UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

SUELLEN DE ARAUJO CADUDA DA SILVA MOTTA

DETECÇÃO DE FALHAS EM DADOS SÍSMICOS 3D UTILIZANDO FUNÇÕES GEOESTATÍSTICAS E SVM

São Luís
2015

SUELLEN DE ARAUJO CADUDA DA SILVA MOTTA

DETECÇÃO DE FALHAS EM DADOS SÍSMICOS 3D UTILIZANDO FUNÇÕES GEOESTATÍSTICAS E SVM

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Eletricidade na área de concentração Ciência da Computação.

Orientador: Prof. Dr. Aristófanes Corrêa Silva
Coorientador: Prof. Dr. Anselmo Cardoso de Paiva

São Luís

2015

Motta, Suellen de Araujo Caduda da Silva.

Detecção de falhas em dados sísmicos 3d utilizando funções geoestatísticas e svm/ Suellen de Araujo Caduda da Silva Motta. – São Luís, 2015.

69 f.

Impresso por computador (fotocópia).

Orientador: Aristófanes Corrêa Silva.

Coorientador: Anselmo Cardoso de Paiva.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2015.

1. Falhas sísmicas. 2. Funções geoestatísticas. 3. Volume sísmico. 4. Máquina de Vetores de suporte. I. Título.

CDU 550.34.06

DETECÇÃO DE FALHAS EM DADOS SÍSMICOS 3D UTILIZANDO FUNÇÕES GEOESTATÍSTICAS E SVM

Suellen de Araújo Caduda da Silva Motta

Dissertação aprovada em 02 de fevereiro de 2015.

Prof. Aristófanes Corrêa Silva, Dr.
(Orientador)

Prof. Anselmo Cardoso de Paiva, Dr.
(Co-orientador)

Profa. Aura Conci, Dra.
(Membro da Banca Examinadora)

Prof. João Viana Fonseca neto, Dr.
(Membro da Banca Examinadora)

AGRADECIMENTOS

À minha família. Em especial, aos meus pais Fernanda Caduda e Ivanildo Motta e à minha irmã Nathalia Caduda, pelo apoio durante esse projeto e sempre.

Ao meu namorado, Igor Costa, pela compreensão, paciência e ajuda, principalmente nos momentos finais deste trabalho.

Ao meu professor e orientador, Aristófanes Silva, por acreditar na minha capacidade e pelos ensinamentos, acadêmicos ou não, passados durante quase seis anos de convivência. Muito obrigada!

Ao Programa de Pós-Graduação em Engenharia de Eletricidade e a todos os seus professores, pelo aprendizado.

A todos os amigos do LabPAI e do NCA, sempre muito dispostos a ajudar a qualquer momento. E a todos os amigos da UFMA, que levarei por toda a vida.

Aos amigos da PUC-Rio, em especial ao professor Marcelo Gattass e ao grupo do v3o2, por me acolherem tão bem e me apresentarem ao mundo da sísmica.

Ao amigo Jéferson Coêlho, pela amizade, disponibilidade e alegria em explicar algoritmos, sísmica, geometria, “mineirês”, ou qualquer assunto, por qualquer meio, de qualquer localidade. Muito obrigada mesmo!

Ao CNPq, pelo suporte financeiro.

Muito obrigada a todos.

RESUMO

Este trabalho apresenta um método automático de detecção de falhas em volumes obtidos através do método de reflexão sísmica. Identificar as falhas geológicas nos dados sísmicos é importante para o conhecimento de um sistema geológico e para o planejamento da exploração de hidrocarbonetos. Sabendo-se que as falhas são descontinuidades presentes nos horizontes sísmicos, propõe-se a utilização de funções geoestatísticas capazes de indicar a variação da amplitude das amostras, em direções e distâncias predeterminadas. Assim, o método baseia-se no uso das funções semivariograma, semimadograma, covariograma e correlograma como características representativas das amostras, que serão classificadas como regiões de falha ou "não falha", através da técnica clássica de Reconhecimento de Padrões conhecida como SVM (*Support Vector Machine* – Máquina de Vetores de Suporte). O método proposto foi validado através de testes realizados com o volume F3 Block, disponibilizado pelo sistema OpendTect, apresentando até 92,15% de sensibilidade e 84,33% de especificidade. Este trabalho também apresenta um método de extração das linhas de falha baseado em crescimento de região e operadores morfológicos, a partir do volume binário resultante da classificação. Também testado sobre o F3 Block, o método foi capaz de extrair satisfatoriamente as falhas, na maioria das fatias do dado.

Palavras-chave: Falhas sísmicas, volume sísmico, funções geoestatísticas, reconhecimento de padrões, máquina de vetores de suporte.

ABSTRACT

This work presents an automatic method for fault detection in data obtained through seismic reflection method. Identifying geological faults in seismic data is critical for better understating a geological system and planning hydrocarbon exploration. Knowing that faults are discontinuities present in seismic horizons, we propose the use of geostatistical functions which are capable of indicating the amplitude variation along the volume samples, in both predetermined distances and directions. Thus, the method is based on semivariogram, semimadogram, covariogram and correlogram functions, used as representative characteristics for the samples, which will be classified as fault or "non fault" regions by the Pattern Recognition technique named Support Vector Machine (SVM). The proposed method was validated by tests made in F3 Block, a seismic data provided by OpendTect system, with up to 92.15% sensitivity and 84.33% specificity. This work also provides an extraction of fault lines method based on region growing segmentation and morphological operators applied on the classification binary resulted volume. Also tested in F3 Block, the method was able to satisfactorily extract the faults in most of the data slices.

Keywords: seismic faults, seismic data, geostatistical functions, pattern recognition, support vector machine.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O método de reflexão sísmica é uma técnica de exploração que permite o estudo de áreas subterrâneas de grande extensão e profundidade. A partir da reflexão de ondas sísmicas ao longo das camadas de subsuperfície captadas por sensores, constroem-se dados 2D ou 3D que representam regiões de interesse, objetivando-se principalmente a exploração de hidrocarbonetos (óleo, gás natural) e outros recursos minerais.

A chamada interpretação sísmica consiste na criação de modelos representativos das regiões exploradas, a partir dos dados sísmicos obtidos pelo método de reflexão. Trata-se de uma tarefa realizada por geólogos e geofísicos especializados, que mapeiam, entre outras estruturas, os horizontes – interfaces contínuas entre as camadas de rochas – e as falhas sísmicas – descontinuidades dos horizontes. As falhas são deslocamentos de rochas causadas principalmente por forças tectônicas. A importância da sua identificação está, por exemplo, no seu papel na formação de *trapas* (“armadilhas” onde óleo pode estar aprisionado), facilitando a compreensão do fluxo de fluidos em um reservatório.

A identificação das falhas é tradicionalmente realizada de forma manual pelos intérpretes, fatia por fatia, falha por falha, o que demanda bastante tempo e exige experiência. Além disso, não há garantia de que o resultado seja repetido, mesmo para uma interpretação realizada pelo mesmo especialista. Por esses motivos, a automatização da detecção das falhas é uma necessidade.

Ao longo dos anos, diversos trabalhos vêm propondo atributos de realce de falhas, no intuito de facilitar a interpretação dessas estruturas. Atributos, segundo Chopra e Marfurt (2007), são quaisquer medidas de dados sísmicos que auxiliam a realçar visualmente ou quantificar características de interesse na interpretação. Assim, desde o início dos anos 90, vários atributos sugiram com o objetivo de realçar as falhas em imagens sísmicas. Chopra e Marfurt (2007) citam alguns deles: atributo de caos (Iske e Randen 2005); coerência (Bahorich e Farmer 1995); variância (Bemmel e Pepper 2000), curvatura (Roberts 2001) e (Boe e Daber 2010), entropia (Cohen, Coult e Vassiliou 2006), entre outros. Outros métodos são baseados em detecção de bordas, como o de Aqrawi et. al. (2011), que segmenta as falhas utilizando uma versão modificada do filtro de Sobel em 3D e o de Pampanelli et. al. (2013), que se baseia no uso de derivadas direcionais de primeira ordem para realçar as descontinuidades ao longo dos horizontes.

Todos esses métodos são baseados na descontinuidade existente nas regiões em que horizontes são interrompidos por falhas sísmicas. Seus resultados são imagens de falhas, cujas superfícies propriamente ditas podem ser extraídas, se desejado. Isso exige um processamento extra, que também pode ser realizado de diferentes maneiras. Por exemplo, Pedersen et. al. (2002) desenvolveram um método de *Ant Tracking* para unir pequenas regiões de baixa continuidade em imagens 3D formando superfícies maiores. Outros trabalhos, como o de Maciel (2014), também utilizam algoritmos baseados em modelos de colônias de formigas para o realce de atributos de falha. E muitos executam o crescimento de superfícies de falhas a partir de uma semente selecionada pelo usuário, como em (Kadlec 2011).

Este trabalho pretende contribuir tanto na identificação visual das falhas nos dados sísmicos, quanto na extração de suas superfícies, compondo um único método automático de detecção. Para a primeira etapa, propõe-se a utilização de funções geoestatísticas como atributos. Estas funções são medidas de variação direcionais e são capazes de indicar a continuidade ou não de dados a uma distância e uma direção pré-determinadas, o que as torna fortes candidatas a representação de textura em imagens 2D ou 3D. As funções semivariograma, semimadograma, covariograma e correlograma, calculadas em diferentes direções no dado sísmico serão utilizadas como características representativas para classificação com SVM (*Support Vector Machine* – Máquina de Vetores de Suporte). Para a segunda etapa, este trabalho propõe uma extração das linhas de falhas em 2D, através de operações morfológicas e segmentação por crescimento de região.

Uma fundamentação teórica sobre os conceitos e técnicas aplicados neste trabalho é apresentada no Capítulo 2. No Capítulo 3, apresentam-se as etapas da metodologia proposta, para a detecção automática de falhas sísmicas, e no Capítulo 4 os resultados numéricos e visuais obtidos com a aplicação da metodologia no dado F3 Block, disponibilizado pelo sistema OpendTect. Por fim, o Capítulo 5 apresenta a conclusão do trabalho e sugere os trabalhos futuros para continuação da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Apresenta-se neste capítulo a base teórica que fundamenta o trabalho. Dividido em quatro seções, o capítulo inicia com uma breve introdução ao método de reflexão sísmica, processo pelo qual é obtido o volume sísmico. A Seção 2.2 traz alguns conceitos básicos de Processamento de Imagens necessários para a compreensão da metodologia desenvolvida. As funções geoestatísticas são apresentadas na Seção 2.3 e, por último, a Seção 2.4 traz importantes conceitos da área de Reconhecimentos de Padrões, incluindo em especial a classificação de dados utilizando a técnica SVM.

2.1 O Método de Reflexão Sísmica

O Método de Reflexão Sísmica é um método indireto de exploração da subsuperfície da terra. Tem como objetivo a construção de um modelo de dados que, após serem processados e interpretados, representem devidamente a geologia de uma região de interesse, para fins de exploração de recursos minerais e de hidrocarbonetos (petróleo, gás natural) (Figueiredo 2007). Este método vem sendo largamente utilizado por ser capaz de cobrir grandes áreas e por ser muito mais econômico em comparação a um método direto, como a perfuração de poços (Silva 2004). A teoria apresentada nesta seção é fortemente baseada nos trabalhos de (Silva 2004), (Machado 2008) e (Figueiredo 2007).

O processo de exploração de hidrocarbonetos pode ser dividido, segundo Robinson e Treitel (1980), em três etapas: *aquisição*, *processamento* e *interpretação*.

2.1.1 Aquisição

O subsolo terrestre é composto por diferentes camadas geológicas de sedimentos, as quais possuem características físicas distintas, como por exemplo, a impedância acústica, que é base do processo de aquisição de dados sísmicos.

No processo de aquisição, ilustrado pela Figura 1, ondas sísmicas são produzidas artificialmente por meio de explosões de dinamite (em terra), ou canhões pneumáticos (em regiões marinhas). Essas ondas se propagam através das camadas rochosas até encontrarem interfaces entre duas camadas de rocha com impedâncias acústicas diferentes. Parte delas então é refratada, continuando a viagem para baixo, e outra parte é refletida, retornando à superfície onde estarão dispostas linhas de receptores (geofones no caso terrestre ou hidrofones no caso marítimo) (Silva 2004).

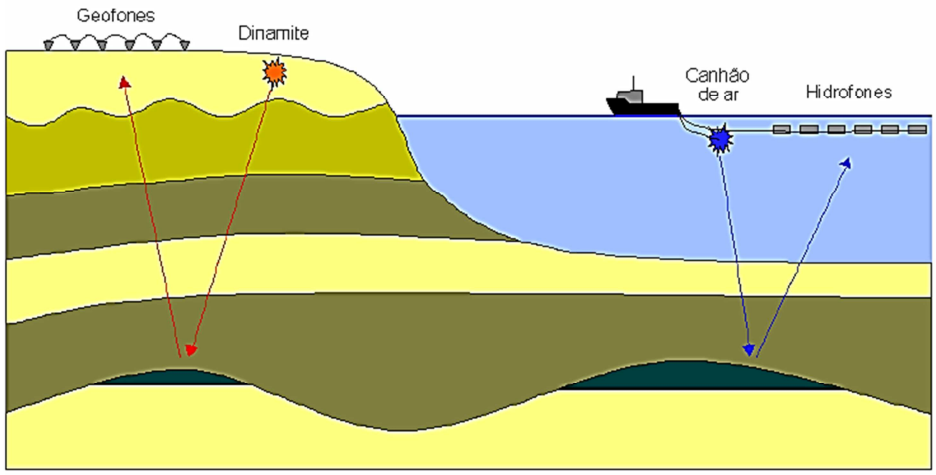


Figura 1 – Processo de aquisição do dado sísmico (Silva 2004).

As informações captadas (valores de amplitude das ondas refletidas e o tempo levado por elas para retornar à superfície) são registradas no sismógrafo e seguem à etapa de processamento.

2.1.2 Processamento

Nesta etapa, alguns erros inerentes ao levantamento sísmico são corrigidos. Além disso, os dados são reorganizados de forma a compor uma grade tridimensional com uma *amostra de amplitude sísmica* em cada vértice da grade. Para isso, algumas transformações são realizadas. Uma delas torna iguais as posições das fontes e dos receptores, como se as ondas tivessem se propagado perfeitamente na vertical (Silva 2004). Outra transformação alinha os receptores, considerando-os todos parte de uma mesma superfície horizontal de referência (Figueiredo 2007).

Assim, o dado tridimensional é composto por duas dimensões espaciais (direções *inline* e *crossline*), que estão relacionadas com o posicionamento das fontes (e dos receptores), e uma dimensão temporal, que corresponde à propagação da onda sísmica, desde a sua geração até o seu registro pelo receptor. Métodos de conversão permitem trocar a variável de tempo por profundidade. Essa conversão depende da velocidade de propagação das ondas em cada tipo de rocha, variando segundo sua porosidade, temperatura, pressão, entre outras características (Figueiredo 2007).

Uma coluna de amostras de mesmas coordenadas espaciais é denominada traço sísmico (Figura 2, à esquerda); uma seção bidimensional, com uma dimensão espacial e outra temporal, é denominada *linha* sísmica (Figura 2, centro). O *volume* sísmico (Figura 2, à direita) é composto por várias linhas sísmicas. Na linha sísmica e no volume sísmico da

figura, o mapa de cores associado relaciona a cor azul aos valores mais negativos de amplitude e a cor vermelha aos mais positivos, passando pelo branco, que indica o valor zero.

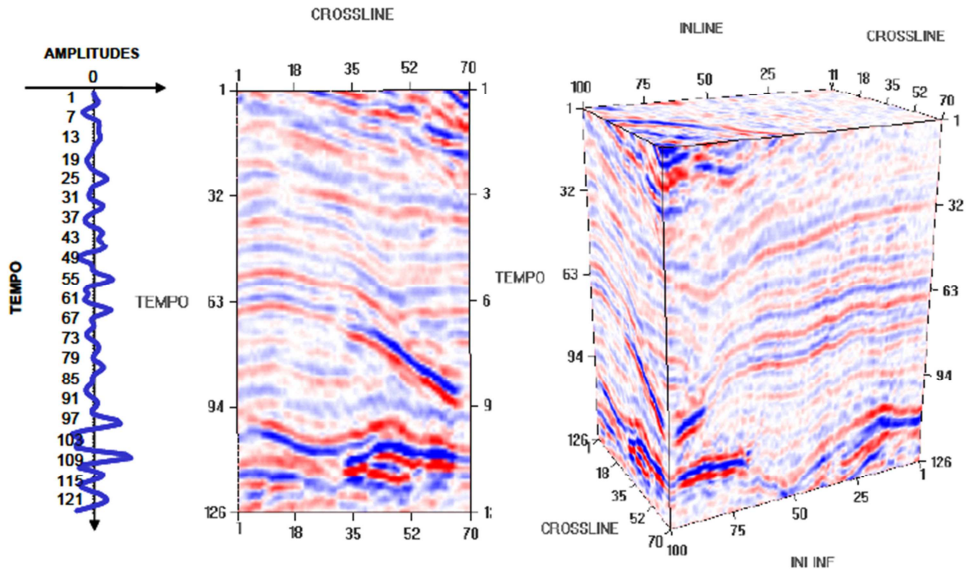


Figura 2 – Dado sísmico. Um traço sísmico (esquerda), uma seção ou linha sísmica (centro) e um volume sísmico (direita) (Silva 2004).

2.1.3 Interpretação

Após a etapa de processamento, o dado sísmico está pronto para ser interpretado por um geólogo ou um geofísico. Nesta etapa, o intérprete visa criar um modelo que represente a geologia contida na área do levantamento. A Figura 3 ilustra dois exemplos de modelos geológicos. Nestes, o especialista identifica as camadas e os elementos presentes na região explorada.

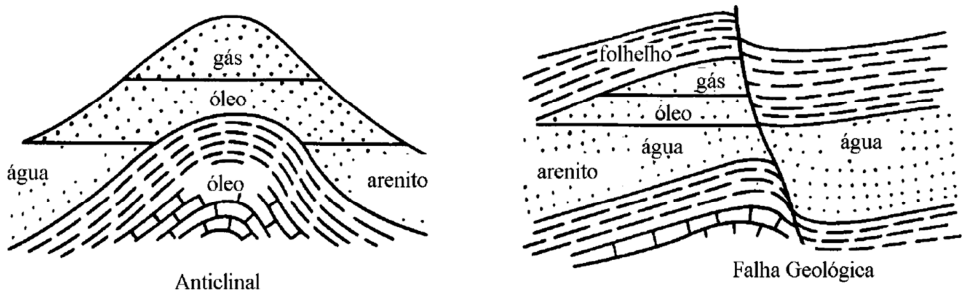


Figura 3 – Exemplo de modelos geológicos (Figueiredo 2007). Adaptado de (Robinson e Treitel 1980).

De acordo com o foco de criação do modelo, classifica-se a interpretação em dois tipos: estrutural e estratigráfica. O objetivo da interpretação estrutural é identificar as camadas geológicas, modelando principalmente horizontes e falhas. Para a interpretação estratigráfica, no entanto, interessa conhecer a maneira como as camadas se formaram no decorrer do tempo.

