



Detecção de Falhas em Dados Sísmicos 3D usando Funções Geoestatísticas e SVM

*Suellen de A. Caduda da S. Motta¹; Aristófanés C. Silva¹; Anselmo C. de Paiva¹; Marcelo Gattass²; Pedro Mário C. e Silva²

¹NCA / Núcleo de Computação Aplicada – Universidade Federal do Maranhão

²PUC-Rio – Pontifícia Universidade Católica do Rio de Janeiro

Copyright 2014, SBGf - Sociedade Brasileira de Geofísica

Este texto foi preparado para a apresentação no VI Simpósio Brasileiro de Geofísica, Porto Alegre, 14 a 16 de outubro de 2014. Seu conteúdo foi revisado pelo Comitê Técnico do VI SimBGf, mas não necessariamente representa a opinião da SBGf ou de seus associados. É proibida a reprodução total ou parcial deste material para propósitos comerciais sem prévia autorização da SBGf.

Abstract

In this work we introduce a methodology for seismic fault detection in three-dimensional seismic volume data using four geostatistical functions. These functions are used as attributes for a Support Vector Machine classifier. Visual and numerical results (up to 89% of accuracy) obtained on tests on Netherlands F3 block indicate this may be a good approach in the fault identification task, for oil fields exploration.

Introdução

A identificação de *falhas sísmicas* é uma etapa importante dentro do processo de exploração de reservatórios de hidrocarbonetos em subsuperfície. Normalmente, a detecção de falhas em dados sísmicos, tal como a de horizontes, é realizada manualmente por um intérprete – um geólogo ou geofísico –, o que demanda bastante tempo e experiência. Por essa razão, diversos trabalhos computacionais têm visado à detecção ou o realce automático das falhas sísmicas, utilizando técnicas de processamento de imagens, especialmente a detecção de bordas. Aqravi et. al. (2011), por exemplo, segmentam as falhas utilizando uma versão modificada do filtro de Sobel em 3D. Song et. al. (2012) apresentam um método de segmentação também baseado em detecção de bordas e aproximação por superfície (*surface fitting*). Com uma técnica parecida, Hale (2007) propôs filtros locais de mergulho para atenuar ou melhorar características baseadas em derivadas direcionais de imagens e em outro trabalho propôs um método para a computação 3D de falhas em imagens, extraindo as superfícies de falhas (HALE, 2009).

Apresenta-se neste trabalho uma metodologia para a detecção de falhas em dados sísmicos tridimensionais, utilizando-se a técnica Máquina de Vetores de Suporte (*Support Vector Machine* – SVM) com atributos baseados em textura obtida a partir de funções geoestatísticas. Estas funções permitem analisar a variabilidade das amostras em conjunto com a sua correlação espacial.

Funções Geoestatísticas

A geoestatística estuda o comportamento das chamadas variáveis regionalizadas. Ou seja, variáveis a que está associada uma localização. As funções geoestatísticas apresentadas a seguir analisam em conjunto a variabilidade e a correlação espacial entre as amostras (SILVA A. C., 2004), definidas por uma função aleatória $Z(u)$ que, no caso de imagens, equivale à intensidade de cada pixel, e no volume sísmico estudado, à amplitude de cada amostra.

As quatro medidas aqui apresentadas são funções de um vetor h , o qual define um ângulo e uma distância para a análise de pares de amostras.

Uma vez que x_i representa as amostras de origem; $x_i + h$ as amostras de extremidade do vetor; e que N equivale ao número total de pares considerados, a medida de *semivariograma* é dada pela Equação 1. Ela mede o grau de dependência entre as amostras, a partir da diferença quadrada de suas intensidades:

$$\gamma(h) = \frac{1}{2N} \sum_{i=1}^N [Z(x_i) - Z(x_i + h)]^2 \quad (1)$$

A função *semimadograma* é semelhante à função semivariograma, exceto que a dependência entre os pixels é medida pela média da diferença absoluta de suas intensidades:

$$m(h) = \frac{1}{2N} \sum_{i=1}^N |Z(x_i) - Z(x_i + h)| \quad (2)$$

A função covariograma, por sua vez, é uma medida de correlação entre duas variáveis. Seu valor tende a ser alto quanto maior for a distância entre os pares analisados e o seu cálculo é dado pela Equação 3:

$$C(h) = \frac{1}{N} \sum_{i=1}^N [Z(x_i)Z(x_i + h) - m_{-h}m_{+h}] \quad (3)$$

Sendo que m_{-h} e m_{+h} equivalem respectivamente à média dos valores das amostras de origem e à média dos valores das amostras das extremidades.

