Pontifícia Universidade Católica
do Rio de Janeiro

Eric da Silva Praxedes

Interpretação Sísmica Quantitativa com uso de Programação Genética

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Marco Aurélio Cavalcanti Pacheco

Rio de Janeiro
Agosto de 2014

Pontifícia Universidade Católica
do Rio de Janeiro

Eric da Silva Praxedes

Interpretação Sísmica Quantitativa com uso de Programação Genética

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Marco Aurélio Cavalcanti Pacheco
Orientador
Departamento de Engenharia Elétrica - PUC-Rio

Prof. Douglas Mota Dias
Departamento de Engenharia Elétrica - PUC-Rio

Prof. Ítalo de Oliveira Matias
Universidade Candido Mendes

Prof. Guilherme Fernandes Vasquez
CENPES - PETROBRAS

Prof. André Bulcão
CENPES - PETROBRAS

Prof. José Eugênio Leal
Coordenador Setorial do Centro Técnico Científico

Rio de Janeiro, 21 de agosto de 2014

Eric da Silva Praxedes

Graduou-se em Ciência da Computação na Universidade Federal do Rio de Janeiro em 2003. Cursou o MBA em Gerenciamento de Projetos da Fundação Getúlio Vargas concluído em 2009.

Ficha Catalográfica

Praxedes, Eric da Silva

Interpretação sísmica quantitativa com uso de programação genética / Eric da Silva Praxedes ; orientador: Marco Aurélio Cavalcanti Pacheco.– 2014.
82 f. : il. (color.) ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2014.
Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Atributos elásticos. 3. Litologia. 4. Programação genética. 5. Classificação. I. Pacheco, Marco Aurélio Cavalcanti. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Agradecimentos

Agradeço a Deus por ter me dado saúde e determinação para a realização deste trabalho.

Agradeço ao meu orientador Professor Marco Aurélio Cavalcanti Pacheco pelo estímulo e parceria para a realização deste trabalho.

À Petrobras, na pessoa do meu gerente Sebastião Cesar Assis Pereira, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos membros da banca examinadora pelas valorosas observações, críticas e sugestões proferidas em relação a este trabalho.

Aos colegas e professores do Laboratório de Inteligência Computacional Aplicada do Departamento de Engenharia Elétrica da PUC-Rio.

Aos colegas das gerências de Tecnologia Geológica (E&P-EXP/GEO/TGEO) e de Estudos Geofísicos Petrosísmica e Inversão (E&P-EXP/GEOF/EGPI).

À geofísica Elita Selmara de Abreu pelos ensinamentos, apoio e paciência dispensados durante a realização deste trabalho.

Aos professores Eugênio Silva e Douglas Mota Dias pela grande ajuda dispensada.

À minha querida Letícia pelo apoio, estímulo e companheirismo, principalmente nas horas mais difíceis.

A todos os amigos e familiares que de alguma forma me estimularam ou me ajudaram.

Resumo

Praxedes, Eric da Silva; Pacheco, Marco Aurélio Cavalcanti. **Interpretação Sísmica Quantitativa com uso de Programação Genética.** Rio de Janeiro, 2014. 82p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Uma das tarefas mais importantes na indústria de exploração e produção de petróleo é a discriminação litológica. Uma das principais fontes de informação para subsidiar a discriminação e caracterização litológica é a perfilagem que é corrida no poço. Porém, na grande maioria dos trabalhos os perfis utilizados na discriminação litológica são apenas aqueles disponíveis no domínio dos poços. Para que modelos de discriminação litológica possam ser extrapolados para além do domínio dos poços, faz-se necessário a utilização de características que estejam presentes tanto nos poços como fora deles. As características mais utilizadas para realizar esta integração rocha-perfil-sísmica são os atributos elásticos. Dentre os atributos elásticos o que mais se destaca é a impedância. O objetivo desta dissertação foi a utilização da programação genética como modelo classificador de atributos elásticos para a discriminação litológica. A proposta se justifica pela característica da programação genética de seleção e construção automática dos atributos ou características utilizadas. Além disso, a programação genética permite a interpretação do classificador, uma vez que é possível customizar o formalismo de representação. Esta classificação foi empregada como parte integrante do fluxo de trabalho estatístico e de física de rochas, metodologia híbrida que integra os conceitos da física de rochas com técnicas de classificação. Os resultados alcançados demonstram que a programação genética atingiu taxas de acertos comparáveis e em alguns casos superiores a outros métodos tradicionais de classificação. Estes resultados foram melhorados com a utilização da técnica de substituição de fluídos de Gassmann da física de rochas.

Palavras-chave

Atributos elásticos; litologia; programação genética; classificação.

Abstract

Praxedes, Eric da Silva; Pacheco, Marco Aurélio Cavalcanti (Advisor). **Quantitative Seismic Interpretation using Genetic Programming.** Rio de Janeiro, 2014. 82p. MSc. Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

One of the most important tasks in the oil exploration and production industry is the lithological discrimination. A major source of information to support discrimination and lithological characterization is the logging raced into the well. However, in most studies the logs used in the lithological discrimination are only those available in the wells. For extrapolating the lithology discrimination models beyond the wells, it is necessary to use features that are present both inside and outside wells. One of the features used to conduct this rock-log-seismic integration are the elastic attributes. The impedance is the elastic attribute that most stands out. The objective of this work was the utilization of genetic programming as a classifier model of elastic attributes for lithological discrimination. The proposal is justified by the characteristic of genetic programming for automatic selection and construction of features. Furthermore, genetic programming allows the interpretation of the classifier once it is possible to customize the representation formalism. This classification was used as part of the statistical rock physics workflow, a hybrid methodology that integrates rock physics concepts with classification techniques. The results achieved demonstrate that genetic programming reached comparable hit rate and in some cases superior to other traditional methods of classification. These results have been improved with the use of Gassmann fluid substitution technique from rock physics.

Keywords

Elastic attributes; lithology; genetic programming; classification.

Sumário

Lista de figuras

Lista de tabelas

1
Introdução

Uma das tarefas mais importantes na indústria de exploração e produção de petróleo é a discriminação litológica. Litologia é a descrição das características físicas macroscópicas de uma rocha tais como cor, textura, tamanho do grão e conteúdo mineral (Schlumberger, 2014; USGS, 2014). Com base nessa descrição, e conhecendo-se a localização de cada tipo de rocha no poço, é possível inferir onde se encontram as formações geradoras, de contenção do hidrocarboneto e principalmente o reservatório, necessários para a ocorrência de um sistema petrolífero.

Para esta tarefa existem diversas fontes de informação: amostras de calha, a descrição petrográfica de amostras laterais, testemunhos, dados correlacionados de outros poços, dentre outros. Além dessas, uma das principais fontes de informação para subsidiar a discriminação e caracterização litológica é a perfilagem que é corrida no poço. A perfilagem consiste em se fazer registros detalhados das formações geológicas através de medidas físicas colhidas por ferramentas que são baixadas dentro do poço. Devido às ferramentas de perfilagem medirem propriedades das rochas no subsolo, seus registros são intrinsecamente geológicos (Doveton, 2002).

Tradicionalmente, técnicas estatísticas são utilizadas para subsidiar a discriminação litológica através do estudo dos perfis. Uma das mais utilizadas é a análise discriminante (Busch *et al.*, 1987). Técnicas de inteligência computacional também vêm sendo utilizadas com relativo sucesso. Trabalhos precursores já defendiam que as redes neurais poderiam ser utilizadas para a análise de perfis de poços com o objetivo de realizar a discriminação litológica (Baldwin *et al*., 1989; Doveton, 1994; Rogers *et al*., 1992). Outros trabalhos utilizaram comitês de redes neurais com o objetivo de incrementar a taxa de litologias corretamente discriminadas (Leite, 2012; Santos *et al*., 2003). A lógica fuzzy também já foi utilizada como ferramenta para a discriminação litológica (Saggaf & Nebrija, 2003).