Os horizontes sísmicos, também chamados refletores, indicam a existência de uma interface entre duas camadas de sedimentos. No dado sísmico, eles se apresentam como séries de reflexões contínuas de intensidades similares (picos ou vales de amplitudes sísmicas), encontradas em vizinhanças laterais ao longo do volume (Silva 2004). As falhas, por sua vez, são fraturas de rochas em subsuperfícies, causadas por forças tectônicas. No dado sísmico, elas se mostram como descontinuidades na estrutura de camadas dos horizontes, como ilustrado na Figura 4.

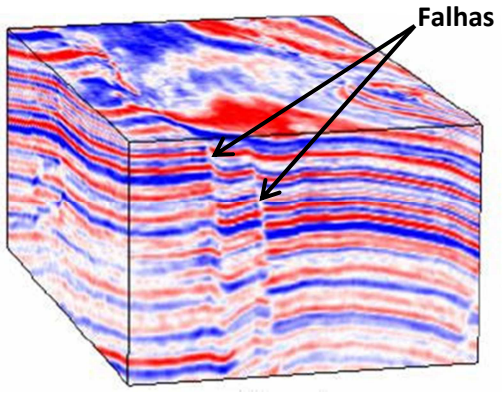


Figura 4 – Exemplo de falhas em um dado sísmico. Adaptada de (Machado 2008).

Para a indústria de petróleo, a identificação de horizontes e falhas permite identificar onde o óleo possa estar aprisionado em um reservatório. A existência de um reservatório não é possível a menos que o óleo esteja impedido de escapar; o que exige a ocorrência de uma camada de rochas impermeáveis, que formem a denominada trapa (armadilha). Assim, como uma falha indica um deslocamento de camadas, a sua presença pode ser responsável tanto pela criação quanto pela ruptura de uma capa selante (camada impermeável), provocando a migração de fluidos (Machado 2008). Na Figura 5(b), tem-se um exemplo de falha que ocasionou um mal selante, já que os fluidos ali presentes podem migrar para fora do reservatório. Enquanto isso, na Figura 5(a), a disposição das camadas impermeáveis, ocasionada pela falha, garante um bom selante.

Existem 33 tipos ou qualificações diferentes para falhas segundo Duarte (2003) apud (Machado 2008, 29). Por exemplo, uma falha pode ser classificada como planar ou curva (ou lístrica), de acordo com a variação da direção da superfície.

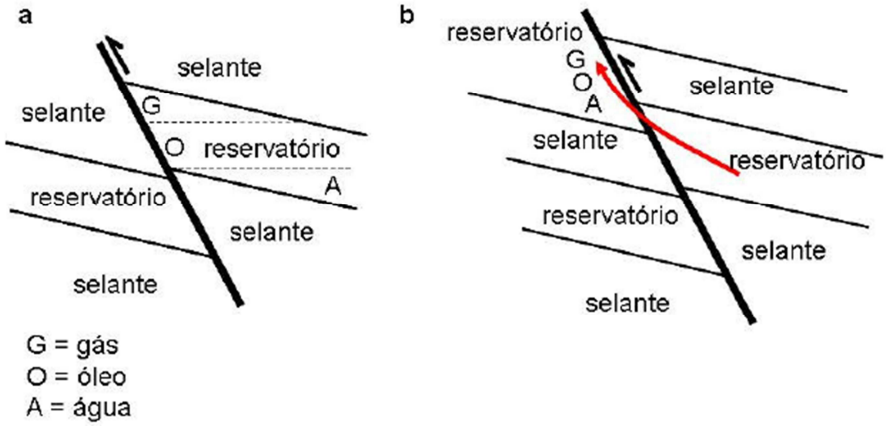
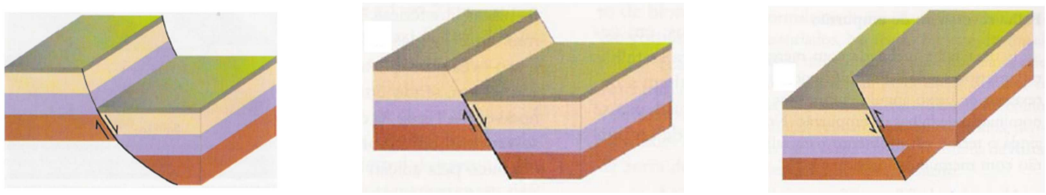


Figura 5 – (a) Exemplo de falha criando um bom selante. (b) Exemplo de falha criando um mal selante. (Machado 2008) Adaptado de (Lines e Newrick 2004).

Também pode ser classificada em normal (de gravidade) ou reversa (de empurrão), considerando o movimento relativo entre os blocos que a originou. Esta é uma classificação difícil de realizar, uma vez que diferentes movimentos de blocos podem resultar em uma mesma configuração final (Machado 2008). Observe, na Figura 6, exemplos desses tipos de falha.



(a) (b) (c)

Figura 6 – Exemplo de falha (a) lístrica, (b) plana normal (ou de gravidade) e (c) plana reversa (de empurrão). Adaptadas de (Machado 2008)

Em geral, no entanto, as falhas geram superfícies aproximadamente verticais, ou seja, com um alto ângulo de mergulho (Machado 2008). As falhas utilizadas neste trabalho são aproximadamente planares e verticais.

A interpretação, e em especial a delimitação das falhas sísmicas, é uma tarefa executada manualmente pelos especialistas, o que demanda bastante tempo e, por fim, faz da automação computacional dessa atividade uma necessidade.

2.2 Conceitos Básicos de Processamento de Imagens

Segundo Gonzalez e Woods (2010), uma imagem digital pode ser definida como uma função bidimensional *f(x, y)*, em que *x* e *y* são coordenadas espaciais discretas de um plano, e a amplitude de *f* em qualquer par de coordenadas *(x, y)* é chamada de intensidade ou nível de cinza da imagem nesse ponto. Uma imagem é, portanto, um conjunto finito de elementos, comumente denominados *pixels*, dispostos regularmente em uma matriz com largura e altura definidas.

Muitas aplicações, entretanto, exigem naturalmente a utilização de uma terceira coordenada, para tornar possível a representação de objetos volumétricos ou áreas geográficas em imagens tridimensionais, como por exemplo, a imagem sísmica[1]. A definição de Gonzalez e Woods necessita então ser estendida para esse caso, qualificando uma imagem como uma função *f(x, y, z)*, onde cada elemento pode ser denominado *voxel* (um *pixel* volumétrico).

O significado físico de cada valor de *pixel* ou *voxel* depende do tipo da imagem e de como ela foi adquirida (Gonzalez e Woods 2010). Por exemplo, um *pixel* de uma imagem fotográfica é obtido através de uma combinação entre a quantidade de iluminação da fonte de luz que incide na cena com a quantidade de iluminação refletida pelos objetos (respectivamente, iluminação e reflectância). O valor de um *voxel* de uma tomografia computadorizada representa a quantidade de raios X absorvidos por um material ou parte do corpo. Em um volume sísmico, cada *voxel* contém o valor da amplitude da onda registrado pelos sensores de ondas sísmicas, no processo de aquisição descrito na Seção 2.1.1.

Visualmente, é uma convenção a representação dos *pixels* ou *voxels* das imagens em tons de cinza, com os menores valores em preto, clareando gradativamente até o branco, que representa os maiores valores. A quantidade de tons possíveis depende da quantidade n de bits utilizados para representar cada *pixel*, e equivale a 2^n.

As subseções seguintes tratam resumidamente de duas técnicas aplicadas em etapas da metodologia deste trabalho: a segmentação por crescimento de região, e as operações baseadas em morfologia matemática para modelagem de objetos e remoção de ruídos. Para mais detalhes sobre essas e outras técnicas, recomenda-se a leitura de (Gonzalez e Woods 2010).

[1] Embora uma das dimensões do volume sísmico seja, em significado, temporal, conforme apresentado na Seção 2.1.2, este fato não será relevante no restante deste trabalho. O volume sísmico será tratado puramente como uma imagem tridimensional, com três dimensões espaciais.

2.2.1 Segmentação por Crescimento de Região

Segmentar uma imagem significa subdividi-la em regiões ou objetos que a compõem (Gonzalez e Woods 2010). Existem muitos tipos de segmentação: alguns baseados na forma, outros na textura dos *pixels*, outros em limiares de intensidade, etc. Segmentar uma imagem não trivial pode ser uma tarefa muito difícil. Por isso é importante escolher adequadamente o tipo de segmentação a ser utilizado em cada aplicação, baseado nos resultados que se deseja obter.

Neste trabalho, utiliza-se uma segmentação simples, conhecida como segmentação por crescimento de região. A partir de um ou mais *pixels* (ou *voxels*) iniciais, denominados sementes, faz-se crescer regiões, anexando pontos vizinhos que obedeçam a uma propriedade pré-estabelecida de similaridade (Gonzalez e Woods 2010).

Existem algumas variáveis a se considerar na execução dessa segmentação, que dependem da natureza do problema e do tipo de imagem utilizada. A primeira questão é a escolha das sementes, ou seja, onde iniciar o crescimento. Outro ponto é a definição de similaridade: o que faz de um novo *pixel* semelhante ao grupo que está em crescimento (pode ser um intervalo predefinido de intensidade ou de cor, por exemplo). Por fim, uma terceira questão é o critério de parada. Por exemplo, pode-se decidir parar quando não houver mais *pixels* semelhantes, ou quando a região crescida atingir um limite de tamanho ou um formato específico, etc (Gonzalez e Woods 2010).

2.2.2 Morfologia Matemática

A teoria dos conjuntos se torna uma ferramenta muito poderosa no processamento de imagens digitais, especialmente quando o campo em questão é o da morfologia matemática. “A palavra morfologia geralmente denota um ramo da biologia que lida com a forma e a estrutura dos animais e das plantas” (Gonzalez e Woods 2010). No Processamento de Imagens, a morfologia matemática lidará com a forma de objetos de interesse, os agrupamentos de *pixels*, com o objetivo de descrever seu comportamento, através da análise de fronteiras, esqueletos, fecho convexo, etc. A morfologia também é largamente utilizada para remover ruídos ou alterar o formato de regiões de acordo com um interesse.

A base da morfologia matemática é a teoria dos conjuntos, os quais são os objetos da imagem. Em uma imagem binária, em que o objeto está em branco e o fundo em preto, por exemplo, o conjunto é formado por todos os *pixels* pertencentes ao objeto, e cada *pixel* é um elemento representado pela sua coordenada *(x,y).*

Duas operações morfológicas básicas são as operações de erosão e dilatação. Ambas são realizadas entre a imagem original do objeto e o chamado “elemento estruturante”. Um elemento estruturante é um conjunto, ou uma subimagem, que será utilizado para examinar uma imagem buscando propriedades de interesse. Ele pode assumir várias formas, dependendo do objetivo da operação. Alguns exemplos de elementos estruturantes são mostrados na Figura 7.

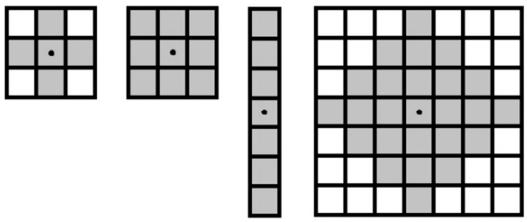


Figura 7 – Elementos estruturantes. Adaptada de (Gonzalez e Woods 2010).

A operação de erosão pode ser definida pela Equação (1). Segundo esta equação, a erosão de *A* por *B* é o conjunto de todos os pontos *z* de forma que *B*, transladado por *z*, está contido em *A* (Gonzalez e Woods 2010).

$$A \ominus B = \{z|(B)_z \subseteq A\} \tag{1}$$

A Figura 8 ilustra o resultado da operação de erosão utilizando uma imagem *A* e um elemento estruturante *B*, com forma primeiramente quadrada e em seguida alongada. Observe que ocorre uma diminuição, ou uma “erosão”, da região do objeto-imagem *A*, completamente controlada pelo tamanho e pelo formato do elemento estruturante utilizado. A erosão é utilizada especialmente para a remoção de elementos.

A dilatação, por sua vez, é definida pela Equação (2), segundo a qual a dilatação de *A* por *B* é o conjunto de todos os deslocamentos, z, de forma que $\hat{B}$ (a reflexão de *B* em torno de sua origem), e A se sobreponham pelo menos por um elemento.

$$A \oplus B = \{z|(\hat{B})_z \cap A \neq \emptyset\} \tag{2}$$

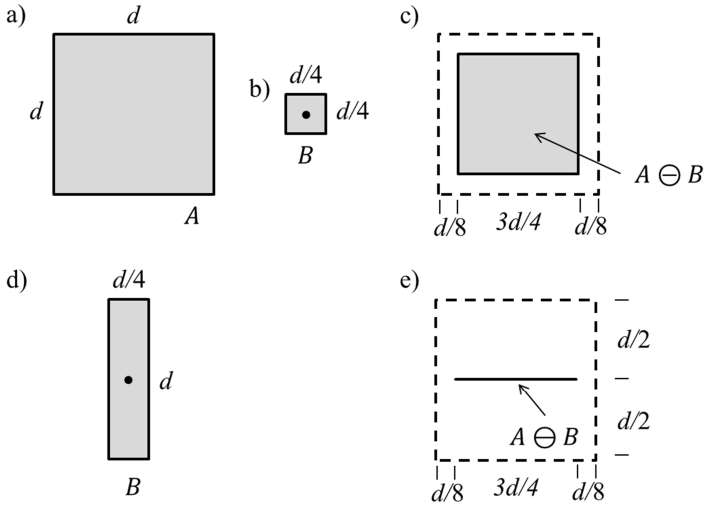


Figura 8 – Exemplos da operação de erosão. a) Conjunto A. b) Elemento estruturante quadrado. c) Erosão de A por B, mostrada sombreada. d) Elemento estruturante alongado. e) Erosão de A por B usando o elemento mostrado em d). Adaptada de (Gonzalez e Woods 2010).

Na Figura 9, pode-se observar o funcionamento da dilatação. A dilatação proporciona o aumento dos objetos e é especialmente utilizada para unir “lacunas” na imagem, geralmente falhas de aquisição que resultam em pequenos buracos. E, da mesma forma que a erosão, seu resultado varia conforme o tamanho e a forma do elemento estruturante utilizado.

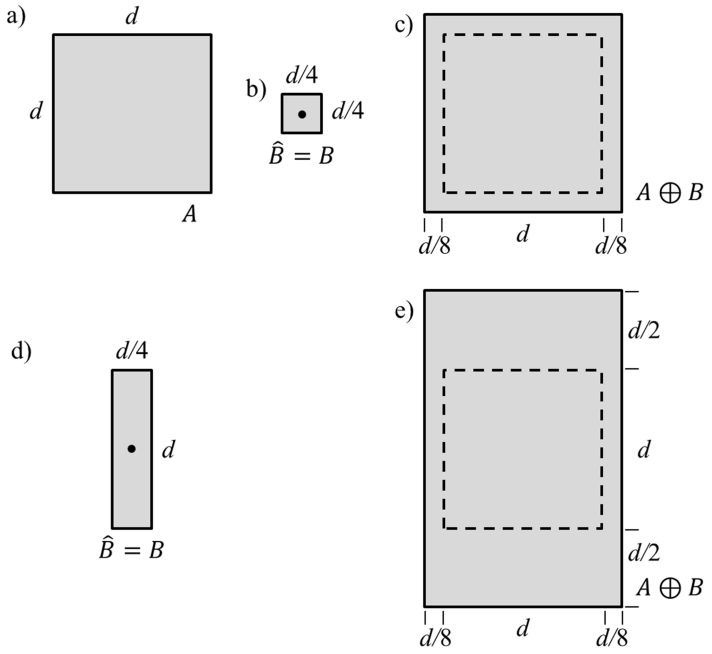


Figura 9 – Exemplos da operação de dilatação. a) Conjunto A. b) Elemento estruturante quadrado. c) Dilatação de A por B, mostrada sombreada. d) Elemento estruturante alongado. e) Dilatação de A por B usando o elemento mostrado em d). Adaptada de (Gonzalez e Woods 2010).

A repetição de seguidas erosões ou seguidas dilatações, ou, ainda, a combinação entre erosões e dilatações pode produzir resultados poderosos. A execução de uma erosão seguida de uma dilatação, utilizando o mesmo elemento estruturante, recebe o nome especial de abertura. A operação na ordem inversa, uma dilatação seguida de erosão, é conhecida por fechamento.

A Figura 10 apresenta um exemplo de aplicação de abertura e fechamento presente neste trabalho. Na Figura 10(a), tem-se uma imagem binária de uma falha sísmica resultante de um processo de classificação. O fechamento da imagem utilizando o elemento estruturante 1x5 da Figura 10(c) une as duas regiões ligeiramente separadas e fecha os demais “buracos” do agrupamento, como mostra a Figura 10(d). E a abertura da imagem, com o mesmo elemento, remove a maior parte dos ruídos presentes na imagem, como mostra a Figura 10(f). Devido ao formato vertical alongado do elemento estruturante, a região final sofre um alongamento também nesta direção.

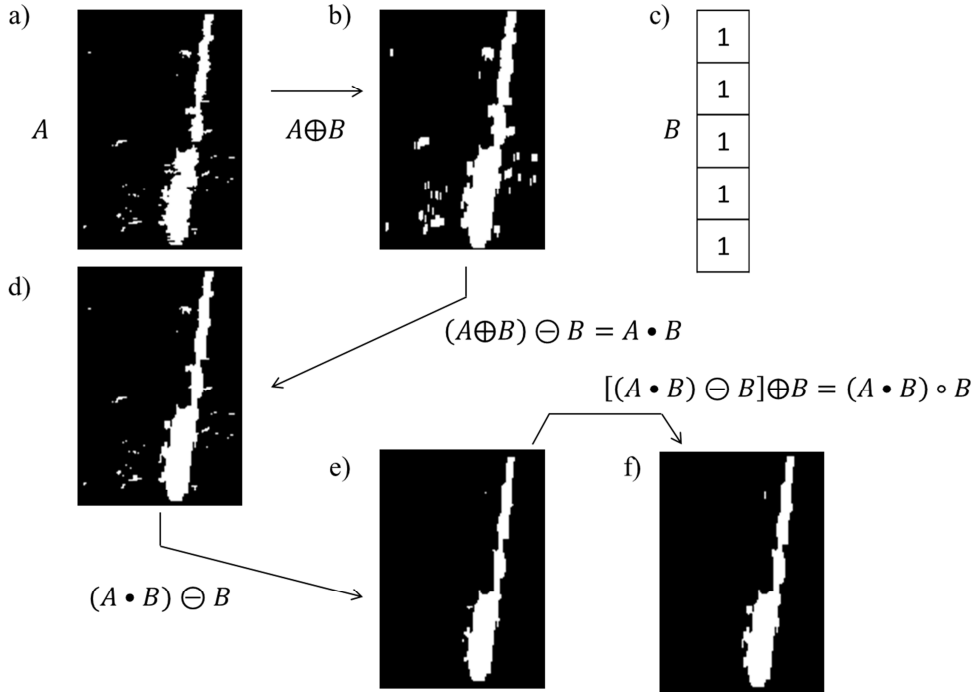


Figura 10 – Exemplo de aplicação de aberturas e fechamentos. a) Imagem original de uma falha. b) Elemento estruturante 1x5. c) Imagem erodida. d) Abertura de A. e) Dilatação da abertura. f) Fechamento da abertura.