$$m_{-h} = \frac{1}{N} \sum_{i=1}^{N(h)} Z(x_i) \quad (4)$$

$$m_{+h} = \frac{1}{N} \sum_{i=1}^N Z(x_i + h) \quad (5)$$

Por último, a função correlograma (Equação 6) é obtida dividindo-se o valor de covariograma pelo produto dos desvios-padrões das amostras de origem e de extremidade. É, portanto, uma versão normalizada do covariograma. Espera-se que a correlação seja alta para unidades próximas (correlograma se aproxima de 1) e baixa quando a distância aumenta (correlograma se aproxima de -1) (SILVA A. C., 2004)

$$\rho(h) = \frac{C(h)}{\sigma_{-h}\sigma_{+h}} \quad (6)$$

Os desvios-padrões são calculados da seguinte forma:

$$\sigma_{-h} = \sqrt{\frac{1}{N} \sum_{i=1}^N [Z(x_i)^2 - m_{-h}^2]} \quad (7)$$

$$\sigma_{+h} = \sqrt{\frac{1}{N} \sum_{i=1}^N [Z(x_i + h)^2 - m_{+h}^2]} \quad (8)$$

Aprendizado de Máquina e SVM

Padrões ou objetos em imagens podem ser descritos através de um conjunto de propriedades ou características extraídas. Técnicas de reconhecimento de padrão são utilizadas para classificar conjuntos de dados como pertencentes a uma ou a outra classe. No caso do presente trabalho, classificam-se amostras extraídas das imagens sísmicas como amostras pertencentes a um dos dois grupos: falhas ou “não-falhas”.

Na fase chamada de treinamento, uma técnica de aprendizado, como o SVM, aprenderá os padrões correspondentes às classes desejadas a partir de características que as representam, e deverá então ser capaz de reconhecer novos elementos (falhas e não-falhas) ainda não “vistos”, na fase de teste.

A técnica SVM é um método de aprendizagem supervisionada. A ideia por trás do SVM é construir um hiperplano como superfície de decisão, de forma que a margem de separação entre as classes que se deseja classificar seja máxima. Um classificador SVM é capaz de encontrar um hiperplano baseado em um conjunto de pontos denominados “vetores de suporte”, o qual maximiza a margem de separação entre as classes. A literatura sobre a técnica é vasta, e melhores descrição e

funcionamento podem ser encontrados por exemplo em (BURGES, 1998) e (FAN et. al., 2008).

Para avaliar a capacidade de aprendizado e generalização da técnica, os testes serão avaliados segundo as taxas de sensibilidade (SB), especificidade (EP) e acurácia (AC), dadas por: (PARIKH et. al, 2008)

$$SB = \frac{VP}{VP + FN} \quad (9)$$

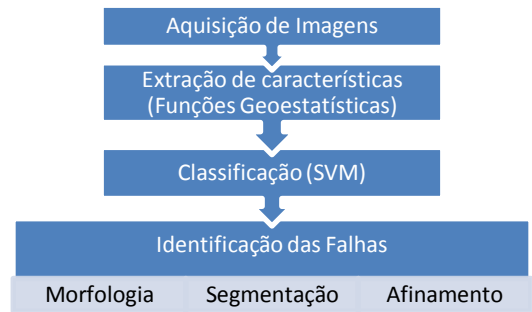
$$EP = \frac{VN}{VN + FP} \quad (10)$$

$$AC = \frac{VP + VN}{VP + VN + FP + FN} \quad (11)$$

onde VP (Verdadeiros Positivos) corresponde à quantidade de amostras de falha classificadas corretamente como falhas; VN (Verdadeiros Negativos) são as amostras de não-falha classificadas corretamente como não-falhas; FP (Falsos positivos) são as amostras de não-falha classificadas erroneamente como falhas e FN (Falsos Negativos) equivale às falhas classificadas erroneamente como não-falhas.

