de acordo com o fluxograma de trabalho apresentado na Figura 2. Os dados dos poços do campo de Namorado foram cedidos pela ANP em formato padrão para perfis geofísicos de poços. Para os objetivos deste trabalho, os dados foram reformatados, excluindo informações excedentes e desnecessárias, sendo salvos em formato final próprio para processamento no programa SAS®.

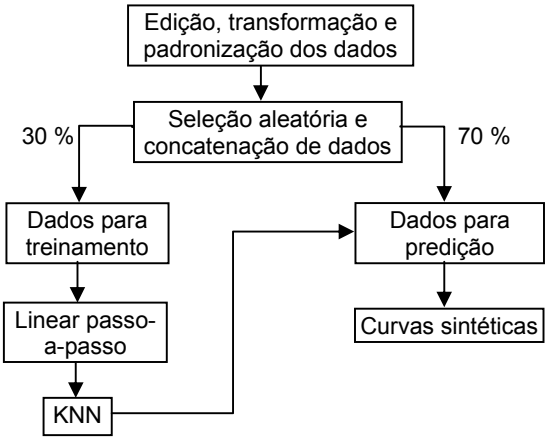

Figura 2 – Fluxograma de trabalho representando o procedimento de geração das curvas sintéticas.

A classificação KNN depende da distância medida, entre amostras, no espaço n-dimensional, definido pelas variáveis de análise (no caso, os perfis de poços). No entanto, os perfis apresentam intervalos de variação bastante distintos. Por exemplo, a curva RHOB em geral apresenta uma variação linear entre 2.0 g/cm^3 e 3.0 g/cm^3, enquanto que a curva ILD pode apresentar uma variação exponencial entre algo como 0.01 ohm.m e 2000 ohm.m. Essa diversidade de escalas dificulta a classificação baseada em uma medida de similaridade (distância). A fim de contornar esse problema, aplica-se um procedimento de transformação e padronização dos dados, conforme descrito em Soares (2005), do qual resultam curvas com mesmo intervalo de variação de valores.

De cada poço, são selecionadas 30% das amostras, escolhidas aleatoriamente em termos de profundidade. A concatenação dessas amostras extraídas de todos os poços gera a base de dados de treinamento. Os 70% restantes dos dados corresponde à base de dados para predição. Sobre a base de treinamento inicialmente aplica-se um procedimento discriminante linear passo-a-passo com o objetivo de identificar os perfis com maior poder de predição para uma determinada curva desejada. Esses perfis correspondem às curvas de entrada para o modelo de predição. Em seguida, ainda sobre a base de treinamento, aplica-se a regra KNN para a definição do modelo de predição. A definição do modelo de predição corresponde à delimitação do subespaço de cada classe (ou valor de perfil), dentro do espaço n-dimensional, conforme a medida de similaridade (distância) adotada.

O passo seguinte é classificar os dados de predição de acordo com o modelo definido pelo classificador KNN quando aplicado sobre a base de treinamento. Quando este procedimento é aplicado sobre todos os dados de cada poço (dados de treinamento + dados de predição) o resultado é o conjunto de todas as curvas sintéticas dos poços. A comparação dessas curvas sintéticas com os perfis reais registrados permite avaliar a precisão da estimativa.

Neste trabalho foi adotado um K =1, o que pode fazer com que o classificador KNN forneça alguns valores com erros locais substanciais. Com o objetivo de atenuar esses erros locais, aplicou-se um filtro de suavização com uma média móvel de três amostras sobre as curvas sintéticas geradas.

Resultados

Foram obtidas cinco curvas sintéticas (para os perfis GR, RHOB, NPHI, DT e ILD) para cada um dos 12 poços analisados, o que resulta em 60 perfis sintéticos.

Com o objetivo de quantificar a precisão na estimativa dos perfis sintéticos, foi calculado o erro percentual médio de cada curva sintética em relação à correspondente curva real:

$$\text{Erro (\%)} = \frac{\sum_{i=1}^{n} \frac{\left|V_{sint} - V_{real}\right|}{V_{real}} \cdot 100}{n} \qquad (3)$$

onde $\mathbf{V_{sint}}$ é o valor da curva sintética, $\mathbf{V_{real}}$ é o valor da correspondente curva real e **n** é o número de amostras do perfil.