Como no exemplo, neste trabalho, as operações morfológicas apresentadas serão aplicadas em imagens binárias. Entretanto, elas podem ser aplicadas também em imagens em níveis de cinza e, além de eficientes na remoção de ruídos e objetos, são também utilizadas para correção de contraste, sombreamento, suavização, etc. Para informações mais

aprofundadas recomenda-se a leitura de Gonzalez e Woods (2010) e de (Soille, Pesaresi e Ouzounis 2011).

2.3 Funções Geoestatísticas

A Geoestatística estuda o comportamento das chamadas variáveis regionalizadas: variáveis às quais está associada uma localização. A maioria dos métodos estatísticos clássicos infelizmente não faz uso da informação espacial inerente aos conjuntos de dados geográficos (Isaaks e Srivastava 1989). Ao longo do tempo, devido ao seu uso quase que exclusivamente prático, porém extremamente bem sucedido em diversas aplicações, a Geoestatística evoluiu do que parecia inicialmente, segundo Isaaks e Srivastava (1989), "aplicações inconsistentes ou adaptações *ad hoc* de modelos bem estabelecidos" para uma área de conhecimento respeitável e de base teórica consistente.

As imagens digitais são funções espaciais e, portanto, naturalmente um campo de aplicação da Geoestatística. As funções geoestatísticas apresentadas a seguir analisam a variabilidade e a correlação espacial entre as amostras (*pixels* ou *voxels*), constituindo assim boas medidas de representação de textura em imagens.

Nas quatro funções geoestatísticas apresentadas, as amostras são definidas por uma função aleatória $Z(u)$. No dado sísmico, $Z(u)$ equivale à amplitude sísmica de cada amostra u do volume, e pode apresentar valores negativos e positivos. Todas as funções são funções de um vetor h, o qual define uma orientação e uma distância para a análise de pares de amostras separadas desta forma.

Sendo x_i a amostra de origem e $x_i + h$ a amostra de extremidade do vetor, a Figura 11 apresenta o exemplo de um possível vetor h, que neste caso representa uma distância de dois *pixels* e um ângulo de 45º em relação à horizontal. Em um espaço tridimensional, a orientação do vetor h pode ser definida por dois ângulos.

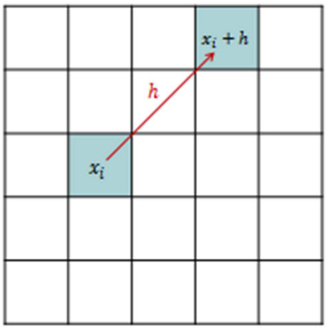


Figura 11 – Ilustração do vetor h.

Nas funções, N equivale ao número total de pares considerados. A primeira função, *semivariograma*, é dada pela Equação (3). Ela mede o grau de dependência entre as amostras, a partir da diferença quadrada de suas intensidades. Com resultado na mesma grandeza da amplitude da função Z, quanto maior o valor de $\gamma(h)$, maior foi constatada a variabilidade entre as amostras separadas por h.

$$\gamma(h) = \frac{1}{2N}\sum_{i=1}^{N}[Z(x_i) - Z(x_i + h)]^2 \tag{3}$$

Considerando o mesmo exemplo da Figura 11, todos os N pares separados por h nesta pequena região são ilustrados na Figura 12.

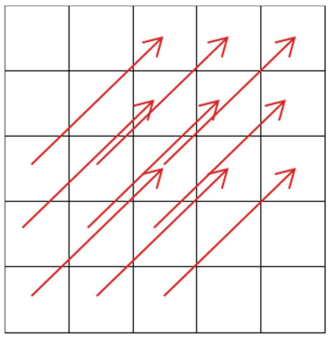


Figura 12 – Todos os pares separados por um determinado vetor h, em uma região.

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:

$$m(h) = \frac{1}{2N}\sum_{i=1}^{N}|Z(x_i) - Z(x_i + h)| \tag{4}$$

A função *covariograma*, por sua vez, é uma medida de correlação entre duas variáveis. Neste caso, três diferentes situações podem ocorrer. Duas variáveis são ditas *positivamente correlacionadas* se altos valores de uma variável tendem a estar associados com altos valores da outra. E da mesma forma, baixos valores associados a baixos valores. Por outro lado, duas variáveis são ditas *negativamente correlacionadas* quando altos valores de uma variável estão associados a valores pequenos da outra, e vice-versa. E por último, duas variáveis podem ser ditas *não correlacionadas*, quando a variação de uma delas aparentemente não implica na alteração da outra.

A função covariograma é dada pela equação:

$$C(h) = \frac{1}{N}\sum_{i=1}^{N}[Z(x_i)Z(x_i + h) - m_{-h}m_{+h}] \tag{5}$$

onde m_{-h} e m_{+h} equivalem respectivamente à média dos valores das amostras de origem e à média dos valores das amostras das extremidades e são calculados através das Equações (6) e (7), respectivamente.

$$m_{-h} = \frac{1}{N}\sum_{i=1}^{N(h)} Z(x_i) \tag{6}$$

$$m_{+h} = \frac{1}{N}\sum_{i=1}^{N} Z(x_i + h) \tag{7}$$

Por último, a função *correlograma*, apresentada na Equação (8), é a versão normalizada da função covariograma, com valores variando entre -1 (variáveis negativamente correlacionadas) e +1 (variáveis positivamente correlacionadas). Isto é obtido dividindo-se o valor de covariograma pelo produto dos desvios-padrões das amostras de origem e de extremidade.

$$\rho(h) = \frac{C(h)}{\sigma_{-h}\sigma_{+h}} \tag{8}$$

Os desvios-padrões são calculados através das equações:

$$\sigma_{-h} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Z(x_i)^2 - {m_{-h}}^2]} \tag{9}$$

$$\sigma_{+h} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Z(x_i + h)^2 - {m_{+h}}^2]} \tag{10}$$

2.4 Reconhecimento de Padrões

A facilidade com que o ser humano distingue naturalmente padrões do mundo real (reconhece faces, entende palavras, lê e escreve caracteres, identifica objetos pelo tato, etc.) tem sido crucial para a sua sobrevivência, há milhares de anos.

É natural que se deseje estender essa capacidade às máquinas, tornando-as capazes de tomar decisões baseadas em padrões observados. Sistemas automáticos de reconhecimento de fala, impressão digital, identificação de sequências de DNA são exemplos da utilidade dessa área.

Os sistemas de reconhecimentos de padrões, segundo Duda et. al. (2001), podem ser divididos nas seguintes etapas, apresentadas na Figura 13:

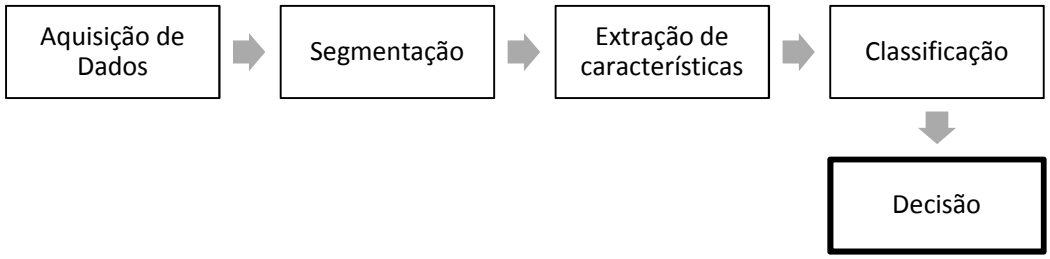


Figura 13 – Etapas de um sistema de reconhecimento de padrões. Adaptada de (Duda, Hart and Stork 2001).

Os dados de entrada são normalmente adquiridos por algum sensor, que depende do tipo da aplicação. Por exemplo, o sensor pode ser um microfone; uma câmera fotográfica; ou um sismógrafo, no caso da aquisição de dados sísmicos.

A segmentação consiste na etapa de separação dos dados de interesse (objetos) do restante dos dados. Ela pode ser realizada manualmente ou de forma automática e, neste caso, constituir uma área complexa dentro de reconhecimento de padrões e extremamente dependente da aplicação. Como exemplos de segmentação, citam-se as tarefas de: separar os diferentes fonemas em uma sequência de áudio; ou diferentes caracteres em uma imagem; separar candidatos a nódulos em uma tomografia pulmonar; entre outros.

Uma vez que os objetos estejam segmentados, a etapa seguinte é a extração de características. O objetivo tradicional dessa etapa, segundo Duda et. al. (2001), é caracterizar um objeto a ser classificado utilizando medidas cujos valores sejam bastante similares para objetos de uma mesma categoria, e bastante diferentes para objetos de categorias diferentes. Quanto melhores as medidas neste sentido, mais fácil se tornará o trabalho do classificador.

A classificação é a atribuição de um objeto a uma categoria. Na maioria dos casos, uma classificação perfeita dos dados é impossível. Isto se deve geralmente a uma variação não

desejada nas medidas de objetos de uma mesma classe: evento que pode ser causado principalmente pela presença de ruídos.

Pode-se acrescentar à etapa de classificação uma subetapa: a de aprendizado (ou *learning*). É praticamente impossível predizer a categoria de determinado objeto sem um conhecimento prévio das características que melhor o representam. Por isso, utiliza-se um conjunto de dados, os quais serão chamados dados de treinamento (*learning data*).

O aprendizado pode acontecer de duas formas: supervisionada ou não supervisionada. No primeiro caso, cada objeto do conjunto de treinamento já tem as suas categorias definidas e o treinamento é realizado de forma a minimizar o erro, com base nessas categorias-resposta. No segundo caso, o classificador tem que descobrir sozinho as relações, regularidades ou categorias que lhe permitam classificar os dados que lhe são apresentados, baseando-se em medidas de qualidade de separação.

2.4.1 Avaliação da Classificação

Existem diversas medidas de avaliação do desempenho de um classificador. A métrica de avaliação mais adequada depende do tipo de dado em questão e da natureza da classificação realizada. Apresentam-se nesta seção algumas medidas estatísticas largamente utilizadas em classificações binárias (duas classes de objetos) realizadas com aprendizado supervisionado. Todas elas são baseadas na chamada matriz de confusão, ilustrada na Tabela 1.

Tabela 1 – Matriz de confusão.

		Valor verdadeiro (mundo real)	
		Classe A (positivo)	**Classe B (negativo)**
Resultado da classificação	**Classe A (positivo)**	VP (verdadeiros positivos)	FP (falsos positivos)
	Classe B (negativo)	FN (falsos negativos)	VN (verdadeiros negativos)

A tabela confronta os valores gerados como resposta por um classificador com os valores reais previamente conhecidos. Considerando duas classes A e B (ou classes positiva e negativa), quatro situações podem ocorrer: a) um objeto A pode ser classificado como pertencente à classe A, constituindo um caso de verdadeiro positivo; b) por outro lado, um objeto A pode ser classificado como pertencente à classe B, sendo um caso de erro e um falso

negativo; c) da mesma forma, um objeto B pode ser classificado como A (falso positivo) ou d) um objeto B pode ser classificado corretamente como B (verdadeiro negativo).

Utilizando o somatório desses casos, podem-se calcular as medidas de acurácia, sensibilidade e especificidade. A taxa de acurácia apresentada na Equação (11) é a medida que indica o total de acertos da classificação, considerando as duas classes.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{11}$$

Observe que uma alta acurácia, somente, não indica uma boa classificação. Considere um caso em que os dados estejam desbalanceados (muito mais objetos da classe A do que da classe B), e o classificador esteja classificando qualquer objeto como A. Obviamente, este não é um bom classificador. No entanto, o número de acertos (acurácia) possivelmente será alto, devido à desproporção dos dados. As medidas de sensibilidade e especificidade são importantes, uma vez que indicam a porcentagem de acertos por classe.

A sensibilidade (também conhecida como taxa de verdadeiros positivos), dada pela Equação (12), mede a proporção de classificações positivas dentre todos os objetos realmente positivos. Similarmente, a especificidade (ou taxa de verdadeiros negativos), dada pela Equação (13), mede a proporção de classificações negativas dentre todos os objetos realmente negativos.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{12}$$

$$Especificidade = \frac{VN}{VN + FP} \tag{13}$$

Um classificador ideal apresentaria tanto 100% de sensibilidade quanto de especificidade. No entanto, essa qualidade dificilmente é atingida na prática e, muitas vezes, é necessário optar pelo resultado mais importante no contexto de uma aplicação. Alterando parâmetros de um classificador, é possível aumentar a sensibilidade ao custo da diminuição da especificidade, e vice-versa.

Um modo de avaliar visualmente esta compensação entre sensibilidade e especificidade é através da chamada Curva ROC (*Receiver Operating Characteristic Curve*). Esta curva é criada projetando-se os valores de sensibilidade e a taxa de falsos positivos (TFP), a qual equivale ao complemento da especificidade, para diferentes configurações de classificação. Cada curva representa o desempenho de um classificador.

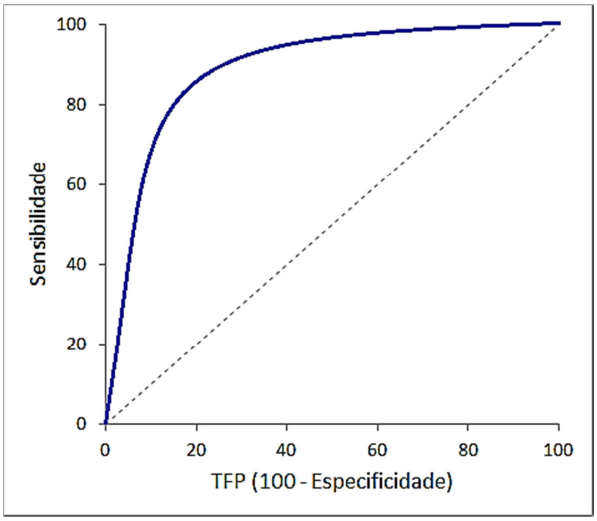


Figura 14 – Exemplos de curvas ROC.

Diferentes limiares de classificação resultam em diferentes valores de sensibilidade e de TFP, sendo o melhor resultado representado na curva ROC pelo ponto mais próximo do canto superior esquerdo do diagrama (0,100), que seria o ponto ideal. Por isso, a exatidão de um teste pode ser quantificada pela área sob a curva: quanto mais a curva se aproxima do canto superior, maior o valor da área sob ela.

2.4.2 Máquina de Vetores de Suporte (SVM)

As Máquinas de Vetores de Suporte, ou SVMs (*Support Vector Machines*), são algoritmos de aprendizagem largamente utilizados para classificação na área de Reconhecimento de Padrões. Sendo uma técnica de aprendizado supervisionado, são utilizados para o treinamento conjuntos de elementos rotulados e representados por suas características numéricas.

Algumas características tornam atraente o uso das SVMs. A primeira delas é a sua base teórica bem estabelecida dentro da Matemática e Estatística. Mas a característica de destaque é com certeza a sua boa capacidade de generalização, demonstrada pelos resultados obtidos em diversas aplicações utilizando SVM. A generalização é a capacidade de classificar corretamente dados não utilizados na base de treino. Em outras palavras, evita-se o *overfitting*: situação em que o classificador se torna muito especializado no conjunto de treinamento, obtendo baixo desempenho quando confronta padrões não conhecidos.

Um conjunto de dados é dito linearmente separável quando é possível separar os padrões das diferentes classes existentes, por ao menos um hiperplano (em duas dimensões, uma reta). Por exemplo, na Figura 15 tem-se um conjunto de dados pertencentes a duas classes: azuis e vermelhas. Inúmeras retas são capazes de separar corretamente os dados de

treinamento. No entanto, a questão é saber qual delas permite aumentar a probabilidade da classificação correta de um novo elemento não conhecido.

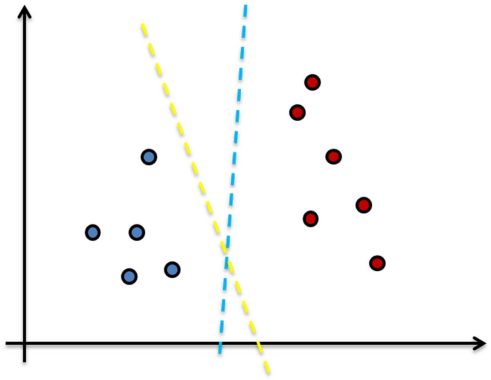


Figura 15 – Duas possíveis separações lineares de um conjunto de dados.

O objetivo do treinamento com SVM é encontrar o hiperplano ótimo (ou hiperplano de margem máxima): aquele que maximiza a margem de separação entre as duas classes, como o ilustrado pela Figura 16. Os pontos sobre as margens máximas são os chamados vetores de suporte.

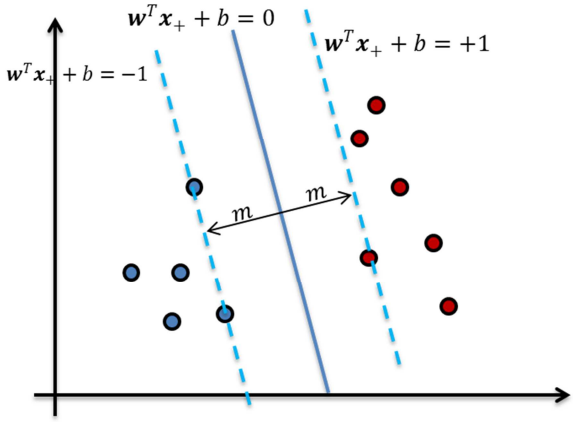


Figura 16 – Hiperplano ótimo de separação entre duas classes de dados.

As definições e equações matemáticas a seguir, que explicam o cálculo do hiperplano de separação SVM são baseadas em (Ben-Hur e Weston 2010), (Chang e Lin 2011) e (Fan, et al. 2008).

O hiperplano de separação pode ser definido pela equação:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0 \tag{14}$$

onde $\boldsymbol{w}^T\boldsymbol{x}$ é o produto escalar entre os vetores $\boldsymbol{w}$ e $\boldsymbol{x}$, em que $\boldsymbol{w}$ é o vetor normal ao hiperplano e *b* é um termo compensador.