Metodologia

O fluxograma a seguir apresenta a sucessão das etapas da metodologia proposta.



Aquisição de Imagens

A metodologia foi aplicada no dado sísmico tridimensional Netherlands offshore F3 block (disponível gratuitamente para download no repositório Open Seismic no site do sistema Opentect <http://opentect.org/index.php/share-seismic-data/osr.html>). Em todo o trabalho, este será referenciado simplesmente como F3 block.

Extração de Características

Na etapa de extração de características, a cada amostra do volume sísmico são atribuídos valores (características, atributos) que serão capazes de representá-lo como uma amostra pertencente a uma região de falha ou de “não-falha”. Esses valores são os resultados das funções semivariograma, semimadograma, covariograma e correlograma, calculadas em janelas tridimensionais de

tamanho $L \times A \times P$ no volume, onde L = largura, A = altura e P = profundidade, parâmetros para o cálculo dos atributos.

O vetor h , também parâmetro das funções, pode assumir treze direções possíveis dentro de cada janela, uma vez que se trata de vetores que ligam pares de voxels (elementos discretos) em um espaço tridimensional. Para ilustrar as treze direções, considere um cubo composto por um voxel central e seus 26 vizinhos, tal como na Figura 1. As direções possíveis a partir da amostra em amarelo estão destacadas em vermelho. Observe que para cada amostra vermelha existe uma oposta, semitransparente. Ambas, no entanto, estão na mesma direção do voxel central.

Além da direção, o vetor h possui uma distância definida. As distâncias possíveis entre quaisquer pares de voxels dependem, é claro, das dimensões da janela, e estão, portanto, definidas como pertencentes ao conjunto $D_h = \{1, 2, \dots, \min(L, A, P) - 1\}$.

Por exemplo, considerando uma janela de dimensões $3 \times 3 \times 3$, o vetor h pode assumir tamanho $d \in D_h = \{1, 2\}$. Assim, na Figura 1, segundo a direção mostrada no primeiro exemplo, há dezoito pares de voxels separados pelo vetor h , caso sua distância seja $d = 1$, e nove pares, caso a distância seja $d = 2$.

O processo completo de extração de características de todo o volume consiste em centralizar a janela sobre todos os voxels (excluindo os da borda, uma vez que toda a janela deve conter voxels da imagem); calcular, em cada posicionamento, os valores de semivariograma, semimadograma, covariograma e correlograma para todas as variações de distância e direção do vetor h desejáveis; e, por fim, atribuir esses valores ao voxel (amostra) central, como seu vetor de características.

Assim, considerando-se o mesmo exemplo de janela ($3 \times 3 \times 3$), com todas as treze direções e distâncias 1 e 2, cada amostra do volume seria representada por um total de 104 características.

Classificação

As amostras, definidas pelos seus atributos geoestatísticos são então submetidas à classificação pelo método SVM. Isto quer dizer que cada voxel do volume sísmico, o qual representa sua região de vizinhança, será classificado como falha ou “não-falha”, com base no seu vetor de características, e a partir de um modelo gerado previamente.

O modelo de classificação é criado a partir do treinamento de subvolumes do dado F3-block, os quais tiveram as suas regiões de falhas demarcadas manualmente por um especialista.

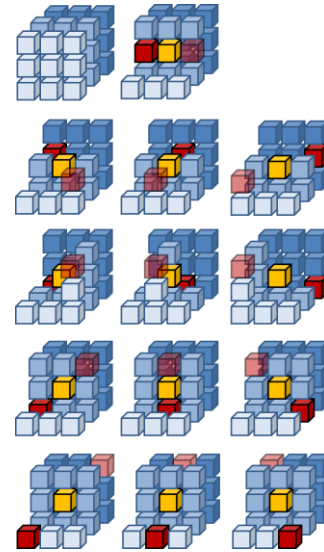


Figura 1 - Possíveis direções para o vetor h em um espaço tridimensional.