De um modo geral, as curvas sintéticas de GR obtidas apresentam a mesma forma e com amplitudes similares às correspondentes curvas reais de GR. Em geral os perfis sintéticos de GR se apresentam mais "nervosos" que os reais. Isto sugere que a suavização através de um filtro de média móvel com janela de três amostras foi insuficiente para o caso dos perfis sintéticos de raios gama. As variáveis de entrada para a geração da curva sintética de GR foram as curvas RHOB, DT, NPHI e ILD. A Figura 3 apresenta a comparação entre a curva sintética e a curva real do perfil GR no poço NA-01A.

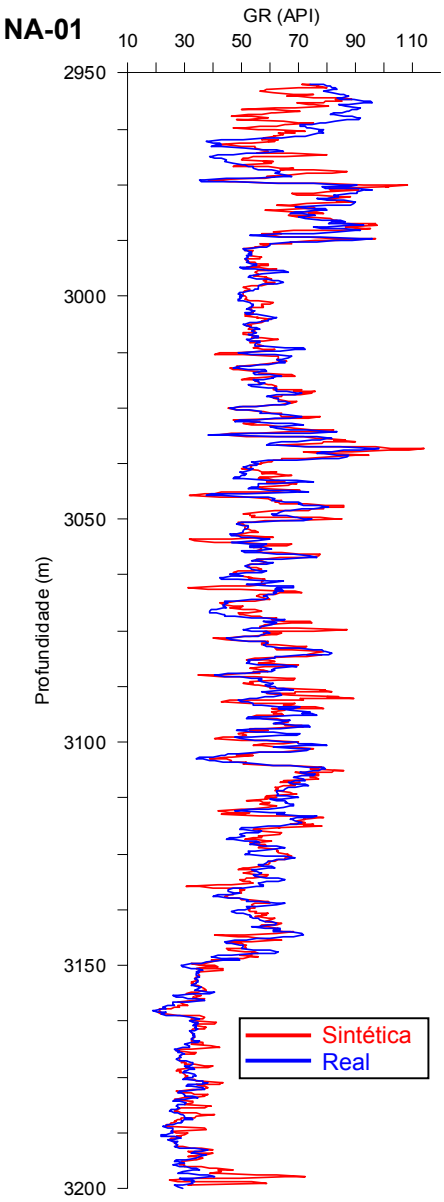

Figura 3 – Comparação entre a curva sintética (vermelha) e a curva real (azul) do perfil GR no poço NA-01A.

As curvas sintéticas para o perfil DT apresentaram, em geral, uma boa correlação, tanto em termos de forma quanto de amplitude, com as correspondentes curvas reais. O nível geral de ruído apresentado pelas curvas sintéticas de DT foi bem inferior ao observado nas curvas sintéticas de GR, sugerindo que o filtro de média móvel com três amostras foi adequado para o perfil DT. As variáveis de entrada para a geração da curva sintética de DT foram as curvas GR, RHOB, NPHI e ILD. A Figura 4 apresenta a comparação entre a curva sintética e a curva real do perfil DT no poço NA-05A.