Considerando dois pontos mais próximos do hiperplano, um do lado positivo e outro do lado negativo, dados por $\boldsymbol{x}_+$e $\boldsymbol{x}_-$, pode-se expressar a margem do hiperplano, que se deseja maximizar, por:

$$\mathrm{m} = \frac{1}{2}\widehat{\boldsymbol{w}}^T(\boldsymbol{x}_+ - \boldsymbol{x}_-) \tag{15}$$

Sendo $\boldsymbol{x}_+$e $\boldsymbol{x}_-$equidistantes do hiperplano, então:

$$\begin{aligned} f(\boldsymbol{x}_+) &= \boldsymbol{w}^T\boldsymbol{x}_+ + b = +a \\ f(\boldsymbol{x}_-) &= \boldsymbol{w}^T\boldsymbol{x}_- + b = -a \end{aligned} \tag{16}$$

para algum $a > 0$, que pode assumir qualquer valor, uma vez que, multiplicar os pontos por um valor fixo aumentará a margem na mesma quantidade. Portanto, a margem não mudará de fato, apenas as unidades em que ela é medida. Assim, assumiremos $a = 1$ na Equação (16). Subtraindo as duas equações e dividindo ambos os lados por $\|\boldsymbol{w}\|$, obtém-se que:

$$\mathrm{m} = \frac{1}{2}\widehat{\boldsymbol{w}}^T(\boldsymbol{x}_+ - \boldsymbol{x}_-) = \frac{1}{\|\boldsymbol{w}\|} \tag{17}$$

O objetivo então é maximizar a margem definida pela Equação (17). Mas existem algumas restrições. Dado que $\boldsymbol{x}_i$ denota o i-ésimo elemento do conjunto de dados $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ onde y_i equivale à classe correspondente ao elemento $\boldsymbol{x}_i$e pode assumir valor $+1$ (elementos positivos) ou -1 (elementos negativos), deseja-se que:

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x}_i + b \geq +1, & se\ y_i = +1 \\ \boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1, & se\ y_i = -1 \end{cases} \tag{18}$$

Ou simplesmente:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \quad i = 1, \ldots, n \tag{19}$$

Portanto, o classificador de margem máxima é a função discriminante que maximiza a margem geométrica definida pela Equação (17), o que equivale a minimizar:

$$\begin{aligned} &\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{Restrições: } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \quad i = 1, \dots, n \end{aligned} \tag{20}$$

A minimização expressa na Equação (20) considera o caso de conjuntos de dados linearmente separáveis, configurando o que se denomina de SVM de margens rígidas. Na prática, entretanto, os dados raramente são linearmente separáveis, e mesmo se o forem, uma margem maior pode ser obtida se for permitido ao classificador a classificação errônea de alguns pontos.

Para permitir uma suavização nas restrições impostas e obter o que se denomina uma SVM linear com margens suaves, introduz-se uma nova variável, ξ_i, chamada variável de relaxamento. Esta variável permite que um exemplo esteja dentro da margem de erro ($0 \leq \xi_i \leq 1$) ou seja classificado incorretamente ($\xi_i > 1$). O termo a ser minimizado é acrescido em $C\sum_{i=1}^{n}\xi_i$, para penalizar os erros, e assim o problema de otimização se torna:

$$\begin{aligned} &\min_{\boldsymbol{w},b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{Restrições: } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0 \end{aligned} \tag{21}$$

A constante C define a importância relativa de maximizar a margem e minimizar a quantidade de folga. Quando os dados são desbalanceados, ou seja, há mais amostras disponíveis de uma determinada classe do que de outra, pode-se utilizar valores diferentes para a constante C, conforme a classe (Chang e Lin 2011), desta forma atribuindo pesos às classes. Assim, a Equação (21) pode ser reescrita da seguinte forma (forma denominada primal), com C^+ como constante de erro para as amostras da classe positiva e C^- como constante de erro para as amostras da classe negativa.

$$\begin{aligned} &\min_{\boldsymbol{w},b,\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C^+\sum_{i=1,y_i=+1}^{n}\xi_i + C^-\sum_{i=1,y_i=-1}^{n}\xi_i \\ &\text{Restrições: } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0 \end{aligned} \tag{22}$$

Utilizando o método de multiplicadores de Lagrange, obtém-se a forma dual da formulação acima:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$
$$\text{Restrições: } \begin{cases} \sum_{j=1}^{n} y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C^+, se\ y_i = +1, \\ 0 \leq \alpha_i \leq C^-, se\ y_i = -1 \end{cases} \tag{23}$$

Essa formulação dual leva a uma expressão do vetor $\boldsymbol{w}$, que depende do conjunto de dados somente através de produtos escalares:

$$\boldsymbol{w} = \sum_{i=1}^{n} y_i \alpha_i \boldsymbol{x}_i \tag{24}$$

A otimização do problema SVM é realizada tradicionalmente utilizando a forma dual, e somente recentemente tem-se mostrado que a formulação primal é capaz de levar a um treinamento eficiente (Ben-Hur e Weston 2010).

A utilização de margens suaves permite o treinamento de dados que não são perfeitamente lineares, tolerando a existência de ruídos. Entretanto, em muitos casos, não é possível separar satisfatoriamente os pontos diretamente utilizando um hiperplano. Para isso, é comum tornar o SVM um método de classificação não linear, através da utilização de funções Kernel.

A ideia fundamental é transformar as entradas para um novo espaço de maior dimensão, em que os dados se tornem linearmente separáveis e seja possível executar normalmente o SVM.

Por exemplo, na Figura 17, os dados de entrada de duas dimensões, à esquerda, são separáveis não linearmente por um elipsoide. Transformando-os utilizando os monômios de segunda ordem ${x_1}^2$, $\sqrt{2}x_1x_2$ e ,${x_2}^2$ os dados se tornam separáveis por um hiperplano, como mostrado no lado direito da figura.

As funções Kernel têm uma forte base teórica, utilizada por vários algoritmos de classificação ou regressão, além das SVMs. Algumas das funções mais utilizadas são apresentadas na Tabela 2. Para maior aprofundamento, recomenda-se a leitura de (Müller, et al. 2001).

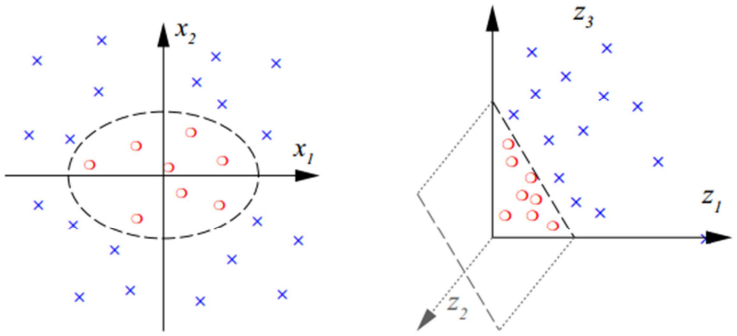


Figura 17 – Exemplo de classificação de dados de duas dimensões. Usando a transformação $({x_1}^2, \sqrt{2}x_1x_2, {x_2}^2)$, os elementos no espaço original à esquerda podem ser separados por um hiperplano (direita). Extraída de (Müller, et al. 2001).

Tabela 2 – Funções Kernel mais utilizadas. (Müller, et al. 2001)

RBF Gaussiano	$k(x,y) = \exp\left(\frac{-\|x-y\|^2}{c}\right)$
Polinomial	$((x \cdot y) + \theta)^d$
Sigmoidal	$\tanh(\kappa(x \cdot y) + \theta)$
Multiquádrico inverso	$\frac{1}{\sqrt{\|x-y\|^2 + c^2}}$

3 METODOLOGIA PROPOSTA

Neste capítulo, apresenta-se a metodologia de detecção de falhas proposta, cujas etapas estão representadas no fluxograma da Figura 18. Como entrada, espera-se uma imagem sísmica tridimensional com cada amostra sísmica (*voxel*) contendo um valor de amplitude, tal como ilustrado na Figura 2. Dessa imagem, são extraídas as características geoestatísticas que serão utilizadas como padrão representativo de textura na etapa de classificação. A Seção 3.1 explica como são extraídas as características do volume e a Seção 3.2 como ocorre a classificação das amostras, com enfoque na criação e validação do modelo SVM. Por fim, a Seção 3.3 apresenta as etapas finais de extração das falhas sísmicas, executada sobre o volume binário resultante da classificação.

Observe que, não coincidentemente, algumas das etapas da metodologia correspondem às etapas clássicas de um sistema baseado em reconhecimento de padrões, apresentadas na Seção 2.4.

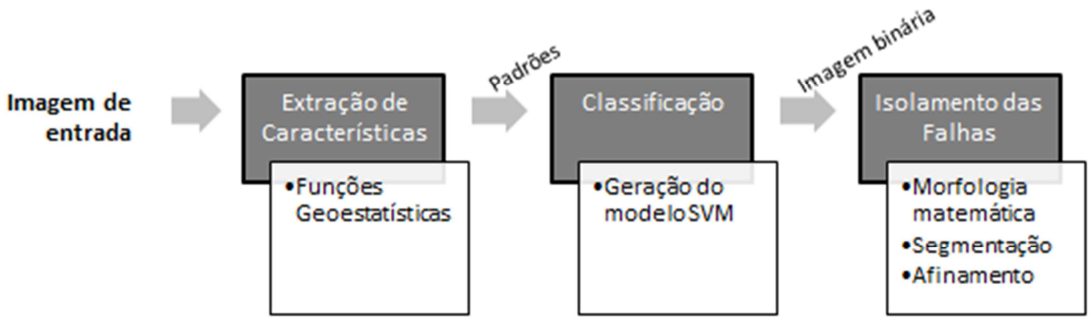


Figura 18 – Fluxograma da metodologia.

3.1 Extração de Características

Na etapa de extração de características, a cada amostra do volume sísmico é atribuído um conjunto de valores capazes de representar a alta variabilidade das amostras vizinhas, no caso de regiões de falhas, ou a baixa variabilidade da vizinhança, no caso de regiões de não falha.

Esses valores são os resultados das funções geoestatísticas de semivariograma, semimadograma, covariograma e correlograma calculados em pares dentro de uma vizinhança definida por uma janela tridimensional de tamanho L x A x P, onde L = largura, A = altura e P = profundidade.

O vetor h, parâmetro das funções geoestatísticas, pode assumir diferentes direções e tamanhos, dentro da janela L x A x P. Treze direções serão consideradas e são ilustradas pela Figura 19. A figura apresenta um cubo composto por um *voxel* central (em amarelo) e seus 26

vizinhos. As direções possíveis a partir desse *voxel* estão destacadas em vermelho. Observe que para cada amostra vermelha existe uma oposta, semitransparente. Ambas, no entanto, estão na mesma direção do *voxel* central (apenas em sentido contrário) e, portanto, contam como uma única direção, totalizando treze possíveis.

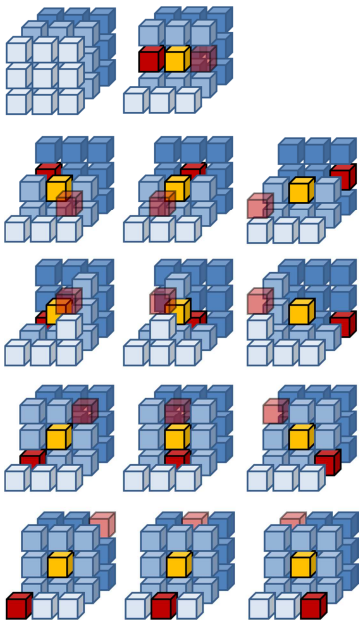


Figura 19 – Possíveis direções do vetor *h* em um espaço discreto tridimensional.

Com relação ao tamanho do vetor h, as distâncias possíveis entre quaisquer pares de *voxels* dependem das dimensões da janela considerada, e estão, portanto, definidas como pertencentes ao conjunto $D_h = \{1, 2, \ldots, min(L, A, P) - 1\}$. Por exemplo, considerando uma janela de dimensões 3x3x3, o vetor h pode assumir um tamanho $d \in D_h = \{1, 2\}$. Assim, no cubo apresentado na Figura 19, segundo a direção horizontal 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 de entrada consiste em:

Para todo *voxel S* da imagem:

1. Centralizar a janela L x A x P sobre S;
2. Calcular, nessa posição e vizinhança, os valores de semivariograma, semimadograma, covariograma e correlograma para todas as variações desejáveis de distância e direção do vetor h;
3. Armazenar os valores encontrados como representantes do *voxel S* (vetor de características).

Assim, considerando o mesmo exemplo de janela (3x3x3), com todas as treze direções e distâncias 1 e 2, cada amostra do volume seria representada por um total de 104 características.

3.2 Classificação

Na etapa de classificação, cada *voxel* do volume sísmico, representado pelo padrão definido pelas funções geoestatísticas, é classificado como um *voxel* de falha ou de "não falha", a partir de um modelo SVM previamente gerado, resultando em um volume binário.

O processo de criação da base de dados para o treinamento e geração do modelo, assim como a sua validação estão descritos na Seção 3.2.1 a seguir.

3.2.1 Geração do Modelo SVM

A Figura 20 apresenta as etapas necessárias para a criação e a validação do modelo de classificação baseado em SVM.

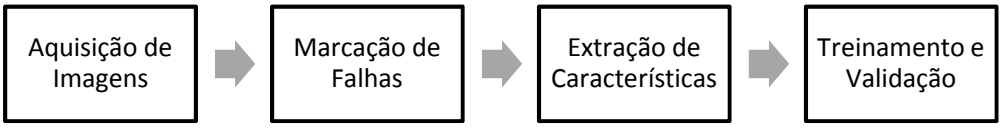


Figura 20 – Etapas de geração e validação do Modelo SVM de classificação de falhas.

3.2.1.1 Aquisição de imagens

Na etapa de aquisição de imagens, foram extraídos dois subvolumes do dado sísmico tridimensional *Netherlands offshore F3-block*[2]. Este dado possui dimensões 950 x 461 x 580, totalizando mais de 254 milhões de amostras sísmicas.

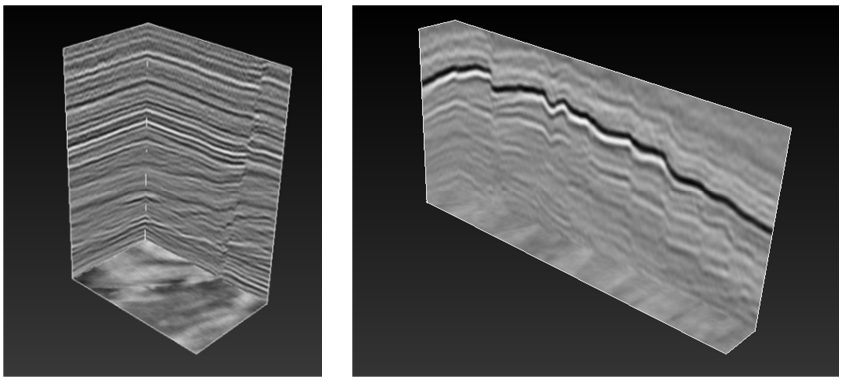
A seleção dos dois subvolumes foi realizada primeiramente com o objetivo de reduzir o número de amostras para treinamento. Uma quantidade tão grande quanto 254 milhões de amostras, além de proporcionar um aprendizado de elevado custo computacional, não é necessária, desde que os subvolumes sejam capazes de representar o restante das amostras, e seu treinamento resulte em uma boa generalização. Além disso, os subvolumes foram estrategicamente selecionados por conter regiões de falhas sísmicas.

[2] Dados públicos e gratuitos disponibilizados no *Open Seismic Repository*, do software OpendTect. (https://opendtect.org/osr/pmwiki.php/Main/HomePage)

As dimensões dos dois subvolumes, identificados por V1 e V2, são:

- V1: 110 x 160 x 70 *voxels*
- V2: 160 x 80 x 20 *voxels*

A Figura 21 apresenta seções dos dois subvolumes extraídos: à esquerda o subvolume V1 e à direita o subvolume V2.

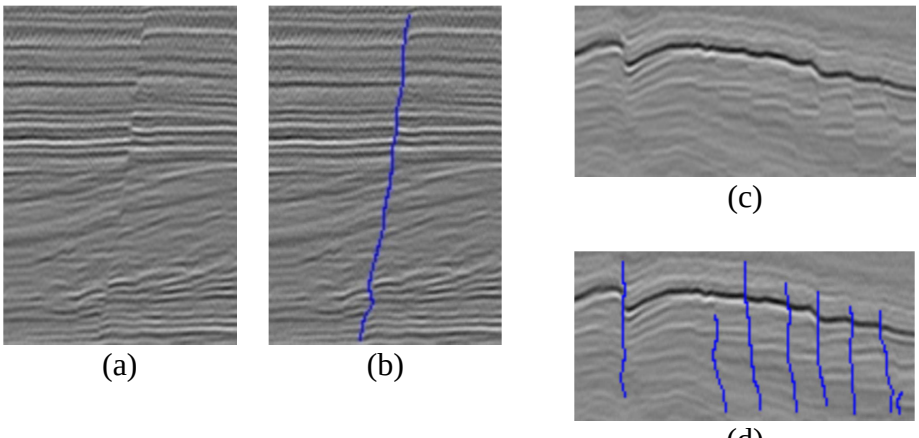


V1 V2

Figura 21 – Os dois subvolumes selecionados para a criação da base de amostras de falha e não falha. (Escalas diferentes)

3.2.1.2 Marcação de Falhas

A marcação de falhas consiste em uma etapa manual de definição por um especialista, fatia a fatia, de quais *voxels* são *voxels* de falha. A imagem V1 contém uma falha sísmica evidente, como se pode ver em uma de suas seções, mostrada na Figura 22(a), e na sua marcação correspondente, na Figura 22(b). Na Figura 22(c), tem-se uma fatia do volume V2 e na Figura 22(d) a sua marcação manual. O volume V2 contém diversas falhas consecutivas, em forma de “degraus”.



(a) (b) (c) (d)

Figura 22 – Seções bidimensionais dos dois subvolumes com as suas falhas marcadas. (a) e (c) Seções originais. (b) e (d) Marcações manuais.

3.2.1.3 Extração de Características

A extração de características para a geração do modelo é realizada exatamente como descrito na Seção 3.1, nos dois subvolumes. A partir das marcações de falhas, é possível criar então um conjunto de dados (vetores de características) de falha e um conjunto de dados de "não falha", muito maior do que o primeiro, uma vez que a quantidade de *voxels* de "não falha" é naturalmente maior.

3.2.1.4 Treinamento e teste

Para validar o modelo SVM gerado nesta etapa, parte dos vetores de características calculados é utilizada para o treinamento e outra parte é reservada para teste. As amostras de treino provêm de blocos de fatias selecionados de posições aleatórias nos dois subvolumes, como ilustrado na Figura 23.

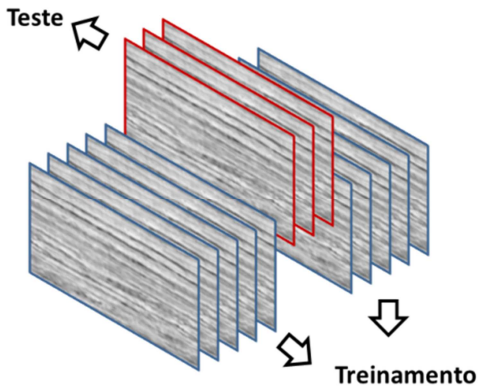


Figura 23 – Separação de amostras para treino e teste.

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). 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, faz-se necessário o uso de pesos diferentes para cada classe a ser classificada, devido ao desbalanceamento entre as classes de falha e "não falha".

3.3 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 as técnicas já abordadas na Seção 2.2.

Na etapa de melhoramento utilizando a morfologia matemática, o volume binário é submetido a aberturas e fechamentos com elemento estruturante de dimensão 1x5 (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. Este formato vertical de elemento estruturante foi escolhido devido à natureza vertical da falha. Ele permite o preenchimento maior de buracos nessa direção e provoca a separação de regiões de falhas diferentes que estejam conectadas.

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 24, 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.

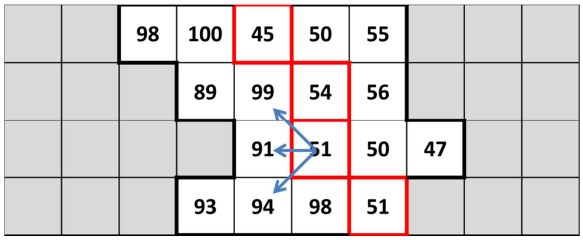


Figura 24 – Ilustração do processo de afinamento.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos utilizando-se a metodologia descrita no Capítulo 3, em duas etapas diferentes. Primeiramente, na Seção 4.1, são apresentados os resultados obtidos com a classificação utilizando SVM, variando-se diversos fatores, como características, porcentagem de treino e pesos entre as classes. Em seguida, na Seção 4.2, são apresentados os resultados visuais das últimas etapas da metodologia: as que compõem a extração das linhas de falha.