O F3-block possui dimensões $950 \times 461 \times 580$, totalizando mais de 254 milhões de voxels. Os subvolumes utilizados para treinamento possuem dimensões:

- V1: $110 \times 160 \times 70$ voxels
- V2: $160 \times 80 \times 20$ voxels

Foram selecionadas para os subvolumes partes das regiões que contém falha, em diferentes direções, dentro do dado F3-block. A Figura 2 apresenta um exemplo de seção de cada volume. A seção do volume V1 contém uma única falha central. A seção do volume V2 contém diversas falhas consecutivas, em forma de “degraus”.

Para validação da classificação, parte das amostras calculadas é utilizada para o treinamento com o SVM e a outra é reservada para teste. O treinamento gera um modelo de classificação, que é então aplicado à base de teste para a avaliação do seu desempenho. Utilizou-se a técnica SVM sem a utilização de kernel através da implementação da biblioteca *liblinear* (FAN et. al., 2008).

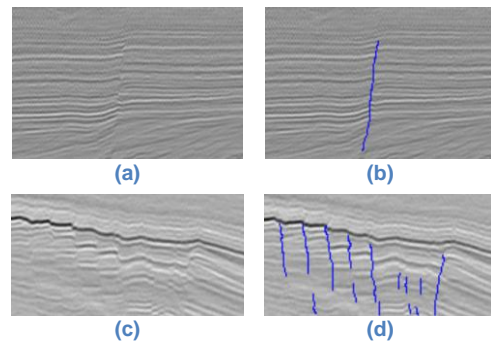


Figura 2 – Seções dos subvolumes. (a) e (b) Seção do volume V1, original e demarcada manualmente. (c) e (d) Seção do volume V2, original e demarcada manualmente.

A classificação linear é preferível quando há um grande número de características e de amostras a serem classificadas (FAN et. al., 2008).

Sobre os parâmetros de treinamento, o valor de custo C permanece com o valor padrão ($C = 1$). É necessário o uso de pesos diferentes para cada classe a ser classificada, devido ao desbalanceamento entre as classes de falha e “não-falha”. No total de amostras obtidas dos subvolumes, a proporção de amostras de falha para “não-falha” é de 60 para 1, valores estes que configuram os pesos das classes.

Identificação das Falhas

Os objetivos desta etapa são a remoção dos falsos positivos resultantes da classificação e a extração das superfícies de falha. Para tal são utilizadas primeiramente duas técnicas amplamente aplicadas em processamento de imagens: morfologia matemática (aberturas e fechamentos); e segmentação por crescimento de região. Ambas didaticamente abordadas em (Gonzalez & Woods, 2010).

Na etapa de melhoramento utilizando a morfologia matemática, o volume binário é submetido a aberturas e fechamentos com elemento estruturante de dimensão 1×5 (largura x altura), de forma a: a) remover pequenas regiões classificadas erroneamente como falhas (ruídos); b) separar regiões de falhas muito próximas que estejam unidas; c) preencher os “buracos” existentes nas regiões.

As falhas são então isoladas através do método de crescimento de região, que consiste no agrupamento de todos os pixels vizinhos que estejam tridimensionalmente conectados. Nesta etapa é possível excluir objetos de tamanho muito pequeno, que são normalmente ruídos restantes da classificação.

O afinamento então é realizado de forma bidimensional, em cada falha individualmente. O processo, ilustrado pela Figura 3, consiste em selecionar, em cada linha da falha, aquela amostra cuja diferença em relação às amostras vizinhas é maior. Na figura, os pixels em branco são aqueles que compõem a falha extraída com o crescimento por região. Os pixels em vermelho são aqueles selecionados com base na sua vizinhança, destacada pelas setas em uma das linhas.

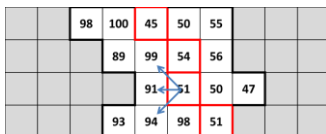


Figura 3 - Ilustração do afinamento.

Resultados

Esta seção apresenta os resultados parciais produzidos por cada etapa da metodologia, aplicada a subvolumes do dado F3 block.

Para validação do modelo de classificação, as amostras dos dois subvolumes do F3-block foram divididas em amostras de treino e teste, em proporções 20%-80%; 40%-60%; 60%-40% e 80%-20%. Isso permite avaliar a capacidade de generalização do modelo de treinamento: o quanto ele será capaz de acertar, com menos ou mais dados conhecidos. Para compor cada base de treinamento, foram selecionados blocos aleatórios de cada subvolume, tal como na Figura 4.