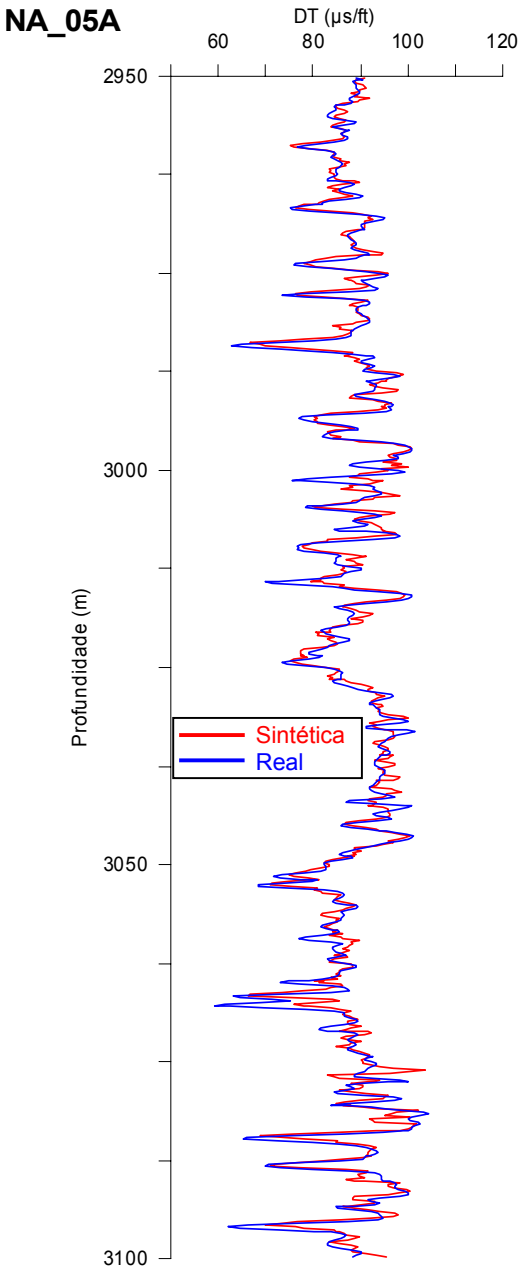

Figura 4 – Comparação entre a curva sintética (vermelha) e a curva real (azul) do perfil DT no poço NA-05A.

No caso dos perfis de porosidade de nêutrons (NPHI) também se observa em geral, uma correlação muito boa com as curvas reais, tanto em termos de forma quanto de amplitude. A presença de picos espúrios, em alguns poços, sugere a adoção de um filtro de média móvel com mais de três amostras. As variáveis de entrada para a geração da curva sintética de NPHI foram as curvas GR, RHOB e DT. A Figura 5 apresenta a comparação entre a curva sintética e a curva real do perfil NPHI no poço NA-05A.

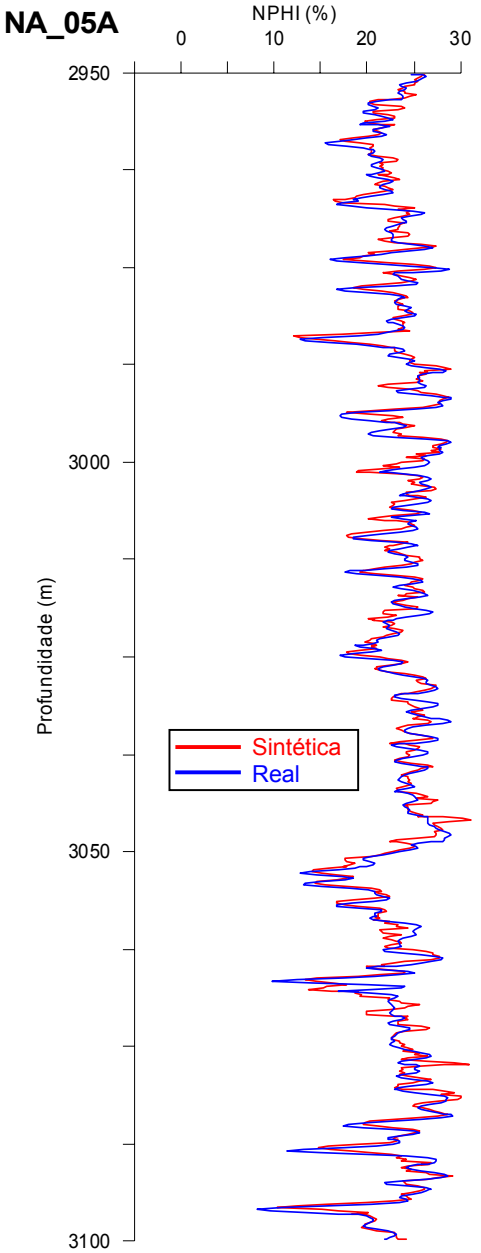

Figura 5 – Comparação entre a curva sintética (vermelha) e a curva real (azul) do perfil NPHI no poço NA-05A.

As curvas sintéticas de RHOB obtidas através do procedimento KNN mostram excelentes resultados. De um modo geral, são os melhores resultados alcançados. A reprodução das formas, das amplitudes e a apresentação de resolução compatível com os perfis reais ocorrem em todos os poços, o que indica que o filtro de média móvel com três amostras foi adequado para esse perfil. Para a simulação dos perfis de RHOB foram utilizados os perfis GR, DT, NPHI e ILD como variáveis independentes. A Figura 6 apresenta a comparação entre a curva sintética e a curva real do perfil RHOB no poço NA-05A.