4.1 Classificação com SVM

O desempenho da classificação utilizando SVM com as funções geoestatísticas como características descritivas foi avaliado através dos valores de sensibilidade, especificidade e acurácia obtidos em diferentes configurações de treinamento.

As funções geoestatísticas foram testadas individualmente e em conjunto. Os resultados são apresentados a seguir em tabelas. Para cada função (semivariograma, semimadograma, covariograma e correlograma), existem quatro tabelas de resultados, cada uma correspondendo a uma diferente proporção de treino e teste: 20% e 80%; 40% e 60%; 60% e 40%; 80% e 20%. Essas proporções foram escolhidas para avaliar a capacidade de generalização do SVM à medida que se aumenta ou diminui a quantidade de amostras utilizadas para treinamento. Também são apresentadas quatro tabelas para a execução dos testes com as quatro funções geoestatísticas combinadas.

As quantidades de amostras de falha e de não falha em cada divisão são apresentadas na Tabela 3. Os valores não são exatamente correspondentes às porcentagens, pois a divisão é feita utilizando as fatias e não as amostras, da forma indicada pela Figura 23.

Todos os resultados foram obtidos utilizando uma janela de dimensões 7x7x7 para a extração de características. Testes com janelas menores (3x3x3 e 5x5x5) apresentaram resultados inferiores de sensibilidade, especificidade e acurácia, evidenciando que melhores são os resultados quanto maiores as dimensões da janela. No entanto, para dimensões maiores (9x9x9 e 11x11x11), as melhorias nos resultados são pouco significativas, se comparadas ao aumento do tempo computacional utilizado para a extração de característica e para o treinamento.

Tabela 3 – Quantidades de amostras por classe e por divisão.

Classe	Total Amostras	Divisão % de treino / % de teste			
		20 / 80	40 / 60	60 / 40	80 / 20
Falhas	17.496	3.566 / 13.930	7.200 / 10.296	10.345 / 7.151	13.891 / 3.605
Não falhas	1.167.072	238.830 / 928.242	477.592 / 689.480	689.431 / 477.641	928.281 / 238.791
Total	1.184.568	242.396 / 942.172	484.792 / 699.776	699.776 / 484.792	942.172 / 242.396

Foram avaliados os resultados para distâncias entre os pixels (tamanho do vetor *h*) variando de 1 a 6, cumulativamente. A menor distância possível é 1 e a maior distância neste caso é 6, uma vez que o tamanho da janela utilizado é 7x7x7. Na primeira linha de cada tabela, são mostrados os resultados obtidos com distância 1; na segunda linha, com distâncias 1 e 2, e assim sucessivamente. Assim, a quantidade de características aumenta com a distância. Esta quantidade, apresentada também nas tabelas, é obtida utilizando-se o seguinte cálculo: quantidade de funções (igual a 1 para as medidas individualmente, e igual a 4 para todas combinadas) vezes a quantidade de distâncias (variando a cada linha da tabela) vezes a quantidade de direções (sempre utilizadas 13 direções).

As médias e os desvios padrão de sensibilidade, especificidade e acurácia, apresentados em cada linha das tabelas, são resultantes de cinco execuções de treinamento e teste. Em cada execução, novos blocos de treinamento são selecionados, tal como indicado pela Figura 23, na metodologia.

4.1.1 Semivariograma

A Tabela 4 apresenta os resultados obtidos com a função semivariograma utilizando somente 20% da base para treino enquanto 80% foram reservados para teste. Na Tabela 5, a divisão feita com 60% para teste. Na Tabela 6, 40% e na Tabela 7, somente 20%.

Para a função semivariograma, os pesos 1 para não falhas e 80 para falhas resultaram em valores de sensibilidade e especificidade mais balanceados, como mostrado nas tabelas. Para todas as funções foram testados pesos para as falhas variando de 1 a valores bastante altos, com o objetivo de avaliar a compensação entre sensibilidade e especificidade e essa avaliação é apresentada na Seção 4.1.6. Nesta seção, tanto para a função semivariograma quanto para as outras funções, os resultados apresentados são aqueles em que os valores de sensibilidade e especificidade se mostraram semelhantes.

Tabela 4 – Resultados da função **semivariograma**, utilizando 80% da base para teste. Apresentam-se as médias e os desvios padrão de sensibilidade, especificidade e acurácia para cinco execuções de cada combinação de distâncias, as quais variam de 1 até 6.

Distâncias	Total de Características	Porcentagem de Teste: 80%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	79,235%	12,706%	73,350%	12,463%	73,459%	12,140%
1-2	26	78,027%	3,551%	82,526%	2,824%	82,462%	2,740%
1-3	39	84,758%	3,341%	78,060%	3,028%	78,157%	2,936%
1-4	52	80,533%	1,690%	82,787%	1,352%	82,754%	1,323%
1-5	65	80,601%	6,829%	82,174%	4,924%	82,166%	4,785%
1-6	**78**	**81,559%**	**5,895%**	**80,908%**	**3,825%**	**80,917%**	**3,683%**

Tabela 5 – Resultados da função semivariograma, utilizando 60% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 60%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	83,559%	4,911%	73,275%	4,688%	73,426%	4,547%
1-2	26	78,764%	2,549%	81,783%	1,385%	81,739%	1,332%
1-3	39	84,400%	2,771%	80,257%	1,513%	80,317%	1,451%
1-4	**52**	**85,322%**	**0,809%**	**80,762%**	**0,472%**	**80,828%**	**0,474%**
1-5	65	81,021%	4,585%	82,553%	1,533%	82,530%	1,448%
1-6	78	80,358%	4,264%	82,791%	1,594%	82,756%	1,511%

Tabela 6 – Resultados da função **semivariograma**, utilizando 40% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 40%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	85,150%	4,272%	70,033%	3,786%	70,262%	3,670%
1-2	26	76,966%	1,567%	82,778%	0,656%	82,691%	0,623%
1-3	**39**	**83,011%**	**1,640%**	**80,930%**	**0,723%**	**80,961%**	**0,717%**
1-4	52	82,643%	1,695%	81,288%	0,321%	81,308%	0,312%
1-5	65	80,300%	1,650%	82,669%	0,520%	82,633%	0,490%
1-6	78	82,340%	0,527%	81,616%	0,244%	81,627%	0,241%

Tabela 7 – Resultados da função **semivariograma**, utilizando 20% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 20%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	84,327%	2,484%	69,073%	1,639%	69,297%	1,593%
1-2	26	77,087%	1,220%	81,881%	0,379%	81,809%	0,365%
1-3	39	83,782%	3,216%	79,673%	0,564%	79,733%	0,520%
1-4	52	84,723%	2,004%	79,721%	0,397%	79,794%	0,392%
1-5	**65**	**83,838%**	**2,666%**	**80,188%**	**0,468%**	**80,241%**	**0,436%**
1-6	78	80,425%	1,317%	80,456%	0,484%	80,456%	0,461%

As linhas destacadas em azul evidenciam os melhores resultados e as linhas em vermelho destacam os piores. Para essa definição, foram considerados melhores resultados aqueles que apresentaram a maior sensibilidade, mas que não apresentaram tão baixa especificidade (menor que 80%).

Os piores resultados foram sempre obtidos utilizando distância igual a 1 ou até 2, e este comportamento se repete para as demais funções. Este fato indica que levar em consideração somente as amostras de *voxels* bastante próximos é insuficiente para separar os dados.

Os melhores resultados, por sua vez, variaram para os diferentes testes. Para a maioria dos testes com a função semivariograma, quanto maior a distância, maiores a sensibilidade e a especificidade. Para outros, como o da Tabela 7, no entanto, existiu um limite: a distância 6 acabou por misturar informações muito distantes e os valores decaíram. Mas, em geral, os resultados são bastante semelhantes utilizando distâncias máximas de 3 a 6.

Outro fator importante que se pode observar nos resultados da função semivariograma e que também se repetirá para as demais funções é a semelhança dos resultados entre as quatro tabelas de diferentes porcentagens de testes, o que indica uma boa generalização, mesmo para uma quantidade tão pequena quanto 20% para treino.

4.1.2 Semimadograma

As tabelas seguintes mostram os resultados para a função semimadograma. Na Tabela 8, foram reservados 80% da base para teste; na Tabela 9, 60%; na Tabela 10, 40% e na Tabela 11, apenas 20%. Para a função semimadograma, o peso 90 para falhas resultou nos valores mais balanceados de sensibilidade e especificidade. E como se pode observar nas tabelas,

utilizar até as distâncias 4 ou 5 resulta nos melhores índices, neste caso. A distância 6 não contribuiu para a separação das classes.

Tabela 8 – Resultados da função **semimadograma**, utilizando 80% da base para teste. Apresentam-se as médias e os desvios padrão de sensibilidade, especificidade e acurácia para cinco execuções de cada combinação de distâncias, as quais variam de 1 até 6.

Distâncias	Total de Características	Porcentagem de Teste: 80%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	80,106%	9,406%	76,907%	7,248%	76,954%	7,009%
1-2	26	83,175%	0,769%	80,981%	0,427%	81,013%	0,413%
1-3	39	82,880%	3,023%	82,766%	1,443%	82,767%	1,380%
1-4	52	83,198%	2,695%	83,467%	0,909%	83,463%	0,857%
1-5	**65**	**84,211%**	**1,161%**	**83,020%**	**1,761%**	**83,037%**	**1,721%**
1-6	78	83,496%	2,405%	83,559%	1,814%	83,558%	1,768%

Tabela 9 – Resultados da função **semimadograma**, utilizando 60% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 60%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	82,711%	7,225%	76,196%	4,742%	76,292%	4,568%
1-2	26	84,781%	1,327%	80,381%	0,252%	80,445%	0,231%
1-3	39	82,673%	1,631%	82,198%	0,387%	82,205%	0,381%
1-4	**52**	**83,283%**	**1,536%**	**83,015%**	**0,456%**	**83,019%**	**0,428%**
1-5	65	80,299%	1,420%	84,119%	0,236%	84,063%	0,218%
1-6	78	82,616%	2,227%	84,073%	0,472%	84,051%	0,446%

Tabela 10 – Resultados da função **semimadograma**, utilizando 40% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 40%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	87,372%	1,881%	72,320%	1,541%	72,546%	1,489%
1-2	26	83,469%	1,101%	80,288%	0,207%	80,336%	0,189%
1-3	39	84,596%	2,416%	80,740%	0,246%	80,798%	0,228%
1-4	**52**	**83,796%**	**1,598%**	**82,573%**	**0,254%**	**82,592%**	**0,249%**
1-5	65	82,514%	2,628%	83,242%	0,323%	83,231%	0,294%
1-6	78	82,667%	2,829%	83,351%	0,446%	83,340%	0,400%

Tabela 11 – Resultados da função **semimadograma**, utilizando 20% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 20%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	83,170%	2,876%	73,259%	1,624%	73,405%	1,561%
1-2	26	82,423%	0,643%	79,213%	0,115%	79,260%	0,121%
1-3	39	81,540%	4,064%	80,251%	0,598%	80,270%	0,536%
1-4	52	81,621%	2,647%	81,421%	0,312%	81,423%	0,271%
1-5	**65**	**83,200%**	**3,567%**	**82,154%**	**0,279%**	**82,170%**	**0,259%**
1-6	78	80,697%	4,149%	82,105%	0,189%	82,085%	0,147%

4.1.3 Covariograma

Os resultados dos testes utilizando a função covariograma são apresentados na Tabela 12 (80% para teste); na Tabela 13 (60% para teste); na Tabela 14 (40%) e na Tabela 15 (20%). Os pesos utilizados foram 1 para não falhas e 80 para falhas. No caso da função covariograma, os melhores resultados de fato foram obtidos utilizando até a distância 6 em todas as proporções de teste.

Tabela 12 – Resultados da função **covariograma**, utilizando 80% da base para teste. Apresentam-se as médias e os desvios padrão de sensibilidade, especificidade e acurácia para cinco execuções de cada combinação de distâncias, as quais variam de 1 até 6.

Distâncias	Total de Características	Porcentagem de Teste: 80%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	77,198%	5,970%	78,823%	4,953%	78,798%	4,793%
1-2	26	79,537%	6,178%	80,818%	4,125%	80,799%	3,975%
1-3	39	81,774%	1,895%	77,440%	1,044%	77,503%	1,012%
1-4	52	81,938%	3,589%	78,415%	3,956%	78,467%	3,847%
1-5	65	80,632%	3,022%	79,865%	2,230%	79,876%	2,159%
1-6	**78**	**82,085%**	**1,056%**	**79,278%**	**1,945%**	**79,319%**	**1,911%**

Tabela 13 – Resultados da função **covariograma**, utilizando 60% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 60%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	74,724%	4,198%	80,180%	3,564%	80,100%	3,454%

1-2	26	77,859%	3,585%	82,012%	1,975%	81,951%	1,895%
1-3	39	79,370%	1,096%	80,566%	0,497%	80,549%	0,487%
1-4	52	80,613%	3,681%	80,326%	1,303%	80,330%	1,235%
1-5	65	78,785%	3,945%	81,568%	1,957%	81,528%	1,873%
1-6	**78**	**82,191%**	**1,363%**	**80,922%**	**0,683%**	**80,940%**	**0,656%**

Tabela 14 – Resultados da função **covariograma**, utilizando 40% da base para teste.

Distâncias	**Total de Características**	**Porcentagem de Teste: 40%**					
		Sensibilidade		**Especificidade**		**Acurácia**	
		Média	**Desvio padrão**	**Média**	**Desvio padrão**	**Média**	**Desvio padrão**
1	13	73,467%	3,636%	80,100%	1,891%	80,001%	1,813%
1-2	26	74,900%	2,067%	84,101%	0,974%	83,964%	0,931%
1-3	39	75,988%	2,411%	82,989%	0,874%	82,884%	0,826%
1-4	52	75,737%	3,166%	83,804%	1,532%	83,684%	1,462%
1-5	65	77,336%	2,619%	83,639%	1,109%	83,544%	1,052%
1-6	**78**	**78,553%**	**2,766%**	**83,422%**	**1,170%**	**83,349%**	**1,113%**

Tabela 15 – Resultados da função **covariograma**, utilizando 20% da base para teste.

Distâncias	**Total de Características**	**Porcentagem de Teste: 20%**					
		Sensibilidade		**Especificidade**		**Acurácia**	
		Média	**Desvio padrão**	**Média**	**Desvio padrão**	**Média**	**Desvio padrão**
1	13	75,889%	2,053%	77,212%	0,500%	77,193%	0,464%
1-2	26	76,392%	1,898%	83,242%	1,151%	83,143%	1,109%
1-3	39	76,742%	2,625%	82,106%	0,762%	82,029%	0,713%
1-4	52	78,229%	2,171%	81,660%	0,756%	81,611%	0,714%
1-5	65	77,833%	2,638%	82,682%	1,065%	82,611%	1,014%
1-6	**78**	**78,868%**	**2,636%**	**81,944%**	**0,965%**	**81,900%**	**0,916%**

4.1.4 Correlograma

A última medida testada isoladamente, a função correlograma, apresentou os resultados mostrados na Tabela 16 (com 80% da base reservados para teste), na Tabela 17 (60% para teste); na Tabela 18 (40%) e na Tabela 19 (20%).

Utilizou-se peso 1 para não falhas e peso 80 para falhas. A distância 1 manteve-se como pior resultado, mas as distâncias máximas de melhores resultados variaram entre 3 e 6. Esta foi a medida isolada com maiores valores simultâneos de sensibilidade e especificidade.

Tabela 16 – Resultados da função **correlograma**, utilizando 80% da base para teste. Apresentam-se as médias e os desvios padrão de sensibilidade, especificidade e acurácia para cinco execuções de cada combinação de distâncias, as quais variam de 1 até 6.

Distâncias	Total de Características	Porcentagem de Teste: 80%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	84,959%	4,410%	80,987%	2,317%	81,046%	2,223%
1-2	26	80,806%	4,916%	85,399%	1,619%	85,332%	1,539%
1-3	39	79,737%	5,938%	85,837%	1,163%	85,748%	1,061%
1-4	52	81,090%	2,358%	85,652%	1,008%	85,584%	0,986%
1-5	65	83,711%	1,734%	85,406%	0,465%	85,381%	0,462%
1-6	**78**	**83,996%**	**1,635%**	**85,612%**	**0,410%**	**85,588%**	**0,390%**

Tabela 17 – Resultados da função **correlograma**, utilizando 60% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 60%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	81,097%	5,124%	82,864%	1,847%	82,838%	1,749%
1-2	26	84,828%	2,410%	84,534%	0,484%	84,539%	0,450%
1-3	39	83,043%	1,891%	85,277%	0,289%	85,244%	0,264%
1-4	52	84,738%	2,660%	85,239%	0,202%	85,232%	0,201%
1-5	**65**	**85,536%**	**1,255%**	**85,290%**	**0,293%**	**85,293%**	**0,289%**
1-6	78	83,910%	3,729%	85,477%	0,865%	85,454%	0,800%

Tabela 18 – Resultados da função **correlograma**, utilizando 40% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 40%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	82,782%	1,336%	81,219%	0,603%	81,243%	0,575%
1-2	26	83,705%	2,052%	84,045%	0,395%	84,040%	0,363%
1-3	**39**	**85,595%**	**0,863%**	**84,543%**	**0,150%**	**84,559%**	**0,156%**
1-4	52	84,135%	1,656%	84,718%	0,183%	84,709%	0,158%
1-5	65	85,185%	1,384%	84,770%	0,121%	84,777%	0,112%
1-6	78	84,540%	2,062%	84,715%	0,281%	84,713%	0,247%

Tabela 19 – Resultados da função **correlograma**, utilizando 20% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 20%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	13	83,344%	5,333%	80,645%	1,161%	80,686%	1,064%
1-2	26	83,207%	1,643%	83,425%	0,349%	83,422%	0,322%
1-3	39	84,657%	3,285%	83,732%	0,473%	83,745%	0,418%
1-4	**52**	**86,353%**	**1,490%**	**83,644%**	**0,245%**	**83,683%**	**0,223%**
1-5	65	83,132%	3,095%	84,034%	0,477%	84,020%	0,432%
1-6	78	85,355%	3,014%	83,534%	0,475%	83,560%	0,428%

4.1.5 Todas as funções

Utilizando-se as quatro funções combinadas para a criação do modelo, obtiveram-se melhores resultados médios de sensibilidade, especificidade e acurácia. Para isso, foram utilizados pesos 1 e 100 para não falhas e falhas, respectivamente. Tal como nas funções isoladas, a distância 1 exclusivamente gera os piores resultados. Mas os melhores resultados foram obtidos com distâncias até 3 ou 4 pixels.

Apresentam-se os resultados obtidos com 80% da base utilizada para teste, na Tabela 20; 60%, na Tabela 21; 40%, na Tabela 22; e 20%, na Tabela 23.