Porém, na grande maioria dos trabalhos os perfis utilizados na discriminação litológica são apenas aqueles disponíveis no domínio dos poços. Dentre eles podemos citar os de raios gama, porosidade neutrônica e resistividade. Ou seja, apesar desses perfis medirem parâmetros diretamente relacionados às propriedades petrofísicas de interesse das rochas, não há como aplicar os modelos produzidos fora dos poços. Fora dos poços, os únicos registros disponíveis são os de subsuperfície que compreendem volumes de atributos derivados dos levantamentos sísmicos. A aplicação desses modelos no domínio dos volumes sísmicos é importante, pois possibilitaria realizar a discriminação litológica entre os poços.

Para que modelos de discriminação litológia possam ser extrapolados para além do domínio dos poços, faz-se necessário a utilização de características que estejam presentes tanto nos poços como fora deles. Nos poços, onde a litologia é conhecida através dos perfis convencionais e de outras informações, pode-se determinar modelos, baseados em características que podem ser obtidas através da sísmica. Uma das características mais utilizadas para realizar esta integração rocha-perfil-sísmica são os atributos elásticos. Os atributos elásticos são aqueles relacionados à teoria da elasticidade, disciplina que relaciona as forças aplicadas a um corpo com as mudanças resultantes em seu tamanho e formato (Villaça & Garcia, 2000).

Dentre os atributos elásticos os que mais se destacam são as impedâncias compressional e cisalhante. Nos poços é possível obter perfis das impedâncias através dos registros de densidade e do tempo de trânsito das ondas compressional e cisalhante, conhecido como perfil sônico. Fora do poço, é possível produzir volumes com valores das impedâncias através do processo denominado inversão sísmica, que utiliza a equação de Aki-Richards e os *angle-gathers* para obter tais volumes (Avseth *et al.*, 2005). Utilizando-se as impedâncias compressional e cisalhante, é possível criar um modelo com dados de poços para discriminar litologias e extrapolá-lo para os dados de sísmica.

Dentre as metodologias comumente utilizadas pelo intérprete para a criação desses modelos, destacam-se aquelas baseadas na teoria de física de rochas ou em métodos estatísticos. A teoria de física de rochas é a disciplina que relaciona os atributos elásticos com as propriedades petrofísicas de reservatórios, como a litologia. As relações utilizadas na física de rochas para traduzir atributos elásticos

em propriedades petrofísicas frequentemente são baseadas em equações empíricas e para um ambiente geológico específico. Os métodos estatísticos, como por exemplo, as técnicas de regressão, agrupamento e as técnicas de reconhecimento de padrões, tentam estabelecer uma relação heurística entre os atributos elásticos e os valores previstos das propriedades petrofísicas somente através da análise dos dados.

Segundo alguns autores (Avseth *et al.*, 2005; Nikravesh *et al.*, 2003), a utilização de metodologias híbridas, combinando a força de técnicas estatísticas e de física de rochas, implicaria na diminuição da incerteza, elevando a confiabilidade das propriedades previstas, contribuindo de forma efetiva no aprimoramento da interpretação de reservatórios.

No entanto, a utilização de modelos mais complexos é por vezes evitada pelo intérprete devido à impossibilidade de interpretar qual regra, equação ou heurística está sendo utilizada para classificar ou extrapolar os atributos elásticos. Esse requisito é fundamental, pois permite avaliar se restrições geológicas e físicas estão sendo respeitadas. A utilização de vários atributos simultaneamente em modelos mais complexos, também levanta questões pelo intérprete, tais como o critério a ser utilizado para a seleção dos atributos e como a previsão será impactada se mais atributos forem adicionados.

1.1. Objetivos

O objetivo deste trabalho é a utilização da programação genética como modelo classificador de atributos elásticos para a discriminação litológica. Esta classificação será empregada como parte integrante do fluxo de trabalho estatístico e de física de rochas (Avseth *et al.*, 2005). O princípio do uso da programação genética é permitir a interpretação do classificador uma vez que é possível extrair o modelo evoluído, além de sua habilidade de seleção automática de atributos. Assim, busca-se um modelo com boa acurácia, aprendizado automático e que proporcione apoio à decisão como também extração e aquisição de conhecimento.

1.2. Organização da Dissertação

No capítulo 2 será apresentado o embasamento teórico que subsidia a discriminação litológica através da interpretação de atributos elásticos. Primeiramente, serão apresentados alguns conceitos da teoria da elasticidade. Em seguida serão apresentadas duas áreas da teoria de física de rochas: a análise de sensibilidade aos fluidos e as relações quantitativas entre a velocidade compressional e cisalhante. Por fim, serão apresentadas duas abordagens utilizadas para a discriminação litológica através de atributos elásticos: a abordagem frequentemente utilizada pelo intérprete e o fluxo de trabalho estatístico e de física de rochas (Avseth *et al.*, 2005).

No capítulo 3 será apresentada uma revisão dos conceitos relacionados à programação genética. Em seguida, será apresentada uma revisão bibliográfica de trabalhos que utilizam a programação genética para o problema de classificação de mais de duas classes. Alguns trabalhos com a comparação entre as principais abordagens utilizadas, decomposição binária e seleção por faixas, também serão apresentados neste capítulo.

No capítulo 4 será apresentado um estudo de caso abordando a classificação de atributos elásticos em poços com o objetivo de discriminar a litologia. Primeiramente, serão apresentados o escopo dos experimentos e a contextualização dos dados utilizados. Em seguida, serão apresentadas as técnicas de classificação empregadas baseadas em programação genética, além das configurações e parâmetros utilizados. Estas técnicas serão utilizadas para classificar a base de dados original, assim como, a base de dados estendida através da técnica de substituição de fluidos de Gassmann.

Por fim, no capítulo 5 serão apresentadas as conclusões finais e algumas sugestões de trabalhos futuros.

2
Interpretação de Atributos Elásticos

Neste capítulo será apresentado o embasamento teórico que subsidia a discriminação litológica através da interpretação dos atributos elásticos. Primeiramente, será revista a teoria da elasticidade, disciplina que relaciona as forças aplicadas a um corpo com as mudanças resultantes em seu tamanho e formato. Em seguida, será discutida a teoria de física de rochas que relaciona os atributos elásticos com as propriedades geológicas de reservatórios. Serão abordadas duas importantes áreas desta teoria: a análise de sensibilidade aos fluidos dos atributos elásticos e as relações quantitativas entre a onda P e a onda S para a discriminação litológica.

Por último, serão apresentadas duas abordagens utilizadas para a discriminação litológica através de atributos elásticos. A primeira é a abordagem frequentemente utilizada pelo intérprete que tem como base a interpretação gráfica de dados e o uso da análise discriminante. A segunda abordagem é o fluxo de trabalho estatístico e de física de rochas (Avseth *et al.*, 2005).

2.1.
Teoria da Elasticidade

O tamanho e o formato de um corpo sólido podem ser modificados pela aplicação de forças externas em sua superfície. A estas forças externas são opostas forças internas que resistem às mudanças em seu tamanho e em seu formato. Como resultado, o corpo tende a retornar para a sua condição original assim que essas forças externas são removidas. De forma análoga, um fluido também resiste às mudanças em seu tamanho (volume), porém neste caso não há resistência em relação ao seu formato. Esta propriedade de resistir às mudanças no tamanho e no formato e retornar a condição de não deformação é chamada de elasticidade (Telford *et al.*, 1990).

As relações entre essas forças e as mudanças na forma dos corpos são expressas por dois conceitos: tensão (σ) e deformação (ε).

Por definição, tensão é a razão da força por unidade de área. Quando uma força é aplicada a um corpo, a tensão é a razão da força pela área a qual a força está sendo aplicada. Se a força é perpendicular à área, ou seja, atua paralelamente a um dos eixos $(x, y\ ou\ z)$ a tensão é dita normal ou pressão $(\sigma_{xx}, \sigma_{yy}\ ou\ \sigma_{zz})$. Quando a tensão é tangencial ao elemento de área, ela é denominada tensão de cisalhamento, atuando em mais de um eixo $(\sigma_{xy},\ \sigma_{xz}\ ou\ \sigma_{yz})$.