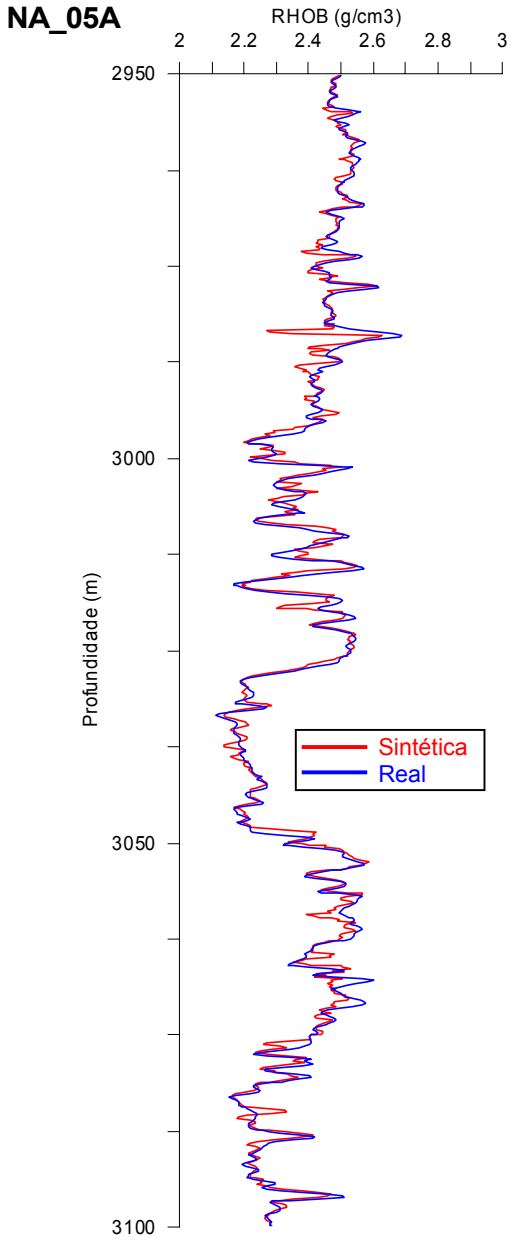

Figura 6 – Comparação entre a curva sintética (vermelha) e a curva real (azul) do perfil RHOB no poço NA-05A.

As curvas sintéticas de ILD apresentam, em geral, formas semelhantes e amplitudes no mesmo intervalo de variação de suas correspondentes curvas reais, no

entanto, apresentam muitos picos espúrios, indicando uma maior dificuldade do algoritmo KNN em realizar uma estimativa precisa da resistividade elétrica. Possivelmente a razão para isto é que a resistividade elétrica apresenta variação logarítmica, enquanto que as demais propriedades medidas pelos perfis variam em escala linear. A adoção de um filtro de suavização com uma janela mais ampla provavelmente seria mais indicado para o caso do perfil ILD. Para a simulação dos perfis de ILD foram utilizados os perfis GR, DT, NPHI e RHOB como variáveis independentes. A Figura 7 apresenta a comparação entre a curva sintética e a curva real do perfil ILD no poço NA-21B.