Tabela 20 – Resultados das quatro funções combinadas: **semivariograma**, **semimadograma**, **covariograma** e **correlograma**, utilizando 80% da base para teste. Apresentam-se as médias e os desvios padrão de sensibilidade, especificidade e acurácia para cinco execuções de cada combinação de distâncias, as quais variam de 1 até 6.

Distâncias	Total de Características	Porcentagem de Teste: 80%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	52	83,804%	5,798%	81,875%	3,229%	81,903%	3,101%
1-2	104	83,759%	5,759%	86,547%	1,467%	86,507%	1,368%
1-3	**156**	**88,610%**	**1,787%**	**85,356%**	**0,442%**	**85,403%**	**0,413%**
1-4	208	84,406%	4,203%	87,086%	1,012%	87,046%	0,935%
1-5	260	84,827%	1,271%	87,156%	0,126%	87,122%	0,115%
1-6	312	84,296%	5,298%	86,938%	1,438%	86,899%	1,340%

Tabela 21 – Resultados das quatro funções combinadas: **semivariograma**, **semimadograma**, **covariograma** e **correlograma**, utilizando 60% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 60%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	52	87,643%	3,078%	80,452%	0,846%	80,558%	0,790%
1-2	104	86,188%	3,273%	85,346%	0,625%	85,359%	0,571%
1-3	**156**	**89,278%**	**1,654%**	**85,485%**	**0,578%**	**85,540%**	**0,584%**
1-4	208	88,646%	1,387%	85,847%	0,564%	85,888%	0,576%
1-5	260	86,025%	3,725%	86,500%	0,416%	86,493%	0,365%
1-6	312	88,120%	1,046%	86,049%	0,420%	86,079%	0,426%

Tabela 22 – Resultados das quatro funções combinadas: **semivariograma**, **semimadograma**, **covariograma** e **correlograma**, utilizando 40% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 40%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	52	86,765%	1,954%	80,085%	0,500%	80,185%	0,471%
1-2	104	88,039%	1,129%	84,292%	0,389%	84,348%	0,376%
1-3	156	87,861%	2,036%	84,558%	0,345%	84,608%	0,330%
1-4	**208**	**88,282%**	**3,669%**	**84,845%**	**0,496%**	**84,897%**	**0,468%**
1-5	260	86,441%	2,578%	85,205%	0,264%	85,224%	0,284%
1-6	312	84,082%	2,729%	85,406%	0,373%	85,387%	0,368%

Tabela 23 – Resultados das quatro funções combinadas: **semivariograma**, **semimadograma**, **covariograma** e **correlograma**, utilizando 20% da base para teste.

Distâncias	Total de Características	Porcentagem de Teste: 20%					
		Sensibilidade		Especificidade		Acurácia	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
1	52	88,547%	2,459%	78,416%	0,576%	78,568%	0,530%
1-2	104	88,067%	2,725%	83,387%	0,526%	83,456%	0,484%
1-3	156	88,637%	2,883%	83,796%	0,492%	83,869%	0,519%
1-4	**208**	**89,673%**	**3,697%**	**83,675%**	**0,228%**	**83,762%**	**0,204%**
1-5	260	84,938%	2,767%	84,314%	0,569%	84,323%	0,536%
1-6	312	87,539%	5,037%	84,304%	0,423%	84,351%	0,395%

4.1.6 Definição dos pesos das classes

Os pesos 1 para não falhas e 80, 90 ou 100 para falhas proporcionaram os resultados mais balanceados de sensibilidade e especificidade, com ambos valores acima de 80% nos melhores casos, conforme apresentado na seção anterior. No entanto, de acordo com a necessidade, pode-se dar a preferência para uma classe ou para outra, variando estes pesos.

Quando pesos iguais são utilizados, existe uma tendência a se classificar as amostras como não falha, dado o natural desbalanceamento do conjunto de dados. À medida que se aumenta o peso para a classe das falhas, no intuito de compensar na classificação a situação do mundo real, obtêm-se maiores valores de sensibilidade ao custo da diminuição da especificidade. Em outras palavras, aumenta-se a quantidade de falsos positivos.

Para analisar esta variação, os mesmos testes apresentados na Seção 4.1 foram realizados, porém fixando-se o peso 1 para as amostras de não falha e variando os pesos para a classe de falha. Os pesos foram variados[3] de 1, que proporcionou *sensibilidade* aproximadamente igual a 0%, até 3000, que proporcionou *especificidade* aproximadamente igual a 0%. Esses valores compõem os gráficos de curvas ROC a seguir, geradas com os testes realizados com a proporção de 40% para treino e 60% para teste.

Cada curva equivale a vários testes (variando os pesos) para uma determinada configuração de distância (1, 1-2, 1-3, 1-4, 1-5 ou 1-6). Assim, também é possível comparar graficamente o desempenho entre as diferentes distâncias. E cada ponto de uma curva equivale ao resultado médio de sensibilidade e especificidade de cinco testes realizados com um determinado par de pesos.

Na Figura 25, tem-se as curvas ROC da função semivariograma. Dado que o melhor classificador é aquele cuja curva mais se aproxima do ponto superior esquerdo (100% de sensibilidade e 0% de TFP), no caso da função semivariograma, esta curva é a resultante das classificações utilizando distâncias 1, 2, 3 até 4, sendo as curvas de distâncias máximas 3, 5 ou 6 bastante próximas a ela.

[3] Variaram-se os pesos de 10 em 10, até o valor 100. A partir deste peso, todas as funções já ultrapassaram os melhores pontos nas curvas ROC (mais próximos de 0, 100), e os restantes dos testes foram então realizados variando-se os pesos de 300 em 300, inicialmente de 300 até o valor 3000, quando a maioria das curvas se aproxima do ponto (100, 100).

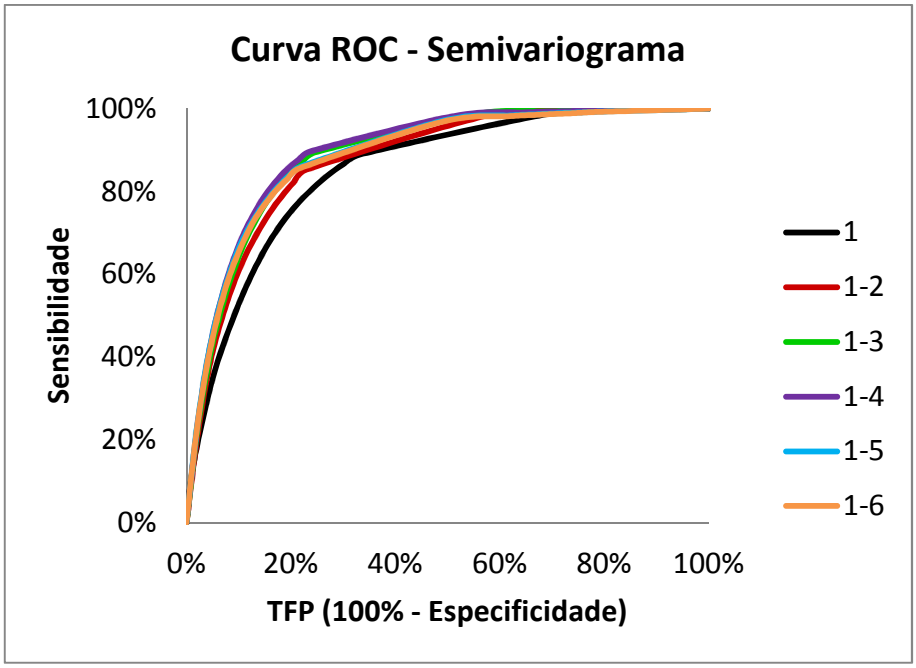


Figura 25 – Curva ROC dos testes com a função semivariograma.

As curvas da função semimadograma são bastante semelhantes entre si exceto somente a distância 1, como se pode observar na Figura 26, onde a curva de distância máxima 6 se destaca ligeiramente entre as outras.

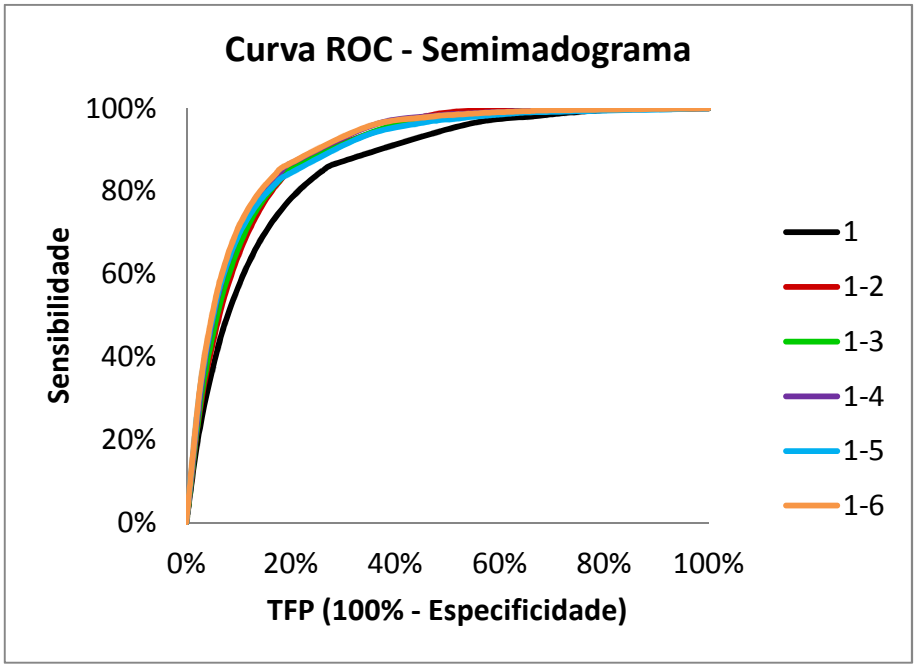


Figura 26 – Curva ROC dos testes com a função semimadograma.

Nos testes com a função covariograma, a distância máxima igual a 6 também proporcionou melhores resultados no geral, como apontam as curvas ROC da Figura 27.

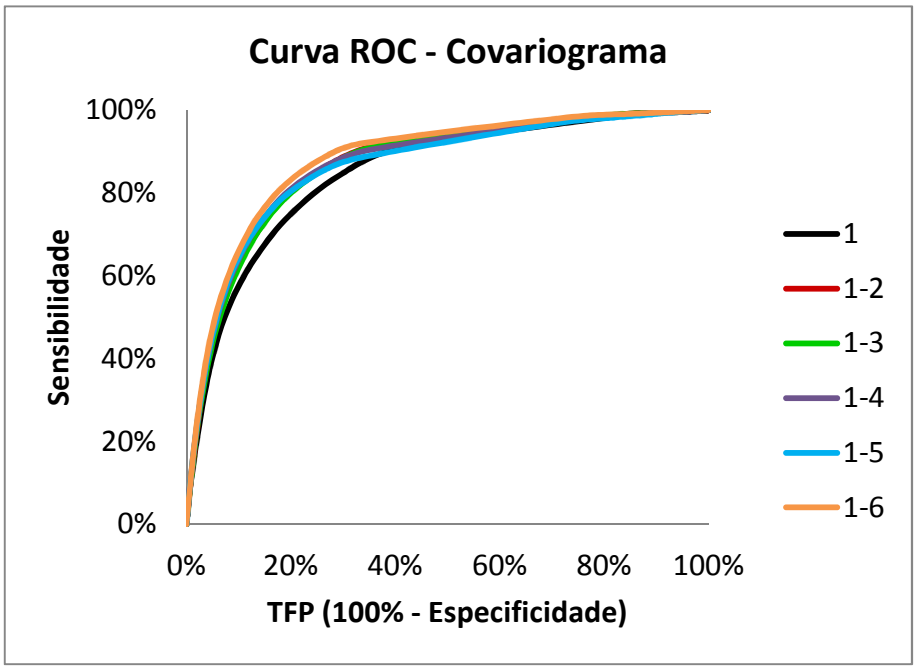


Figura 27 – Curva ROC dos testes com a função covariograma.

A função correlograma apresentou os melhores resultados isoladamente. Na Figura 28, pode-se ver que a curva de distância máxima 5 apresentou o melhor desempenho, mas novamente muito semelhante ao das outras distâncias, exceto ao da curva de distância unicamente igual a 1: a pior classificação, como nas outras funções.

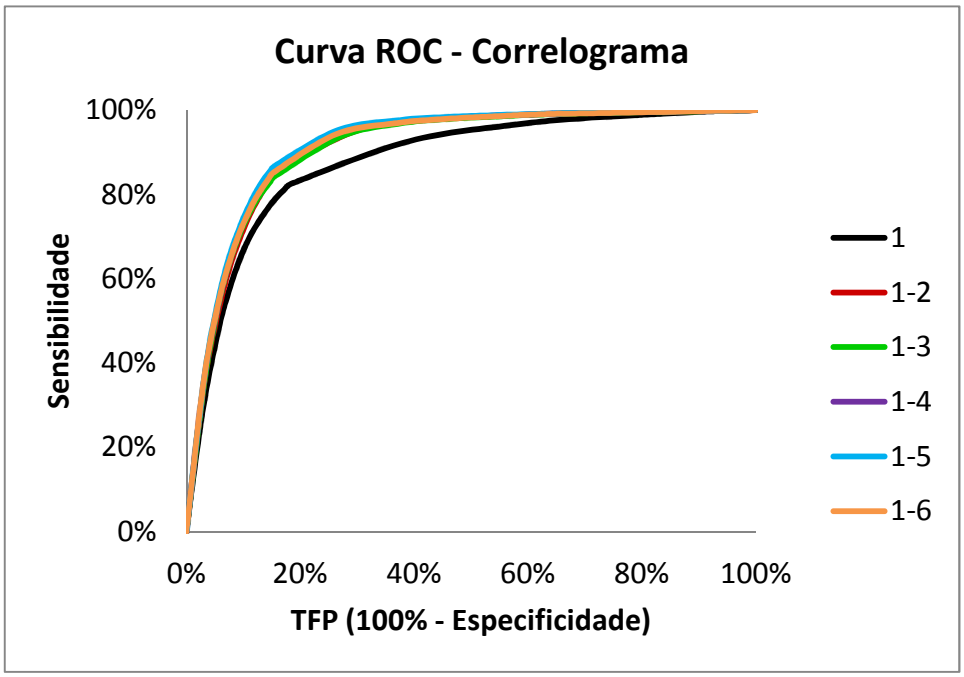


Figura 28 – Curva ROC dos testes com a função correlograma.

Gerando modelos SVM utilizando como características todas as funções geoestatísticas apresentadas neste trabalho, obtiveram-se os melhores resultados, como indicam as curvas ROC da Figura 29, mais próximas do ponto ideal de classificação.

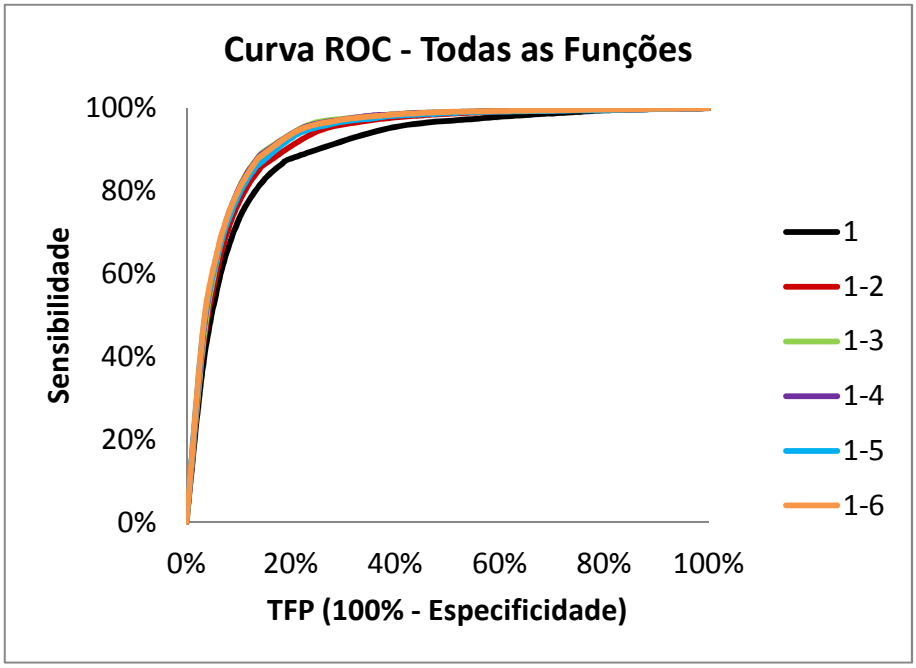


Figura 29 – Curva ROC dos testes com todas as funções geoestatísticas.

A melhor curva de cada gráfico apresentado foi redesenhada na Figura 30. Assim é possível avaliar visualmente quais funções constituem melhores medidas de classificação de falhas e não falhas, sendo a combinação de todas elas a melhor opção.

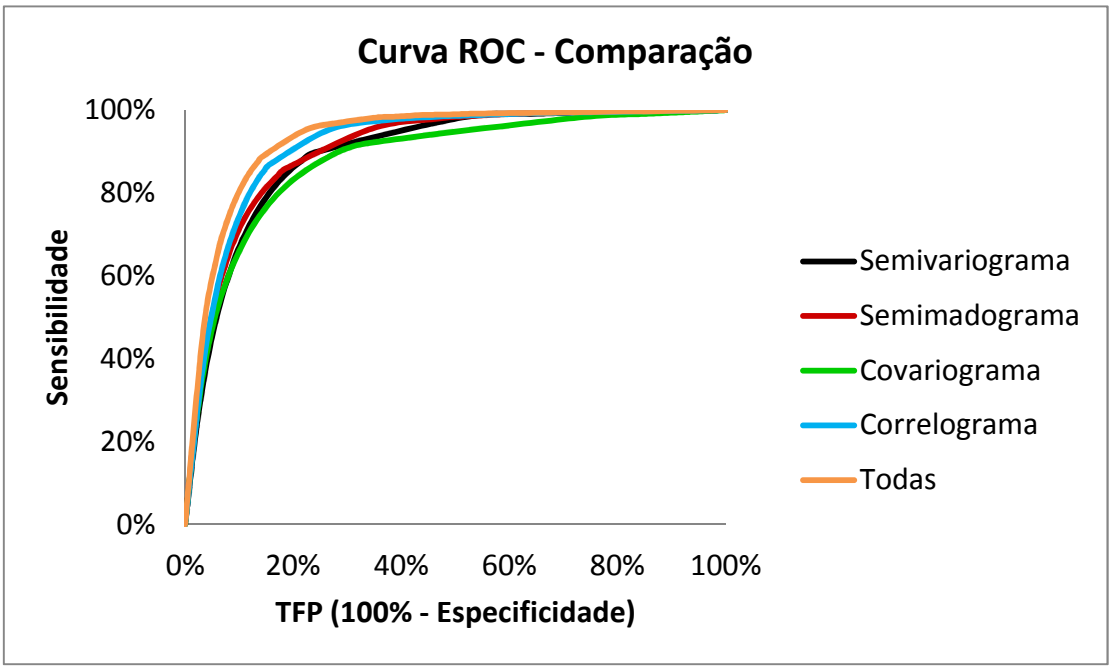


Figura 30 – Comparação dos melhores resultados entre as funções.