Quando um corpo elástico é submetido a tensões, ocorrem mudanças no tamanho e no formato deste corpo que são denominadas deformações. A deformação é definida como uma mudança relativa, isto é, fracionada em cada dimensão de um corpo. Quando as variações de comprimento se dão em apenas uma dimensão $(\varepsilon_{xx},\ \varepsilon_{yy}\ ou\ \varepsilon_{zz})$, elas são chamadas de deformações normais. Quando as variações de comprimento se dão em mais de uma dimensão $(\varepsilon_{xy} = \varepsilon_{yx}, \varepsilon_{yz} = \varepsilon_{zy}\ ou\ \varepsilon_{zx} = \varepsilon_{xz})$, temos uma mudança na forma do corpo e essa deformação é chamada de cisalhante.

As mudanças de dimensões devido às tensões normais resultam em mudanças volumétricas. Essas alterações volumétricas relativas à unidade de volume são denominadas de dilatação (Δ), sendo expressa pela equação (2.1).

$$\Delta = \varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz} \tag{2.1}$$

Para se calcular a deformação em função da tensão é necessário saber a relação entre esses dois conceitos. Quando a deformação é pequena, esta relação é dada pela lei de Hooke (Villaça & Garcia, 2000) que atesta que a tensão aplicada a um corpo é diretamente proporcional à deformação sofrida por este. Em geral a lei de Hooke envolve relações muito elaboradas, porém quando o meio é homogêneo e isotrópico, ela pode ser expressa pelas equações (2.2) e (2.3).

$$\sigma_{ii} = \lambda \cdot \Delta + 2 \cdot \mu \cdot \varepsilon_{ii} \qquad i = x, y, z \tag{2.2}$$

$$\sigma_{ij} = \mu \cdot \varepsilon_{ij} \qquad i, j = x, y, z, i \neq j \tag{2.3}$$

Os valores λ e μ são conhecidos como os parâmetros de Lamé. O primeiro valor (λ) é conhecido como primeiro parâmetro de Lamé, enquanto o segundo (μ) é chamado de módulo cisalhante, uma medida da resistência à tensão de cisalhamento.

Constantes elásticas, como os parâmetros de Lamé, são valores que caracterizam um determinado material em função da deformação sofrida por este quando uma tensão é aplicada. Além das constantes de Lamé, a Lei de Hooke pode ser expressa em função de outras constantes elásticas que também caracterizam o corpo, tais como o módulo de Young (E), a razão de Poisson (ν) e o módulo bulk (K). O módulo de Young é definido como a razão entre a tensão e a deformação ao longo do mesmo eixo. A razão de Poisson é a relação entre as deformações paralelas à tensão aplicada. O módulo bulk é definido como a razão entre a tensão normal e a dilatação. As equações (2.4), (2.5) e (2.6) apresentam as três constantes elásticas descritas acima, tanto em função das tensões e deformações quanto das constantes de Lamé.

$$E = \frac{\sigma_{xx}}{\varepsilon_{xx}} = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu} \tag{2.4}$$

$$\nu = -\frac{\varepsilon_{yy}}{\varepsilon_{xx}} = -\frac{\varepsilon_{zz}}{\varepsilon_{xx}} = \frac{\lambda}{2(\lambda + \mu)} \tag{2.5}$$

$$K = -\frac{\sigma_{xx}}{\Delta} = \frac{3\lambda + 2\mu}{3} \tag{2.6}$$

As constantes elásticas são definidas de tal maneira que elas sempre são números positivos, daí o uso do termo “módulo” para denomina-las. Os sinais negativos na definição da razão de Poisson e do módulo bulk também se prestam a esta finalidade. A razão de Poisson possui valores teóricos entre -1 e 0,5. No entanto, poucos são os materiais que apresentam razão de Poisson negativas. Sendo assim, para fins práticos, adota-se que os limites da razão de Poisson estão entre 0 e 0,5 variando entre 0,05 para rochas muito rígidas e 0,45 para materiais pouco consolidados. Os fluidos possuem uma razão de Poisson de 0,5, uma vez que não possuem resistência ao cisalhamento $(\mu = 0)$.

O desequilíbrio entre as tensões aplicadas a um corpo leva à geração de dois tipos de onda. A primeira corresponde às mudanças de dilatação (Δ) e a segunda é responsável pelas rotações em um dos eixos do corpo. Estas ondas, que viajam no interior de um corpo, são chamadas ondas de corpo (*bodywaves*). O primeiro tipo de onda é conhecido como ondas longitudinais, compressionais ou ondas P. O segundo tipo é conhecido como ondas transversais, de cisalhamento ou ondas S[1].

[1] O “P” da onda compressional é devido ao fato desse tipo de onda ser o primeiro (*primary*) evento registrado em um terremoto, enquanto o “S” da onda de cisalhamento deve-se ao fato de ser o segundo (*second*) evento registrado.

As velocidades das ondas $P\ (VP)$ e $S\ (VS)$ são apresentadas nas equações (2.7) e (2.8).

$$VP = \sqrt{\frac{\lambda + 2\mu}{\rho}} = \sqrt{\frac{\mathrm{K} + \frac{4}{3}\mu}{\rho}} \tag{2.7}$$

$$VS = \sqrt{\frac{\mu}{\rho}} \tag{2.8}$$

As equações (2.7) e (2.8) mostram que VP e VS, em um sólido homogêneo e isotrópico, são funções apenas das constantes elásticas $(K\ e\ \mu)$ e da densidade (ρ). Outro atributo elástico relacionado às velocidades VP e VS é a impedância. A impedância de um meio elástico é a razão entre a tensão e a velocidade da partícula (Mavko *et al.*, 1998). A impedância compressional (IP) pode ser expressa pelo produto da densidade e VP (equação (2.9)), enquanto a impedância cisalhante (IS) pode ser expressa pelo produto da densidade e VS (equação (2.10)).

$$IP = \rho\ .VP \tag{2.9}$$

$$IS = \rho\ .VS \tag{2.10}$$

As rochas sedimentares, onde são encontrados os acúmulos de hidrocarbonetos, possuem uma estrutura granular com espaço entre os grãos. Estes espaços são responsáveis pela porosidade das rochas, sendo este um importante fator para determinar a velocidade. Em um reservatório, o espaço poroso é preenchido com um fluido cujas constantes elásticas e densidade também afetam as velocidades. O óleo é ligeiramente mais compressível do que a água, portanto poros preenchidos com óleo resultam em velocidades ligeiramente menores do que poros preenchidos com água. O gás é consideravelmente mais compressível do que a água, portanto poros preenchidos com gás geralmente resultam em velocidades bem menores. Estes efeitos são utilizados como indicadores de hidrocarbonetos.

Na próxima seção serão apresentadas em maiores detalhes as relações entre os atributos elásticos e as propriedades geológicas de reservatórios. Este estudo é o objeto de uma área da geofísica denominada física de rochas, sendo de grande utilização para a discriminação litológica.

2.2. Física de Rochas

A física de rochas fornece as conexões entre os atributos elásticos medidos a partir da superfície da Terra, de dentro dos poços ou em laboratório com as propriedades geológicas das rochas como a mineralogia, a porosidade, o tipo dos fluidos, a pressão de poros, a permeabilidade, a viscosidade, dentre outras (Avseth *et al.*, 2005). Ela fornece o entendimento e as ferramentas teóricas e empíricas para aperfeiçoar a caracterização de reservatórios baseada em dados elásticos.