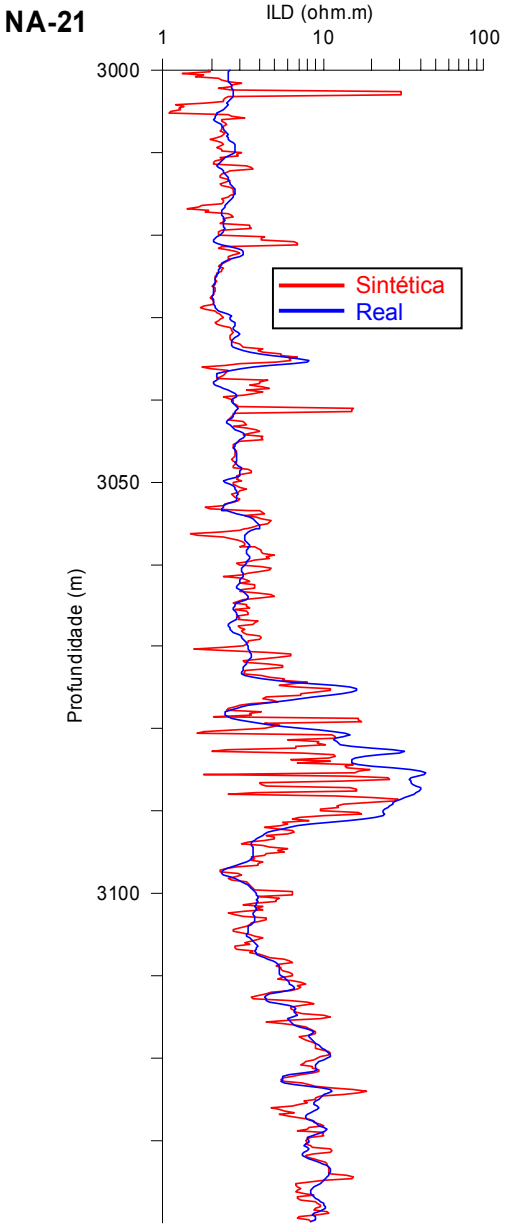

Figura 7 – Comparação entre a curva sintética (vermelha) e a curva real (azul) do perfil ILD no poço NA-21B.

A Tabela 1 apresenta o erro médio percentual obtido na predição dos perfis em todos os poços estudados. Observa-se que, com a exceção do perfil ILD, o classificador KNN estimou perfis sintéticos com erros médios aceitáveis em relação aos perfis reais dos poços.

Tabela 1 – Erros médios de predição.

Poço	Erro médio de predição (%)				
	GR	DT	NPHI	RHOB	ILD
NA-01A	12,0	4,6	19,5	1,6	460
NA-02	13,5	6,2	14,4	2,3	230
NA-04	14,0	3,6	9,5	1,1	29
NA-05	10,5	3,3	7,0	1,2	30
NA-07	12,0	4,5	16,5	1,8	290
NA-11A	17,6	3,0	13,2	1,1	28
NA-12	14,2	3,9	10,1	1,6	32
NA-17A	13,3	4,6	15,0	1,7	190
NA-21B	12,4	2,8	13,6	0,9	15
RJS-19	14,1	3,5	7,6	1,2	51
RJS-42	12,3	4,9	12,3	1,7	193
RJS-234	12,6	4,8	12,4	1,9	695
Geral	**13,2**	**4,1**	**12,6**	**1,5**	**186,9**

Conclusões

Para a grande maioria dos perfis a técnica KNN se mostrou adequada para a estimação de perfis sintéticos, em termos de correlação com a forma, amplitude de valores e nível de resolução, em relação às curvas reais. No entanto, o desempenho observado não foi o mesmo para todas as curvas. O erro médio geral de predição para o perfil ILD (186,9%) não pode ser considerado aceitável. No entanto, a predição das demais curvas foi satisfatória. Os melhores resultados obtidos foram para a curva RHOB (erro de 1,5%) e DT (erro de 4,1%), seguidos pelos resultados para as curvas NPHI (erro de 12,6%) e GR (erro de 13,2%). Devido à elevada disponibilidade de dados, é possível adotar K = 1. A adoção de um filtro de suavização de média móvel com janela de três amostras foi suficiente para a atenuação de picos espúrios na maioria dos perfis sintéticos, especialmente para os perfis de densidade e sônico.

Agradecimentos

O primeiro autor agradece a concessão de bolsa de estudos pelo PRH-02 da ANP.

Referências

Hair, F.H.J.; Anderson, R.E.; Tathanm, R.L.; Black, W.C. (2005) Análise Multivariada de Dados. 5ª ed., Porto Alegre, Bookman.

Ribeiro, F.S.A. (2008) Técnicas de classificação supervisionada na predição de perfis faltantes de poços. Dissertação de mestrado. Programa de Pós-graduação em Engenharia Civil. COPPE/UFRJ.

Soares, J.A. (2005) Um fluxo de trabalho para modelagem de eletrofácies com entrelaçamento de técnicas de classificação supervisionada e não-supervisionada. 9º Congresso Internacional da SBGf. Salvador.