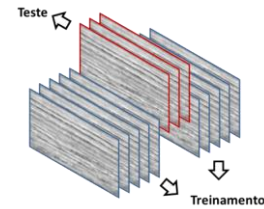


Figura 4 - Divisão das seções do volume sísmico para treinamento e teste.

O tamanho de janela escolhido, com base em resultados obtidos previamente, foi de $7 \times 7 \times 7$. Variou-se a distância do vetor h de 1 até 6 de forma cumulativa. Ou seja, o último teste contém vetores de todas as distâncias.

Como os blocos de treinamento são selecionados aleatoriamente, treino e teste foram realizados cinco vezes para evitar o selecionamento de uma base ruim para treino. As Tabelas 1 a 4 apresentam as médias e os desvios-padrão das taxas de sensibilidade, especificidade e acurácia para essas cinco execuções de cada combinação de distâncias. Para cada porcentagem de teste, os resultados são mostrados, com o melhor resultado destacado em azul e o pior em vermelho. Foram considerados melhores resultados aqueles que proporcionaram de forma balanceada altos valores para as três medidas de avaliação.

Observe que em todas as tabelas, os melhores resultados foram obtidos agrupando-se distâncias maiores. Os piores resultados foram sempre aqueles obtidos com uma única distância, igual a 1. Isto demonstra a importância de considerar como vizinhos não somente os voxels adjacentes, mas sim também aqueles distantes dentro do limite da janela estabelecida.

A taxa de sensibilidade aumenta com a proporção da base de treinamento. No entanto, os melhores valores em todas as tabelas são próximos, evidenciando uma boa capacidade de generalização do classificador para as medidas geoestatísticas.

A Figura 5 apresenta visualmente um dos melhores resultados, obtido com porcentagem de treino de 60% dos dados, janela $7 \times 7 \times 7$, e distâncias 1, 2, 3 e 4.

As imagens são seções dos dois subvolumes, com suas versões originais mostradas em a).

Distância	Porcentagem de teste: 80%					
	Sensibilidade		Especificidade		Acurácia	
	MD	DP	MD	DP	MD	DP
1	74,97%	0,079	86,70%	0,019	86,51%	0,017
1-2	79,92%	0,060	88,40%	0,014	88,27%	0,013
1-3	79,16%	0,038	88,79%	0,007	88,64%	0,006
1-4	76,96%	0,043	89,34%	0,010	89,15%	0,009
1-5	73,88%	0,059	90,03%	0,012	89,78%	0,011
1-6	75,66%	0,045	89,61%	0,009	89,39%	0,009

Tabela 1 - Resultados com 80% da base para teste.

Distância	Porcentagem de teste: 60%					
	Sensibilidade		Especificidade		Acurácia	
	MD	DP	MD	DP	MD	DP
1	79,47%	0,053	85,20%	0,021	85,10%	0,020
1-2	79,75%	0,039	87,83%	0,010	87,70%	0,009
1-3	79,84%	0,041	88,10%	0,010	87,96%	0,009
1-4	81,13%	0,016	87,90%	0,006	87,79%	0,005
1-5	79,74%	0,019	88,00%	0,004	87,87%	0,003
1-6	79,82%	0,014	88,34%	0,004	88,20%	0,004

Tabela 2 - Resultados com 40% da base para teste.

Distância	Porcentagem de teste: 40%					
	Sensibilidade		Especificidade		Acurácia	
	MD	DP	MD	DP	MD	DP
1	79,31%	0,030	84,88%	0,010	84,79%	0,010
1-2	83,34%	0,027	86,31%	0,006	86,26%	0,006
1-3	83,08%	0,010	86,50%	0,004	86,45%	0,004
1-4	83,05%	0,006	86,66%	0,004	86,60%	0,004
1-5	82,23%	0,014	86,91%	0,003	86,83%	0,003
1-6	82,41%	0,004	86,71%	0,001	86,64%	0,001

Tabela 3 - Resultados com 60% da base para teste.