Um dos usos mais importantes da física de rochas é a possibilidade de extrapolação de seus resultados para um conjunto de dados em particular. No poço, assumindo-se que a qualidade dos dados é boa, as propriedades geológicas das rochas são muito bem conhecidas. Testemunhos, amostras laterais e os perfis de poços trazem informações valiosas sobre a litologia, a porosidade, a permeabilidade e os tipos de fluidos nos poros. No entanto, estas são medidas locais que muitas vezes não representam as características principais da área de estudo. Através de estudos de física de rochas é possível extrapolar os dados de poços para condições geológicas plausíveis que podem existir para além deste, bem como modelar diferentes cenários e explorar como os atributos elásticos se alteram de acordo com o cenário previsto. Estes estudos são particularmente úteis quando se deseja entender os padrões de atributos elásticos de fluidos e fácies[2] que não estão representados nos poços.

A sensibilidade das velocidades VP e VS aos parâmetros de reservatórios já é conhecida há muitos anos. Com o enorme incremento da aquisição, do processamento sísmico e a necessidade de interpretar os atributos elásticos para a detecção de hidrocarbonetos, identificação e caracterização litológica e o monitoramento de reservatórios, houve a necessidade prática de quantificar a relação entre os atributos elásticos com as propriedades das rochas. Descobrir e interpretar as relações entre os atributos elásticos e as propriedades de reservatório tem sido o foco da pesquisa na área de física de rochas ao longo de décadas.

Serão apresentadas nas seções seguintes, duas importantes áreas da física de rochas que são muito utilizadas para a discriminação litológica. A primeira é a

[2]O termo "fácies" originalmente significava a mudança lateral no aspecto litológico de uma unidade estratigráfica. Contudo, seu significado foi ampliado para expressar uma ampla gama de conceitos geológicos: ambiente de deposição, composição litológica, associação geográfica, climática ou tectônica, dentre outros (ICS, 2014). O termo é sempre usado no plural.

análise de sensibilidade dos atributos elásticos para diferentes fluidos. A segunda é a área que estuda as relações entre VP e VS e outros atributos derivados dos mesmos, como as impedâncias IP e IS.

2.2.1. Análise de sensibilidade aos fluidos

A análise de sensibilidade aos fluidos é a parte da física de rocha que estuda e tenta prever as velocidades sísmicas de uma rocha saturada com um fluido, a partir de rochas saturadas com outro fluido (Mavko *et al.*, 1998). Ou, de forma equivalente, prever as velocidades sísmicas de rochas saturadas com um fluido a partir de velocidades com a rocha seca. Neste tipo de problema, são utilizadas as relações da equação de Gassmann (1951)-Biot (1956), que prevê como o módulo bulk da rocha (K_{sat}) se modifica com as mudanças nos fluidos dos poros.

$$\frac{K_{sat}}{K_{mineral} - K_{sat}} = \frac{K_{rocha}}{K_{mineral} - K_{rocha}} + \frac{K_{fluido}}{\phi\left(K_{mineral} - K_{fluido}\right)} \qquad (2.11)$$

Na equação de Gassmann (2.11), K_{rocha}, $K_{mineral}$ e K_{fluido} são respectivamente os módulos bulk da rocha seca, do mineral que compõe a rocha e do fluido. K_{sat} é o módulo bulk da rocha saturada com o fluido e ϕ é a porosidade.

Embora a equação de Gassmann seja originalmente descrita no sentido de prever o módulo bulk da rocha saturada a partir da rocha seca, ela é mais utilizada para prever o módulo bulk da rocha saturada com um fluido a partir de outro fluido. O procedimento envolve a aplicação da equação duas vezes: a primeira calcula o módulo bulk da rocha seca a partir do módulo bulk da rocha saturada com o fluido 1; em seguida o módulo bulk da rocha saturada com o fluido 2 é calculado a partir do módulo bulk da rocha seca. A equação (2.11) é reescrita na equação (2.12), eliminando-se algebricamente K_{rocha} e colocando os módulos bulk da rocha saturada com os fluidos 1 e 2 $(K_{sat1}\ e\ K_{sat2})$ em função dos módulos bulk dos mesmos fluidos $\left(K_{fluido1}\ e\ K_{fluido2}\right)$.

$$\frac{K_{sat2}}{K_{mineral} - K_{sat2}} - \frac{K_{fluido2}}{\phi\left(K_{mineral} - K_{fluido2}\right)} = \frac{K_{sat1}}{K_{mineral} - K_{sat1}} - \frac{K_{fluido1}}{\phi\left(K_{mineral} - K_{fluido1}\right)} \quad (2.12)$$

Para o problema de substituição de fluidos existem dois efeitos de fluidos que devem ser considerados: a mudança na densidade da rocha e a mudança na compressibilidade da rocha. Rochas pouco consolidadas, ou com micro fissuras são geralmente menos duras e possuem pequena rigidez no espaço poroso. Rochas mais duras, que são bastante cimentadas, sem micro fissuras, possuem alta rigidez no espaço poroso. É importante salientar que a sensibilidade sísmica aos fluidos nos poros não está unicamente relacionada à porosidade. A sensibilidade sísmica aos fluidos é determinada por uma combinação que envolve a porosidade, a constituição mineral e a rigidez do espaço poroso.

Para realizar a análise de sensibilidade aos fluidos, o cenário mais comum é começar com um conjunto inicial de velocidades ($VP\ e\ VS$) e de densidades, correspondentes a rocha com um fluido inicial, que pode ser a condição *in situ*. Estas velocidades geralmente se originam de perfis de poços, porém, também podem advir de medidas de laboratório e modelos teóricos.

O procedimento de substituição do fluido se dá em cinco passos (Avseth *et al.*, 2005):

1. No primeiro passo, devem-se calcular os módulos *bulk* e cisalhante, a partir das velocidades e da densidade, pelas seguintes equações:

$$K_{sat1} = \rho\left((VP_1)^2 - \frac{4}{3}(VS_1)^2\right)$$

$$\mu_{sat1} = \rho(VS_1)^2$$

Onde K_{sat1} e μ_{sat1}são respectivamente os módulos bulk e cisalhante da rocha saturada com o fluido 1.

2. Em seguida a equação de Gassmann deve ser aplicada para calcular o novo módulo bulk da rocha saturada com o fluido 2 (K_{sat2}):

$$K_{sat2} = \frac{\left(\frac{K_{sat1}}{K_{mineral} - K_{sat1}} - \frac{K_{fluído1}}{\emptyset\left(K_{mineral} - K_{fluído1}\right)} + \frac{K_{fluído2}}{\emptyset\left(K_{mineral} - K_{fluído2}\right)}\right) \cdot K_{mineral}}{1 + \left(\frac{K_{sat1}}{K_{mineral} - K_{sat1}} - \frac{K_{fluído1}}{\emptyset\left(K_{mineral} - K_{fluído1}\right)} + \frac{K_{fluído2}}{\emptyset\left(K_{mineral} - K_{fluído2}\right)}\right)}$$

3. O módulo cisalhante permanece inalterado:

$$\mu_{sat2} = \mu_{sat1}$$

4. Neste passo deve-se corrigir a nova densidade da rocha para o fluido 2, através da equação:

$$\rho_2 = \rho_1 + \phi(\rho_{fluido2} - \rho_{fluido1})$$

5. Por fim, calculam-se as novas velocidades para o fluido 2.

$$VP_2 = \sqrt{\frac{K_{sat2} + \frac{4}{3}\mu_{sat2}}{\rho_2}}$$

$$VS_2 = \sqrt{\frac{\mu_{sat2}}{\rho_2}}$$

Quando se realiza o cálculo da substituição de fluidos, alguns cuidados devem ser observados. Um deles é utilizar as corretas propriedades dos fluidos nas equações. Outro cuidado deve ser tomado quando a rocha está saturada de gás, neste caso, ele também deve ser considerado como um fluido. A equação de Gassmann também só se aplica a rochas homogêneas e isotrópicas. Além destes, outras premissas não citadas também devem ser observadas.

2.2.2. Relações entre VP e VS

Outra parte importante da física de rochas é o uso da onda S na caracterização e monitoração de reservatórios. Ao adicionar a informação trazida pela onda S à informação da onda P, geralmente obtém-se uma melhor separação nos padrões dos atributos elásticos das diferentes litologias. As relações entre VP e VS são chaves para a determinação da litologia e do fluido, a partir de dados sísmicos ou a partir do perfil sônico.