Os mesmos resultados apresentados nas curvas ROC da Figura 30 são mostrados na Tabela 24, onde as funções são identificadas pelas siglas: SMV (semivariograma), SMM (semimadograma), COV (covariograma) e COR (correlograma). Pela tabela, confirma-se o melhor desempenho do teste com todas as curvas, observando-se que a sensibilidade cresce rapidamente, quando a taxa de falsos positivos ainda se mantem baixa, mesmo até o último peso testado.

Tabela 24 - Comparação dos melhores resultados de cada função: semivariograma (SMV) com distância máxima 4; semimadograma (SMM), com distância máxima 6; covariograma (COV), com distância máxima 6; correlograma (COR), com distância máxima 5; e todas as funções, com distância máxima 4.

Peso	**SMV**		**SMM**		**COV**		**COR**		**Todas**	
Falhas	**Sens.**	**TFP**	**Sens.**	**TFP**	**Sens.**	**TFP**	**Sens.**	**TFP.**	**Sens.**	**TFP**
1	1,9%	0,2%	1,1%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
10	24,7%	2,2%	32,7%	2,6%	17,7%	1,4%	37,4%	3,2%	46,7%	3,3%
20	43,2%	4,6%	52,7%	5,4%	31,2%	2,7%	57,0%	5,9%	64,9%	5,9%
30	55,9%	7,1%	63,5%	7,6%	42,0%	4,1%	66,7%	7,9%	73,0%	7,7%
40	65,3%	9,5%	70,4%	9,6%	51,1%	5,7%	72,4%	9,4%	77,8%	9,1%
50	72,3%	12,0%	74,8%	11,4%	58,8%	7,7%	76,6%	10,7%	81,0%	10,2%
60	77,8%	14,4%	78,2%	13,1%	66,5%	10,2%	79,7%	11,9%	83,4%	11,2%
70	82,1%	16,9%	80,7%	14,5%	74,5%	13,8%	82,1%	12,9%	85,2%	12,1%
80	85,3%	19,2%	82,6%	15,9%	82,2%	19,1%	84,1%	13,8%	86,5%	12,8%
90	87,8%	21,5%	84,3%	17,2%	88,4%	26,2%	85,5%	14,7%	87,6%	13,5%
100	89,7%	23,8%	85,8%	18,4%	91,9%	33,2%	86,8%	15,5%	88,6%	14,2%
300	98,2%	51,1%	95,5%	34,5%	98,2%	72,8%	95,3%	26,6%	95,2%	22,2%
600	99,1%	65,3%	97,9%	46,5%	99,1%	85,3%	97,6%	37,4%	97,1%	29,1%
900	99,4%	71,2%	98,6%	52,9%	99,4%	90,8%	98,3%	45,3%	98,3%	35,9%
1200	99,5%	74,7%	98,9%	57,3%	99,6%	93,5%	98,7%	51,5%	98,4%	39,0%
1500	99,5%	77,0%	99,1%	60,3%	99,6%	94,1%	98,9%	56,2%	98,8%	43,1%
1800	99,5%	78,8%	99,2%	62,5%	99,8%	97,3%	99,0%	59,7%	98,9%	46,2%
1100	99,6%	80,0%	99,3%	64,4%	99,9%	98,1%	99,2%	62,8%	98,9%	47,7%
1400	99,6%	81,0%	99,4%	66,2%	99,9%	99,3%	99,3%	66,5%	99,0%	50,3%
1700	99,6%	81,8%	99,4%	67,0%	99,9%	99,5%	99,3%	67,6%	99,2%	53,4%
3000	99,6%	82,6%	99,5%	68,6%	100%	99,7%	99,4%	70,4%	99,2%	55,4%

4.2 Extração das Falhas

Esta seção apresenta os resultados visuais da classificação com SVM e da última etapa da metodologia, a qual visa a extração automática das superfícies de falha. Utilizou-se para isso, primeiramente, um dos melhores modelos gerados, obtido com as seguintes configurações:

- Seleção aleatória de 60% da base para treino. Os testes nas 40% amostras restantes resultaram em **92,15% de sensibilidade e 84,33% de especificidade**;
- Funções geoestatísticas semivariograma, semimadograma, covariograma e correlograma, calculadas em janelas de tamanho 7x7x7, todas as direções tridimensionais, e distâncias 1, 2, 3 e 4 entre os voxels;
- Pesos 1 para não falhas e 100 para falhas.

As imagens apresentadas na Figura 31 exemplificam o resultado da classificação para quatro fatias do subvolume V1. Em a) tem-se as fatias originais; em b) o resultado da classificação em uma imagem binária, onde os pixels brancos indicam falhas e os pixels pretos indicam não falhas; e em c) o mesmo resultado, no entanto, explicitando as quantidades de verdadeiros positivos (em azul), falsos positivos (em verde) e falsos negativos (em vermelho) em comparação às marcações manuais utilizadas para a criação da base. Os pixels com os valores originais indicam os verdadeiros negativos.

Da mesma forma, na Figura 32, tem-se os resultados da classificação para três das fatias do subvolume V2. Observe que, no caso deste subvolume, que possui várias pequenas falhas próximas umas às outras, a quantidade de falsos positivos é alta, provocando a união das falhas, o que dificultará a execução das etapas seguintes. Este é um caso, portanto, em que pode ser preferível um modelo gerado com menor sensibilidade em favor da diminuição de falsos positivos, utilizando, como demonstrado na Seção 4.1.6, pesos menores para a classe das falhas.

A Figura 33 e a Figura 34 apresentam respectivamente fatias do subvolume V1 e do subvolume V2 classificadas com um modelo gerado conforme o anterior, alterando somente o peso das falhas de 100 para 40, o que proporcionou **90,02% de especificidade e 81,54% de sensibilidade**. A principal consequência positiva desta alteração são regiões de falhas menos conexas e mais facilmente separáveis. Em contrapartida, acaba-se por perder alguns voxels de falha, gerando buracos.

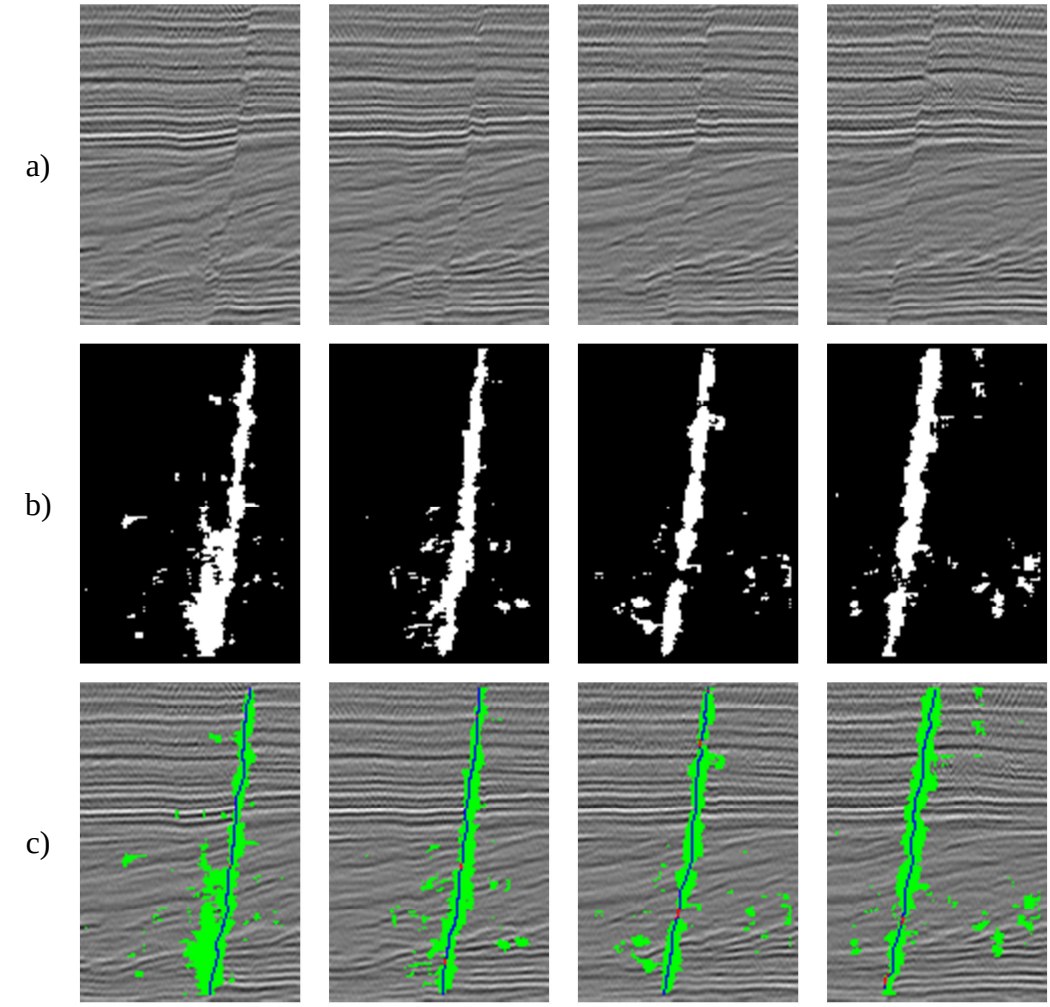


Figura 31 – Fatias classificadas do subvolume V1, com modelo gerado com peso 100 para falhas. a) Fatias originais. b) Imagens binárias. Falhas: voxels brancos, não falhas: voxels pretos. c) Representação dos verdadeiros positivos (em azul), falsos positivos (em verde) e falsos negativos (em vermelho).

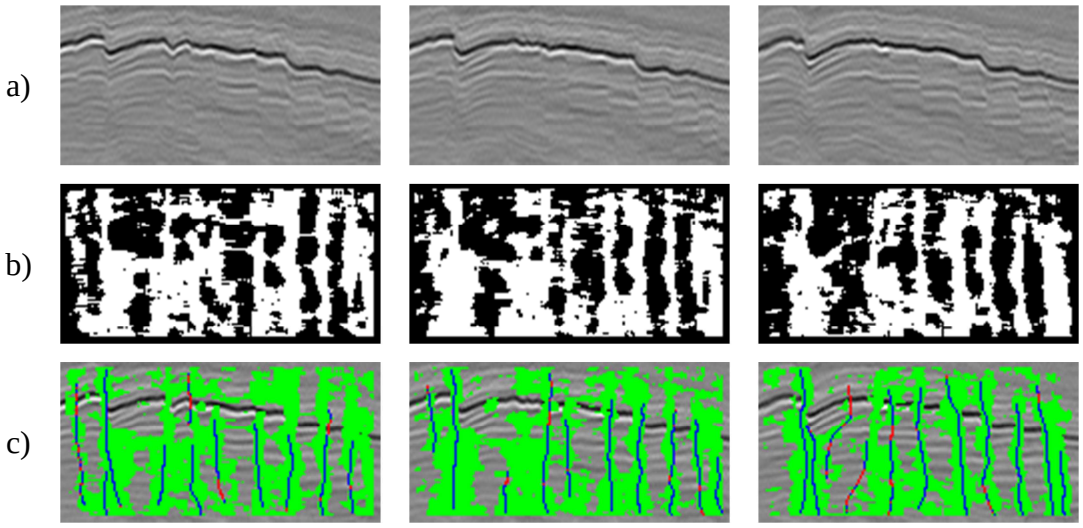


Figura 32 – Fatias classificadas do subvolume V2, com modelo gerado com peso 100 para falhas. a) Fatias originais. b) Imagens binárias. Falhas: voxels brancos, não falhas: voxels pretos. c) Representação dos verdadeiros positivos (em azul), falsos positivos (em verde) e falsos negativos (em vermelho).

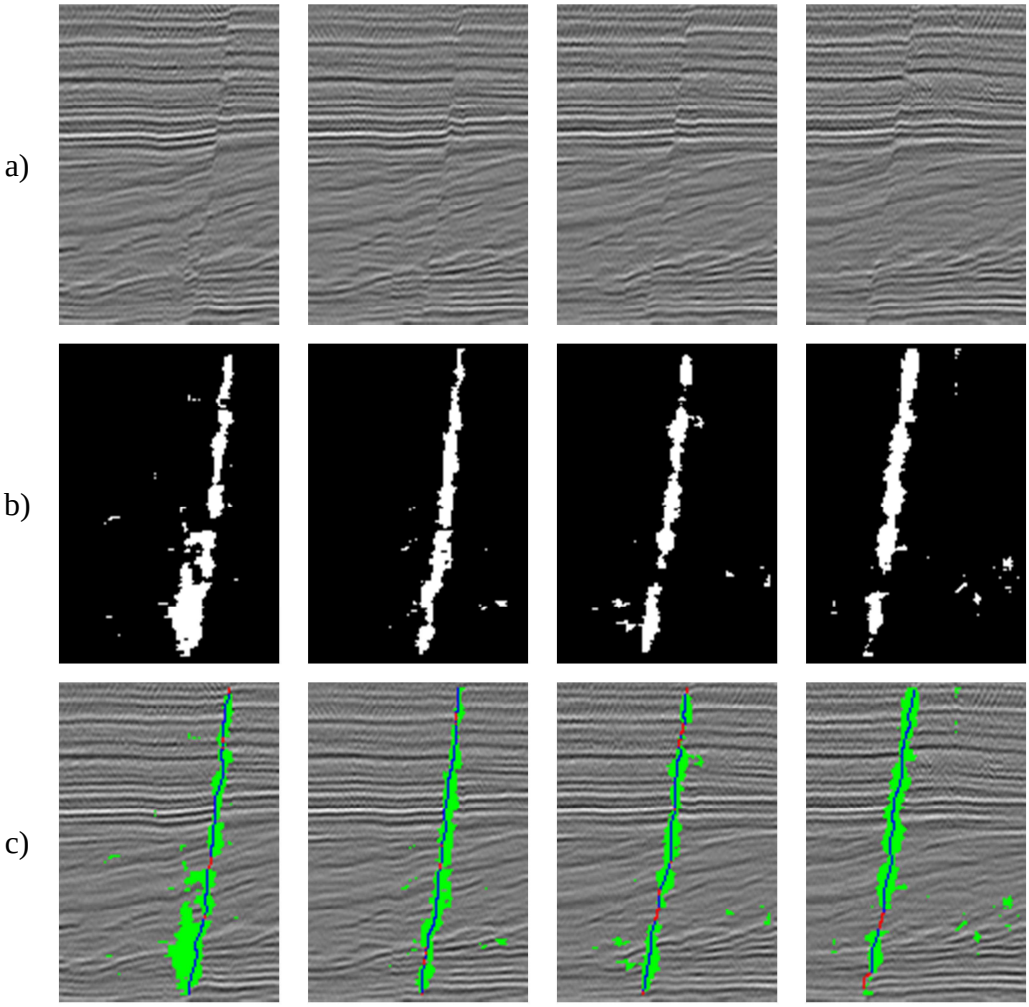


Figura 33 – Fatias classificadas do subvolume V1, com modelo gerado com peso 40 para falhas. a) Fatias originais. b) Imagens binárias. Falhas: voxels brancos, não falhas: voxels pretos. c) Representação dos verdadeiros positivos (em azul), falsos positivos (em verde) e falsos negativos (em vermelho).

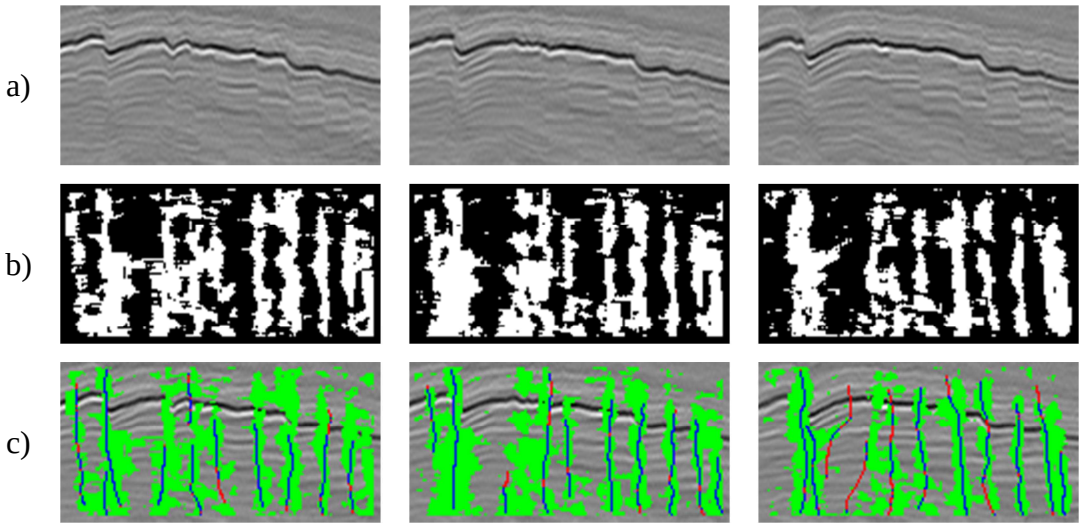
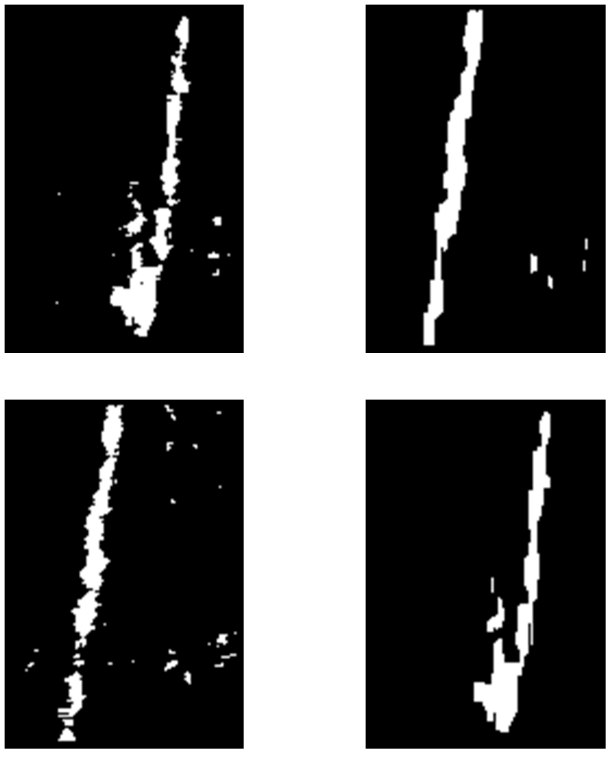


Figura 34 – Fatias classificadas do subvolume V2, com modelo gerado com peso 40 para falhas. a) Fatias originais. b) Imagens binárias. Falhas: voxels brancos, não falhas: voxels pretos. c) Representação dos verdadeiros positivos (em azul), falsos positivos (em verde) e falsos negativos (em vermelho).

Pela Figura 33, pode-se observar que a mudança do modelo não ocasionou diferenças significativas no subvolume V1, a não ser pela diminuição dos ruídos. Na Figura 34, no entanto, é visível a maior separação entre as falhas, e o aumento da quantidade de falsos negativos (linhas vermelhas).