Distância	Porcentagem de teste: 20%					
	Sensibilidade		Especificidade		Acurácia	
	MD	DP	MD	DP	MD	DP
1	78,54%	0,014	83,54%	0,007	83,45%	0,007
1-2	83,23%	0,008	84,70%	0,002	84,67%	0,002
1-3	83,10%	0,003	85,00%	0,001	84,97%	0,001
1-4	81,53%	0,003	85,11%	0,000	85,04%	0,001
1-5	80,46%	0,007	85,37%	0,002	85,27%	0,001
1-6	80,98%	0,008	85,23%	0,001	85,15%	0,001

Tabela 4 - Resultados com 80% da base para teste.

Em b) e em c) são mostrados os resultados da classificação com SVM, sendo que em b) estão destacados os pixels correspondentes aos verdadeiros positivos (em azul), aos falsos negativos (em vermelho), e aos falsos positivos (em verde). O restante das imagens, nos tons originais, corresponde aos verdadeiros negativos. Em c) são mostrados os pixels classificados como falha (brancos).

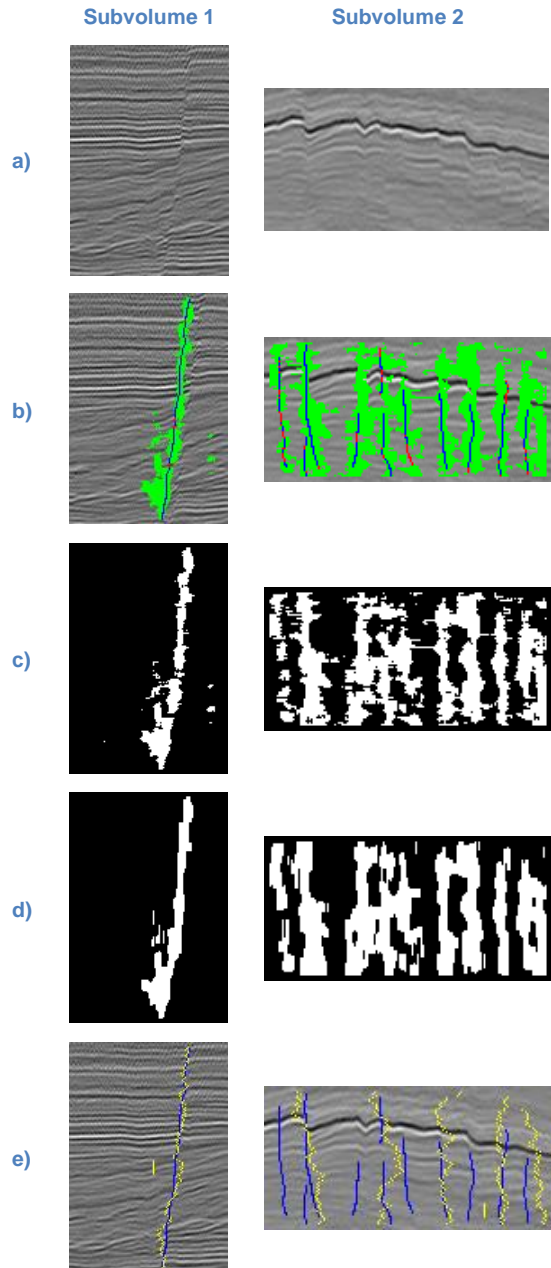


Figura 5 - Resultados. (a) Seções originais dos subvolumes (b) Resultado da classificação com SVM, ressaltando a quantidade de VP (azul), VN (cores originais), FP (verde) e FN (vermelho) (c) Resultado binário da classificação com SVM (branco: falha; preto: não-falha) (d) Seções após aplicação da morfologia matemática (e) Imagem final após afinamento das falhas (Resultado em amarelo, marcação em azul).

As imagens em d) são resultantes da aplicação da morfologia matemática: um fechamento seguido de uma abertura com elemento estruturante retangular 1x5. Observe como os pequenos elementos falsamente classificados como falha são removidos neste passo. E também os buracos contidos nas falhas são fechados.