Há uma grande variedade de relações publicadas entre VP e VS, além de técnicas de previsão de VS, que a princípio parecem ser bastante distintas. No entanto, a maioria delas adota sempre os mesmos dois passos:

- Estabelecer relações empíricas entre VP, VS e a porosidade para um fluido de referência, geralmente a rocha saturada de água ou a rocha seca.
- Utilizar a equação de Gassmann para mapear essas relações empíricas para outros fluidos.

Pickett (1963) e Castagna *et al.* (1993) publicaram relações empíricas entre VP e VS para os calcários saturados de água. Para velocidades mais elevadas, a linha reta de Pickett se adapta melhor aos dados. Contudo, a velocidades mais baixas (porosidades mais elevadas), os dados desviam-se da linha reta e tendem as velocidades na água ($VP = 1{,}5\ km/s;\ VS = 0\ km/s$). Nos mesmos trabalhos foram publicadas relações entre VP e VS para a dolomita, também saturada de água.

Para os arenitos e folhelhos Castagna *et al.* (1985, 1993) e Han (1986) também publicaram relações empíricas entre VP e VS. Dentre as três equações, as de Han e Castagna *et al.* (1993) são essencialmente a mesma e fornecem o melhor ajuste global para os arenitos. Castagna *et al.* (1993) sugere que se a litologia é bem conhecida, é possível ajustar essas relações diminuindo ligeiramente a razão VP/VS quando há maiores conteúdos de argila e aumentando-o quando há areias mais limpas. Quando a litologia não é bem conhecida, as equações de reta de Han (1986) e Castagna *et al.* (1993) fornecem uma média razoável. A Tabela 2.1 apresenta as equações citadas anteriormente.

Tabela 2.1 - Relações entre $\boldsymbol{VP}$ e $\boldsymbol{VS}$ por litologia.

Litologia	Relação entre VP e VS (km/s)	Autor
Calcário	$VS = VP/1{,}9$	Pickett (1963)
Calcário	$VS = -0{,}05508\mathrm{VP}^2 + 1{,}0168\mathrm{VP} - 1{,}0305$	Castagna *et al.* (1993)
Dolomita	$VS = VP/1{,}8$	Pickett (1963)
Dolomita	$VS = 0{,}5832\mathrm{VP} - 0{,}0777$	Castagna *et al.* (1993)
Arenito	$VS = 0{,}8042\mathrm{VP} - 0{,}8559$	Castagna *et al.* (1993)
Arenito	$VS = 0{,}7936\mathrm{VP} - 0{,}7868$	Han (1986)
Arenito/Folhelho	$VS = 0{,}8621\mathrm{VP} - 1{,}1724$	Castagna *et al.* (1985)

Greenberg e Castagna (1992) forneceram um método empírico para estimar VS a partir de VP para rochas com múltiplos minerais e saturadas com água salgada, baseados em relações empíricas polinomiais de rochas com apenas um mineral. Para cada litologia, VS foi estimado a partir da equação (2.13).

$$VS = a_{i2} \cdot VP^2 + a_{i1} \cdot VP + a_{i0} \tag{2.13}$$

Onde VP e VS são respectivamente as velocidades da onda P e S em rocha saturada de água. Tanto VP quanto VS estão em $(\mathrm{km/s})$ e os coeficientes da equação para cada litologia estão listados na Tabela 2.2.

Tabela 2.2 - Coeficientes da equação de Greenberg e Castagna (1992) por litologia.

Litologia/Coeficientes	a_{i2}	a_{i1}	a_{i0}
Arenito	0	0,80416	−0,85588
Calcário	−0,05508	1,01677	−1,03049
Dolomita	0	0,58321	−0,07775
Folhelho	0	0,76969	−0,86735

2.3. Fluxo de trabalho do intérprete

Nesta seção, será apresentada a abordagem frequentemente utilizada pelo intérprete para realização da discriminação litológica a partir de atributos elásticos. O trabalho do intérprete, realizado por geólogos e geofísicos, tem por objetivo extrapolar o modelo geológico concebido através da interpretação dos dados sísmicos e da integração das informações nos poços disponíveis, para toda a extensão da área de estudo.

Evidentemente, na indústria do petróleo, esta caracterização geológica tem por objetivo final a descoberta de acumulações de hidrocarbonetos, óleo ou gás. Os locais de um campo considerados com maior probabilidade de encontrar essas acumulações são chamados de "*sweet-spots*". Estes locais são reservatórios que possuem características tais como, baixa saturação de água, alta porosidade, alta permeabilidade, dentre outros que são favoráveis à presença de acumulação e prospecção de hidrocarbonetos.

O primeiro passo realizado pelo intérprete da área envolve a seleção dos poços que farão parte da interpretação. Além do critério de localização geográfica,

um poço normalmente é selecionado por apresentar disponibilidade de dados, principalmente aqueles utilizados na discriminação litológica através de atributos elásticos. Dentre esses dados, destacam-se os perfis de poços sônico compressional (DT) e sônico cisalhante (DTS), que permitem o cálculo das respectivas velocidades sísmicas, o perfil de densidade, que permite o cálculo das impedâncias ($IP\ e\ IS$), além do perfil de porosidade que é utilizado na análise de sensibilidade aos fluidos.

Escolhidos os poços, o próximo passo é o controle de qualidade dos perfis. Nesta fase é realizada uma inspeção com o objetivo de verificar se os perfis possuem medições confiáveis. Além dos principais perfis para a discriminação litológica, também são verificados perfis auxiliares como o "*caliper*" que mede o diâmetro do poço. Através deste perfil é possível verificar se houve arrombamento na parede do poço o que poderia levar a medições incorretas. Caso seja constatado algum problema na qualidade dos perfis, eles podem ser corrigidos através de algum cálculo teórico ou podem ser simplesmente descartados. A Figura 2.1 apresenta a seção de um poço onde são realizadas as inspeções para o controle de qualidade dos perfis.

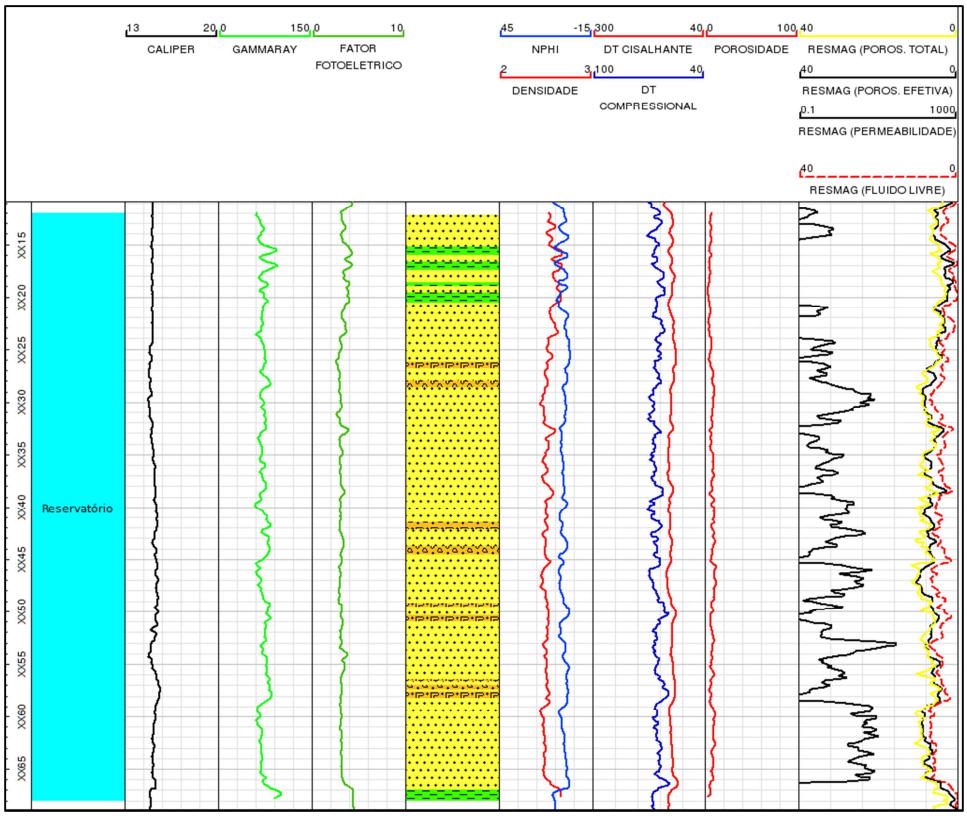


Figura 2.1 - Seção de poço utilizada para inspeção visual dos perfis.