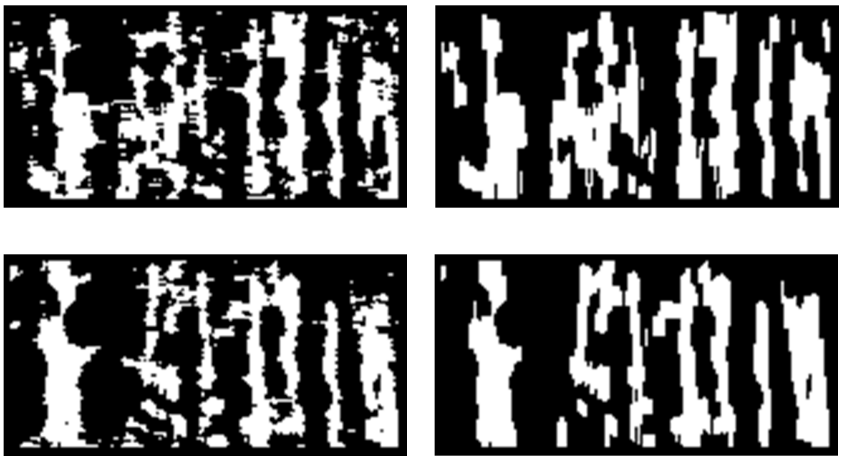
Seguindo a metodologia, na etapa seguinte o volume classificado (binário) será submetido a erosões e dilatações fatia a fatia, com o objetivo de remover ruídos, fechar buracos nas falhas e de desunir falhas diferentes que estejam conectadas, após a classificação. Quanto melhor o desempenho da classificação, menor a importância dessa etapa. Em outras palavras, esta etapa é importante para suprir os erros da classificação com SVM.

A execução de uma abertura seguida de um fechamento utilizando um elemento estruturante retangular de 1 pixel de largura por 5 pixels de altura atingiu bons resultados na maioria dos casos. A Figura 35 apresenta o resultado de duas fatias do subvolume V1 e a Figura 36 do subvolume V2.



(a) (b)

Figura 35 – Operações morfológicas realizadas sobre fatias do subvolume V1. a) Fatias resultantes da classificação e em b) Fatias após uma abertura e um fechamento com elemento estruturante 1x5.



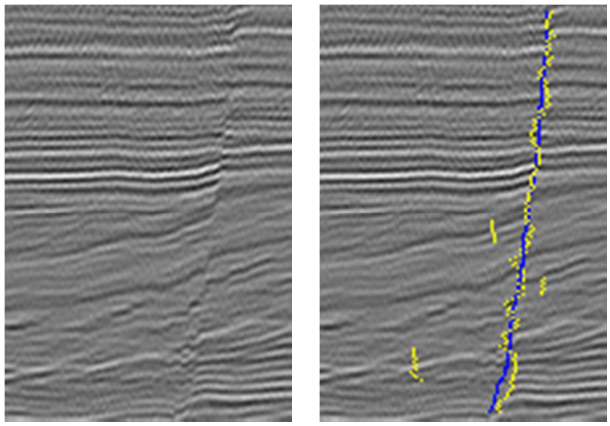
(a) (b)

Figura 36 – Operações morfológicas realizadas sobre fatias do subvolume V2. a) Fatias resultantes da classificação e em b) Fatias após uma abertura e um fechamento com elemento estruturante 1x5.

Nas duas fatias superiores a maior parte dos ruídos (pequenas regiões brancas) foi removida. Além disso, a continuidade da região de falha foi fortalecida devido à forma vertical do elemento estruturante. Nas duas fatias inferiores, é possível perceber que muitas das conexões entre diferentes agrupamentos de falhas foram quebradas.

O crescimento de região é então executado sobre o volume operado morfologicamente de forma tridimensional. Ou seja, serão agrupados todos os voxels conectados espacialmente, para que a última etapa da metodologia, o afinamento descrito na Seção 3.3, seja realizada em cada falha separadamente.

Os resultados finais, com as falhas afinadas sobrepostas de volta aos volumes são apresentados na Figura 37 (subvolume V1) e na Figura 38 (subvolume V2), ao lado das fatias originais correspondentes. As falhas encontradas pela metodologia estão destacadas em amarelo, sobre as falhas marcadas manualmente, em azul.



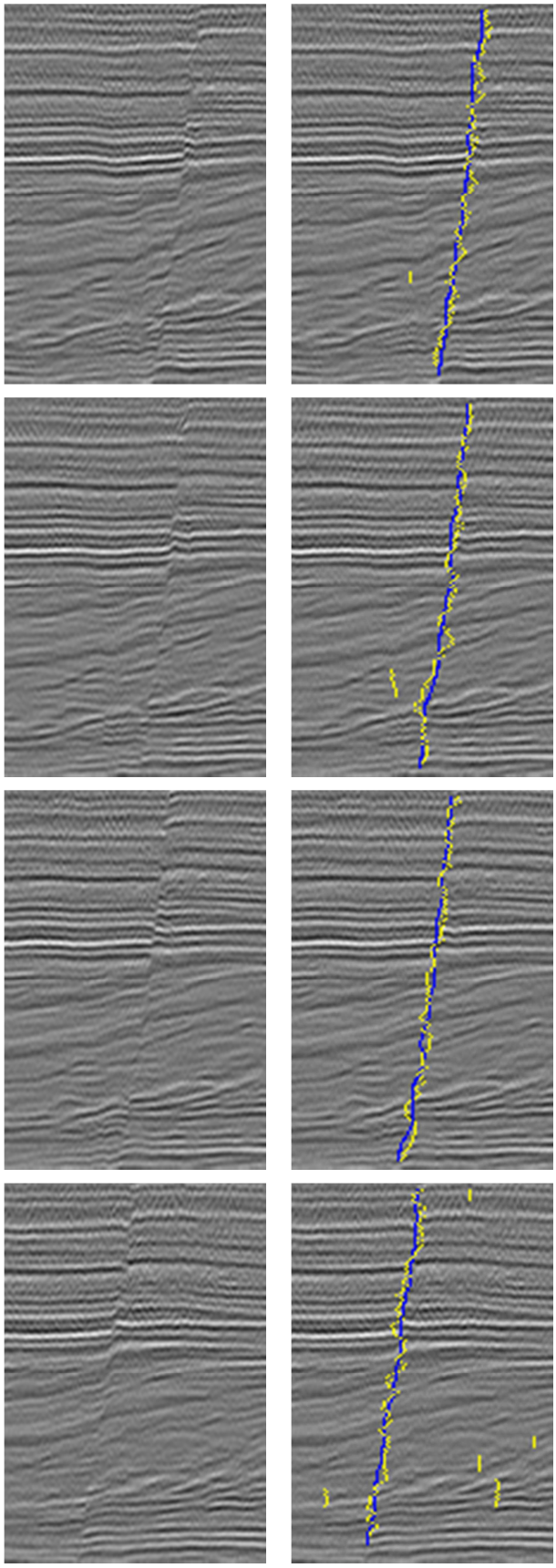


Figura 37 – Resultado final em fatias do subvolume V1. Em amarelo, a falha extraída pelo método proposto, e em azul a marcação manual.

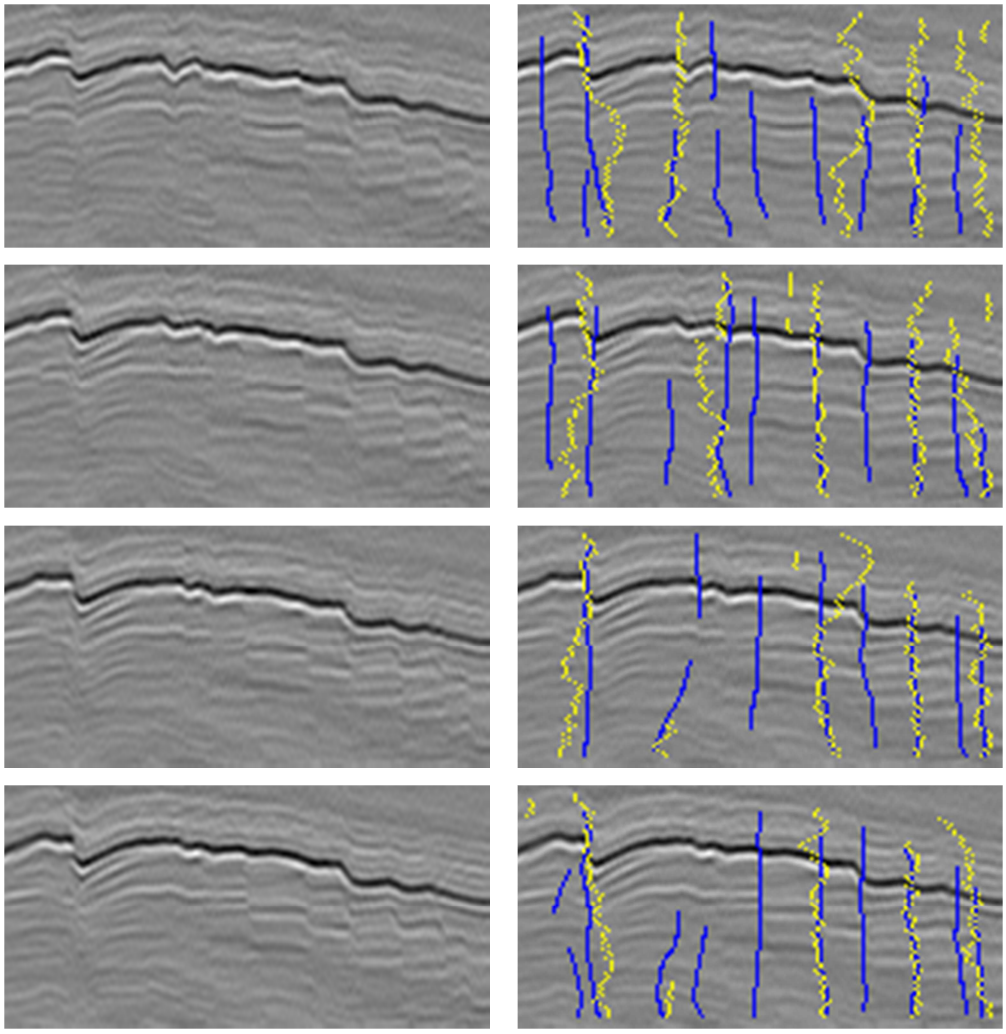


Figura 38 – Resultado final em fatias do subvolume V2. Em amarelo, a falha extraída pelo método proposto, e em azul a marcação manual.

A detecção das falhas no subvolume V1 ocorreu como o esperado. A grande falha que o volume contém foi identificada em todas as fatias, e poucos ruídos permaneceram.

Com relação ao subvolume V2, também poucos ruídos foram mantidos. Entretanto, o problema da proximidade entre as falhas permaneceu até as últimas etapas, fazendo com que algumas falhas fossem perdidas.

A comparação com outros trabalhos se torna uma tarefa difícil, uma vez que nenhum trabalho científico forneceu dados numéricos de acerto, e a própria definição de acerto é subjetiva, já que não existe uma marcação de falhas universal. Além disso, a comparação visual é irrelevante, dado que os resultados binários da classificação não serão comparáveis aos resultados contínuos de um método de realce. Por isso, achou-se mais válida a comparação do método proposto com as marcações manuais, de forma tanto numérica quanto visualmente.

5 CONCLUSÃO

Este trabalho apresentou uma metodologia de detecção de falhas sísmicas baseada na utilização das funções geoestatísticas de semivariograma, semimadograma, covariograma e correlograma como características representativas para uma classificação com Máquinas de Vetores de Suporte, as SVMs.

A metodologia foi desenvolvida e testada utilizando-se subvolumes extraídos do dado F3 Block, disponibilizados pelo sistema OpendTect. Os altos valores de sensibilidade e especificidade obtidos nos testes indicam a sua eficácia. Índices de até 92,15% de sensibilidade e 84,33% de especificidade foram obtidos nas classificações utilizando-se todas as funções simultaneamente. Entretanto, para a etapa de extração de falhas apresentada, é preferível um menor índice de falsos positivos. Assim, com diferentes parâmetros de classificação, obteve-se até 84,33% de sensibilidade com 90,02% de especificidade, também utilizando todas as funções.

A etapa de extração apresentada, por sua vez, se mostrou bastante eficaz em casos de falhas isoladas. No entanto, não foi capaz de separar completamente os casos em que existem falhas muito próximas umas às outras. O aumento da especificidade contribuiu para a diminuição deste problema, mas não foi o suficiente. Sempre existirão falsos positivos, dada a enorme semelhança entre as características calculadas sobre os *voxels* manualmente marcados como falhas e os seus vizinhos imediatos. Sendo assim, devem ser avaliadas formas de separação das regiões agrupadas pelo crescimento de região, separando, nesta etapa, falhas que estejam possivelmente interligadas.

No geral, o trabalho apresentou resultados satisfatórios, que proporcionam muito mais rapidez na interpretação das falhas sísmicas do dado F3 Block. Em alguns minutos de execução, um intérprete obtém as falhas marcadas visualmente, economizando esforço e tempo de trabalho, com resultados próximos ao desejado. Uma ferramenta de interação, por exemplo, poderia permitir ao usuário a correção manual, caso desejado.

Assim, este trabalho forneceu como contribuição principal a aplicação de funções geoestatísticas, já utilizadas em outras áreas de processamento de imagens como características de textura, como atributos de falhas. Além disso, eles são todos utilizados em conjunto dentro de um sistema de reconhecimento de padrões, para uma classificação precisa, utilizando SVM, do que é falha e do que não é falha em um volume sísmico.

Como trabalhos futuros, propõe-se primeiramente o teste dos modelos gerados em outros dados sísmicos reais para avaliar os resultados. Caso estes se mostrem inferiores aos

apresentados neste trabalho, um treinamento pode ser realizado novamente para incluir novas informações fornecidas por especialistas e tornar o método mais generalizável.

Outra proposta é a seleção da base e das características que gerem um modelo ótimo de classificação. Como a seleção das amostras de treinamento foi realizada de forma aleatória, não existe a garantia de que estas amostras sejam as mais representativas de todo o volume. Buscas pela melhor base com o auxílio, por exemplo, de algoritmos genéticos, podem ser realizadas.

Da mesma forma, é possível selecionar dentre todas as características geoestatísticas (variando função, ângulo e distância) aquelas que melhor representam as amostras. Uma menor quantidade de características, inclusive, proporcionará a diminuição do tempo de extração e de classificação. O tempo de execução também pode ser reduzido paralelizando-se essas duas tarefas. A forma que é realizada a extração de características, utilizando janelas, e a classificação dessas janelas representadas pelas funções torna a sua execução fortemente paralelizável.

REFERÊNCIAS

Aqrawi, A., e T. Boe. “Improved Fault Segmentation using a Dip-guided and modified 3D Sobel Filter.” *81st Annual International Meeting.* SEG, Expanded Abstracts, 2011.

Bahorich, M. S., e S. L. Farmer. “3-D Seismic Coherency for Faults and Stratigraphic Features: The Coherence Cube.” *The Leading Edge, 14*, 1995: 1053-1058.

Bemmel, P. P., e R. E. F. Pepper. Seismic Signal Processing and Apparatus for Generating a Cube of Variance Values. U.S. Patente 6151555. 2000.

Ben-Hur, A., e J. Weston. “A User's Guide to Support Vector Machines.” In: *Data Mining Technicques for the Life Sciences,* por O. Carugo e F. Eisenhaber, 223-239. Humana Press, 2010.

Boe, T. H., e R. Daber. “Seismic Features and the Human Eye: RGB Blending of Azimuthal Curvatures for Enhancement of Fault and Fracture Interpretation.” *80th Annual International Meeting, Expanded Abstracts.* SEG, 2010. 1535-1539.

Bradiski, G., e A. Kaehler. *Learning OpenCV.* Sebastopol, CA: O'Reilly Media, 2008.

Chang, C., e C. Lin. “LIBSVM: A Library for Support Vector Machines.” *ACM Transactions on Intelligent Systems and Technology,* Abril de 2011.

Chopra, S., e K. J. Marfurt. “Seismic Attributes for Fault/Facture Characterization.” *CSPG CSEG Convention.* 2007.

—. *Seismic Attributes for Prospect Identification and Reservoir Characterization.* SEG Books, 2007.

Cohen, I., N. Coult, e A. Vassiliou. “Detection and Extraction of Fault Surfaces in 3D Seismic Data.” *Geophysics, 71*, 2006: 21-27.

Duda, R. O., P. E. Hart, and D. G. Stork. *Pattern Classification.* 2ª. New York: Wiley-Interscience, 2001.

Fan, R., K Chang, C. Hsieh, X. Wang, e C. Lin. “LIBLINEAR: A Library for Large Linear Classification.” *Journal of Machine Learning Research 9*, 2008: 1871-1874.

Figueiredo, A. M. “Mapeamento Automático de Horizontes e Falhas em Dados Sísmicos 3D baseado no Algoritmo de Gás Neural Evolutivo.” *Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.* Rio de Janeiro, 2007.

Gibson, D., M. Spann, e J. Turner. “Automatic Fault Detection for 3D Seismic Data.” *Proc. VIIth Digital Image Computing: Techniques and Applications*, 2013: 10-12.

Gonzalez, R. C., e R. C. Woods. *Processamento Digital de Imagens.* 3ª Edição. São Paulo: Pearson Prentice Hall, 2010.

Isaaks, E., e R. M. Srivastava. *An Introduction to Applied Geostatistics.* New York: Oxford University Press, 1989.

Iske, A., e T. Randen. "Mathematical Methods and Modeling in Hydrocarbon Exploration and Production." *Springer Verlag*, 2005.

Kadlec, B. Visualization of Geologic Feature using Data Representations thereof: US Patent Application. Patente 0.115.787. 2011.

Lines, L. R., e R. T. Newrick. *Fundamentals of Geophysical Interpretation.* Society of Exploration Geophysicists, 2004.

Machado, M. C. "Determinação de Malhas de Falhas em Dados Sísmicos por Aprendizado Competitivo." *Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.* Rio de Janeiro, 2008.

Maciel, W. A. G. L. "Um Estudo sobre o Realce de Atributos de Falha em Dados Sísmicos baseado em Modelos de Colônias de Formiga." *Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.* Rio de Janeiro, Junho de 2014.

Müller, K., S. Mika, G. Rätsch, K. Tsuda, e B. Schölkopf. "An Introduction to Kernel-Based Learning Algorithms." *IEEE Transactions on Neural Networks 12(2)*, Março 2001: 181-201.

Pampanelli, P., P. M. Silva, e M. Gattass. "A New Volumetric Fault Attribute based on First Order Directional Derivative." *13th International Congress of The Brazilian Geophysical Society.* SBGf, 2013.

Pedersen, S., T. Randen, L. Sonneland, e O. Steen. "Automatic 3D Fault Interpretation by Artificial Ants." *72nd Annual International Meeting.* SEG, Expanded Abstracts, 2002.

Roberts, A. "Curvature Attributes and their Application to 3D Interpreted Horizons." *First Break, 19*, 2001: 85-100.

Robinson, E. A., e Sven Treitel. *Geophysical Signal Analysis.* Englewood Cliffs: Prentice-Hall, 1980.

Silva, P. M. "Visualização Volumétrica de Horizontes em Dados Sísmicos 3D." *Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.* Rio de Janeiro, 2004.

Smola, A. J., P. L. Bartlett, B. Schölkopf, e D. Schuurmans. *Advances in Large Margin Classifiers.* Cambridge: MIT Press, 2000.