Por fim, o resultado final é mostrado nas imagens em e). Estas são resultantes da etapa de afinamento (resultado em amarelo), executado em cada falha isolada (segmentada com crescimento de região). A marcação do especialista é mostrada em azul. A seção do subvolume 1 apresentada exemplifica um caso muito bem sucedido. O subvolume 2 já possui várias pequenas falhas muito próximas, dentre as quais nem todas foram identificadas corretamente. As duas primeiras, por exemplo, acabam sendo identificadas como uma única falha, e, portanto, a primeira não é destacada na última imagem.

Quanto melhor o resultado na fase de classificação (especialmente sensibilidade), melhores são os resultados nas etapas seguintes. E a fase de aplicação da morfologia visa suprir os problemas provenientes da etapa de classificação, em alguns casos, separando falhas diferentes que estejam unidas, sempre completando buracos dentro de falhas e eliminando pequenas regiões solitárias.

Conclusão e Trabalhos Futuros

Com base nos resultados até então obtidos em subvolumes extraídos do dado F3-block, pode-se concluir que as funções geoestatísticas provaram ser bons atributos para a detecção de falhas em dados sísmicos. A fase de classificação é a fase mais importante da metodologia, e principal foco de melhorias para os trabalhos futuros.

Uma primeira mudança seria o modo na seleção da base de treinamento. Neste trabalho, isto foi feito repetidamente, de forma aleatória. Porém, a otimização nesse processo utilizando técnicas de seleção de melhores indivíduos pode vir a melhorar os resultados.

Por último, embora um único dado sísmico de grandes dimensões, como o F3 block, seja capaz de fornecer uma enorme base de dados, a metodologia será aplicada em outros dados, visando avaliar a sua eficiência na detecção de falhas.

Referências

AQRAWI, A. A.; e BOE, T. H., 2011, Improved fault segmentation using a dip guided and modified 3D Sobel filter: 81st Annual International Meeting, SEG, Expanded Abstracts, p. 999–1003, 2011.

BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2):1–43, 1998.

CHOPRA, Satinder; e MARFURT, J. Kurt. Seismic Attributes for Prospect Identification and Reservoir Characterization. *SEG Geophysical Developments* no. 11, 2007.

FAN, Rong-En; CHANG, Kai-Wei; HSIEH, CHO-Jui; WANG, Xiang-Rui; e LIN, Chih-Jen., LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research* 9 (2008), 1871-1874. <http://www.csie.ntu.edu.tw/~cjlin/liblinear>

GONZALEZ, R. C., & WOODS, R. C. (2010). *Processamento Digital de Imagens* (3ª Edição ed.). São Paulo: Pearson Prentice Hall HALE, Dave. Local dip filtering with directional laplacians. Technical report, Center for Wave Phenomena, Colorado School of Mines, 2007.

HALE, Dave. Structure-oriented smoothing and semblance. Technical report, Center for Wave Phenomena, Colorado School of Mines, 2009.

MACHADO, Marcos de Carvalho. Determinação de Malhas de Falhas em Dados Sísmicos por Aprendizado Competitivo. Tese de Doutorado. PUC-Rio. Rio de Janeiro, 2008.

PARIKH, Rajul, MATHAI, Annie, PARIKH, Shefali, SEKHAR G Chandra, THOMAS, Ravi. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*. 2008 Jan-Feb; 56(1): 45–50.

SILVA, Aristófanos Corrêa; CARVALHO, Paulo Cezar Pinto; GATTASS, Marcelo. Diagnosis of lung nodule using semivariogram and geometric measures in computerized tomography images. *Computer Methods and Programs in Biomedicine*, v. 79, p. 31-38, 2005.

SILVA, Pedro Mário Cruz e. Visualização Volumétrica de Horizontes em Dados Sísmicos 3D. Tese de doutorado - PUC-Rio. Rio de Janeiro, 2004.

SONG, Jianguo; MU, Xing; LI, Zhe; WANG, Changjiang; e SUN, Yongzhuang. A faults identification method using dip guided facet model edge detector. *SEG Technical Program Expanded Abstracts 2012*, p. 1-5, 2012.

TUCERYAN, M. e JAIN, A. K. *The Handbook of Pattern Recognition and Computer Vision* (2nd Edition) by C. H. Chen, L. F. Pau, P. S. P. Wang. Cap. 2.1 – Texture Analysis, pp. 207-248, World Scientific Publishing Co., 1998.