Em seguida, uma série de gráficos de dispersão (*cross-plots*) é analisada pelo intérprete com o objetivo de encontrar alguma correlação entre as litologias e os atributos elásticos, baseados nas equações de física de rochas. A escolha das variáveis que serão analisadas nos gráficos depende de qual propriedade geológica pretende-se extrapolar com os atributos elásticos. No caso da discriminação litológica, os gráficos normalmente utilizados são o de $VP\ x\ VS$ e de $IP\ x\ IS$. Neles, são utilizadas as equações fornecidas pela física de rochas que relacionam VP, VS e a litologia. Caso as equações estejam em uma mesma tendência que os dados de poços, elas são utilizadas para extrapolar o conhecimento da litologia para fora do poço. Na Figura 2.2 é apresentado um *cross-plot* de $IP\ x\ IS$ dividido em três litologias: conglomerado, arenito e folhelho. No gráfico também são apresentadas as linhas de tendência lineares dos dados e as equações de física de rochas para arenitos (Castagna *et al.*, 1993) e folhelhos (Greenberg & Castagna, 1992).

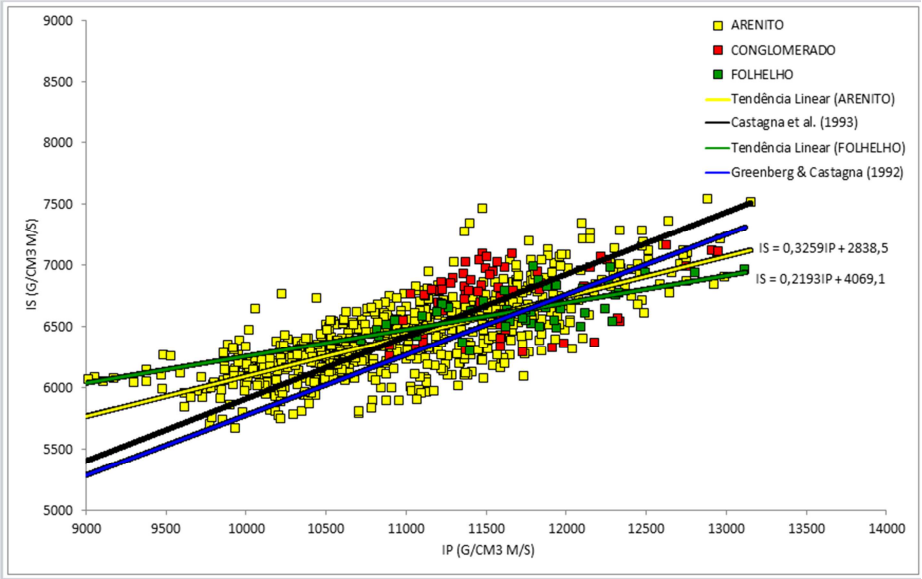


Figura 2.2 - *Cross-plot* de IP x IS a partir de dados de perfis de poços com linhas de tendências lineares e equações de física de rochas para cada litologia.

Nesta etapa, é importante salientar que a escolha final dos atributos realmente utilizados na extrapolação se dá de maneira qualitativa. Ou seja, ela depende do conhecimento e da experiência do intérprete, que seleciona os atributos que lhe parecem mais apropriados. Outro fator importante é a limitação no número de atributos utilizados. Como a seleção é baseada em *cross-plots* de duas dimensões, o número máximo de atributos elásticos selecionados geralmente

é de no máximo dois. Além disso, a densidade obtida da sísmica em geral é pouco confiável, de modo que só dois volumes de atributos elásticos são disponíveis.

Em muitos casos, a tendência dos dados de poços não pode ser representada por nenhuma equação de física de rochas conhecida, como é o caso da Figura 2.2. Isso acontece, porque as equações de física de rochas são baseadas em relações empíricas e para um ambiente geológico específico. Neste caso, o intérprete frequentemente utiliza a técnica estatística da análise discriminante para extrapolar o conhecimento presente nos dados de poços. Neste caso, a extrapolação ocorre através de um classificador estatístico e não através de uma equação que funciona como uma regressão. Novamente, é o intérprete quem seleciona, qualitativamente, os atributos que serão utilizados no classificador. Neste caso, esta escolha se dá através da análise dos histogramas de cada atributo elástico, agrupados pelas litologias de interesse. Na Figura 2.3 são apresentados dois histogramas utilizados na seleção de atributos. Além deles, os gráficos das respectivas distribuições gaussianas também são apresentados. O primeiro histograma é relativo aos dados de IP e o segundo de IP/IS. Neste exemplo, o intérprete selecionou o atributo IP como melhor discriminante litológico em relação ao atributo IP/IS. Esta seleção foi baseada na análise gráfica das distâncias entre as médias das respectivas distribuições gaussianas. Quanto maior forem estas distâncias melhor será o atributo para discriminar as litologias.

No caso do uso da análise discriminante, apesar da técnica não limitar o número de atributos envolvidos na classificação, eles são limitados pelo intérprete a no máximo dois ou três. Isso acontece, pois há a necessidade de entendimento pelo intérprete da “lei” ou regra que está sendo utilizada para extrapolar o conhecimento adquirido nos poços. Utilizando-se poucos atributos elásticos é possível interpretar graficamente, como os dados estão sendo extrapolados através da classificação estatística realizada pela análise discriminante.

Entender de que maneira está sendo feita a extrapolação é primordial para o intérprete, pois desta maneira é possível saber de antemão se restrições geológicas e físicas estão sendo respeitadas. De outra forma, a extrapolação estaria apenas utilizando-se de correlações estatísticas entre os atributos elásticos e a litologia. Evidentemente, restrições tanto no número de atributos quanto na utilização de técnicas de extrapolação ou classificação mais simples, que possam ser mais facilmente interpretáveis, podem levar a uma menor taxa de acerto na previsão.

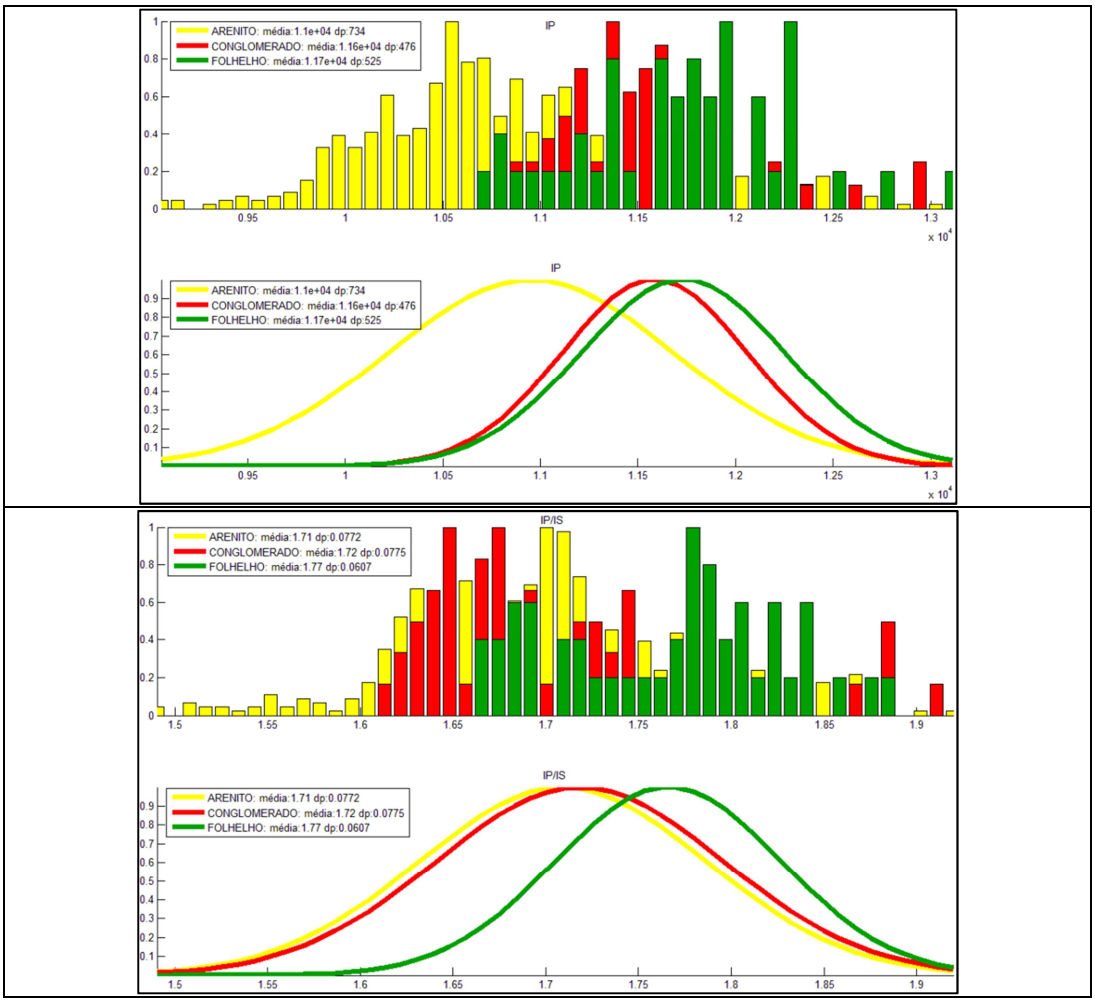


Figura 2.3 - Histogramas de $\boldsymbol{IP}$ e $\boldsymbol{IP/IS}$ para cada fácies identificada no(s) poço(s), além das respectivas distribuições gaussianas. Neste caso, o atributo $\boldsymbol{IP}$ foi considerado melhor discriminante do que o atributo $\boldsymbol{IP/IS}$.

2.4. Fluxo de trabalho estatístico e de física de rochas

Neste seção será apresentado o fluxo de trabalho híbrido denominado estatístico e de física de rochas (Avseth *et al.*, 2005). Nesta abordagem os autores propõem uma metodologia que envolve desde a seleção dos melhores atributos através da teoria da informação, passando pelo uso da teoria de física de rochas e de métodos estatísticos para a interpretação quantitativa de atributos elásticos para a previsão de propriedades de reservatórios. Esta abordagem, ao contrário da abordagem do intérprete que utiliza de forma independente os métodos de física de rochas e as técnicas estatísticas, propõem um fluxo de trabalho híbrido que integram ferramentas dessas duas áreas. Além disso, a metodologia faz uso de

procedimentos mais quantitativos, ao contrário da abordagem do intérprete que é mais qualitativa.

Segundo alguns autores (Avseth *et al.*, 2005) (Nikravesh *et al.*, 2003), a utilização de metodologias híbridas implicaria na diminuição da incerteza, elevando a confiabilidade das propriedades geológicas previstas, contribuindo de forma efetiva no aprimoramento da interpretação de reservatórios. Nikravesh *et al.* (2003) argumenta que métodos híbridos, combinando a força de técnicas estatísticas e de física de rochas, seriam mais eficazes. Na Figura 2.4 são apresentados esquematicamente os três métodos propostos: estatísticos, física de rochas e hibrido.

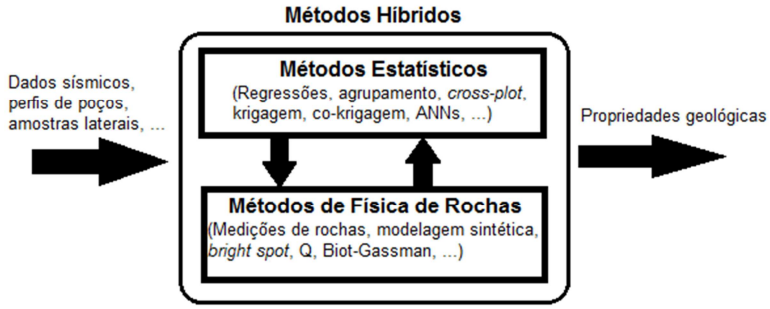


Figura 2.4 - Descrição esquemática dos métodos estatísticos, de física de rochas e híbridos para relacionar propriedades geológicas com atributos elásticos. (alterado de Nikravesh *et al.*, 2003).

O fluxo de trabalho estatístico e de física de rochas se divide em quatro etapas principais. A primeira etapa tem por objetivo a identificação de fácies a partir de perfis de poços e da geologia. A segunda envolve a utilização da teoria de física de rochas e da simulação de Monte Carlo para a geração de atributos teóricos. A terceira etapa envolve a criação de um modelo classificador, que a partir dos atributos elásticos derivados dos perfis de poços, possa prever propriedades geológicas em um volume sísmico. O quarto e último passo é a utilização de técnicas geoestatísticas que tem por objetivo levar em conta correlações espaciais geologicamente realísticas e a incerteza espacial das propriedades do reservatório.

Como o foco do presente trabalho é a discriminação litológica através de atributos elásticos derivados de perfis de poços, serão vistos em maiores detalhes a primeira, segunda e terceira etapa desta abordagem. O objetivo é apresentar os

procedimentos que vão desde a definição de fácies nos poços até a criação do modelo classificador. Portanto, não serão abordados a previsão de propriedades geológicas em volumes sísmicos, tão pouco a utilização de técnicas geoestatísticas.

Identificação de fácies a partir de perfis de poços e da geologia

Normalmente, a informação que se origina a partir dos poços é a observação mais direta disponível do reservatório. Em muitos projetos de caracterização de reservatório o primeiro passo é definir, com base nas informações de poços, as fácies desejáveis de serem previstas no reservatório. O termo fácies é utilizado neste contexto para definir grupos categóricos, não necessariamente apenas para tipos de litologias, mas também para alguma propriedade ou conjunto de propriedades. Por exemplo, uma combinação entre litologia e os fluidos intraporos como arenitos saturados de água e arenitos saturados com óleo seriam consideradas duas fácies ou categorias diferentes.

Utilizando a informação disponível nos poços como amostras laterais, testemunhos, amostras de calha e perfis de poços, o intérprete identifica cada profundidade com uma das fácies. É conveniente realizar este procedimento com poços onde os dados e a interpretação estão mais completos e confiáveis. O critério para definir as fácies depende do objetivo a ser alcançado. Este pode ser o mapeamento de diferentes litologias, a delimitação de fraturas, a identificação de fluidos ou a monitoração das mudanças na pressão e temperatura do reservatório.

Física de rochas e simulação de Monte Carlo

As diferentes condições físicas ou fácies de interesse que se pretende identificar, nem sempre estarão adequadamente amostradas nos dados iniciais dos poços de treinamento. Geralmente é necessário estender os dados de treinamento, utilizando a física de rochas para simular diferentes condições físicas teóricas.

Um pressuposto fundamental no processo de calibração em poços é que os perfis de poços, estendidos através dos modelos de física de rochas, serão estatisticamente representativos de todos os valores possíveis dos atributos elásticos que possam ser encontrados na área de estudo. A decisão subjetiva de que os dados de treinamento são um conjunto estatisticamente representativo

influencia o desempenho de todas as técnicas de classificação. A partir deste pressuposto é possível explorar a variabilidade intrínseca de cada fácies no espaço dos atributos elásticos utilizando simulações de Monte Carlo.

A fim de se estabelecer uma relação com a informação sísmica, atributos elásticos são teoricamente calculados utilizando os perfis dos poços de treinamento. Simulações de Monte Carlo são projetadas a partir da distribuição de cada fácies definida anteriormente sendo utilizados modelos determinísticos para calcular os atributos elásticos. Apesar de esta metodologia ser de caráter geral e poder ser aplicada a qualquer conjunto de atributos matematicamente calculados, neste caso somente atributos elásticos que possuam significado físico são considerados.

O próximo passo é a seleção do melhor atributo elástico, ou conjunto de atributos. Este passo depende do reservatório em questão e do problema a ser resolvido, que pode ser uma discriminação litológica, a detecção de fluidos ou a identificação de uma zona fraturada, dentre outros. Tradicionalmente, esta seleção é feita através da análise visual de gráficos como histogramas e de dispersão (*cross-plots*), como foi visto na abordagem do intérprete. Uma técnica mais quantitativa proposta por esta metodologia é a utilização dos conceitos originados na teoria de informação, especialmente o conceito de informação mútua, pois eles medem estatisticamente quais são as variáveis que mais contribuem para a solução de um problema específico (Cover & Thomas, 2006)(Takahashi *et al.*, 1999).

Classificação dos atributos elásticos

A próxima etapa é a criação de um modelo classificador, que a partir dos atributos elásticos derivados dos perfis de poços discrimine um volume ou horizonte de atributos elásticos em diferentes classes. Estas classes são as fácies definidas nos poços de treinamento na primeira etapa, que podem representar diferentes litologias, tipos de fluidos, rochas fraturadas e não fraturadas, dentre outros. Neste ponto a abordagem não determina qual técnica de classificação deverá ser empregada. Alguns exemplos citados incluem a análise discriminante, a classificação através dos k vizinhos mais próximos, redes neurais, árvores de classificação e a classificação Bayesiana.

3
Programação Genética

Neste capítulo será apresentada uma revisão dos conceitos relacionados à programação genética. Alguns desses conceitos envolvem o modo de representação, a inicialização da população, os modos de seleção, as funções de aptidão, os operadores genéticos, dentre outros. Também será apresentada uma revisão bibliográfica do uso da programação genética para o problema de classificação.

A Programação Genética (PG) é uma técnica de computação evolucionária que resolve problemas automaticamente, sem a exigência prévia de o usuário conhecer ou especificar a forma ou a estrutura de uma solução (Koza, 1992). Em PG uma população de programas de computador é evoluída, isto é, de geração em geração a PG transforma populações de programas de forma estocástica em novas populações de programas, através de determinadas operações com o objetivo de que essas novas populações geradas sejam melhores do que as anteriores.

Um sistema baseado em PG avalia o quão bem um programa resolve determinado problema executando-o (Poli *et al.*, 2008). Em seguida, o sistema compara o comportamento do programa com um objetivo definido. Esta comparação é quantificada no sentido de atribuir um valor numérico ao programa, denominado aptidão. Aqueles programas da população que possuem as melhores aptidões são selecionados para produzirem descendentes, ou seja, novos programas para a próxima geração. Os dois principais operadores genéticos utilizados na criação de novos programas, a partir de programas já existentes, são o cruzamento e a mutação. O cruzamento consiste na criação de um novo programa através da combinação de partes aleatoriamente selecionadas de dois programas genitores, enquanto a mutação é a criação de um novo programa através da alteração de uma parte aleatoriamente selecionada de um programa genitor. As etapas básicas de um sistema baseado em PG são apresentadas no fluxograma da Figura 3.1.

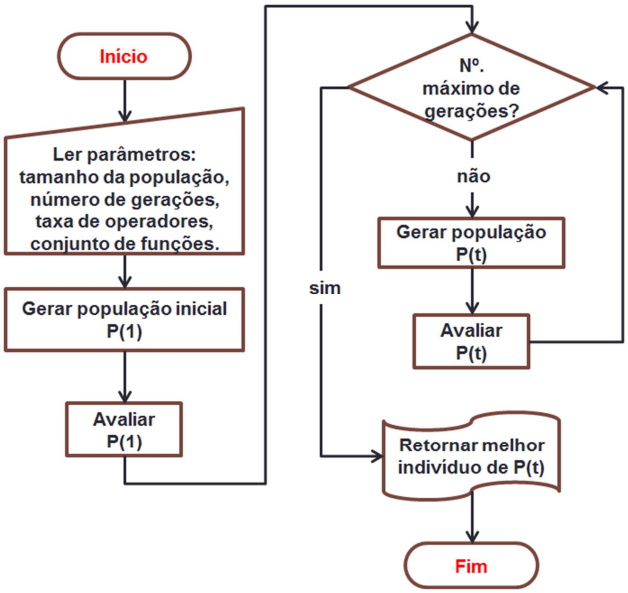


Figura 3.1 - Etapas básicas de um sistema de Programação Genética.

Na programação genética, os programas são usualmente expressos como árvores sintáticas, em vez de linhas de código (Espejo *et al.*, 2010). A Figura 3.2 apresenta a árvore que representa o programa $x^2 + x + 1$. As variáveis e constantes do programa $(x\ e\ 1)$, que formam as folhas da árvore, na terminologia de PG são denominados terminais, enquanto as operações aritméticas $(+\ e\ *)$, que formam os nós internos, são chamadas de funções. As funções e os terminais formam o conjunto de primitivas de um sistema de PG. A escolha dos conjuntos de terminais e funções são os dois primeiros passos preparatórios para a aplicação de PG em um problema. Juntos eles definem quais serão os elementos disponíveis para a criação dos programas.

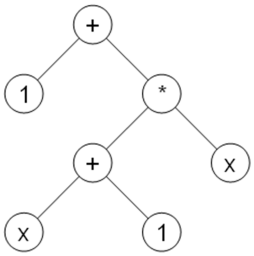


Figura 3.2 - Árvore sintática de PG representando o programa $\boldsymbol{x^2 + x + 1}$.

Segundo Koza (1992), para que um algoritmo baseado em PG funcione adequadamente, as funções devem possuir uma importante propriedade denominada fechamento. Esta propriedade pode ser dividida em dois conceitos: consistência de tipo e segurança no cálculo da aptidão.

A consistência de tipo é necessária, pois a operação de cruzamento pode misturar e agregar nós arbitrariamente. Então, faz-se necessário que qualquer subárvore possa ser usada como argumento de qualquer função do conjunto de funções. Para assegurar que isto possa ser feito é exigido que todas as funções apresentem consistência de tipo, ou seja, que todas retornem valores do mesmo tipo e que todos os seus argumentos também sejam deste mesmo tipo. O outro conceito da propriedade de fechamento é a segurança no cálculo da aptidão. Ela é exigida, pois muitas das funções comumente utilizadas podem falhar em tempo de execução. A segurança geralmente é obtida pela modificação do comportamento normal da função por versões protegidas. Estas versões protegidas das funções, primeiramente testam seus argumentos de entrada procurando por eventuais problemas antes de executar o cálculo propriamente dito. Se um problema é detectado, um valor padrão é retornado.

Assim como em outros algoritmos evolucionários, em PG os indivíduos da população inicial são gerados aleatoriamente. As abordagens mais amplamente utilizadas são os métodos *full*, *grow* e uma combinação desses dois, chamado *Ramped half-and-half* (Poli *et al.*, 2008). O método *full* possui esse nome devido ao fato dele gerar árvores cheias, ou seja, todas as folhas estão numa mesma profundidade. O método *grow*, ao contrário do método *full*, permite a criação de árvores de formatos e tamanhos mais variados. Devido aos métodos *full* e *grow* fornecerem um leque muito variado de tamanhos e formatos por si mesmos, Koza (1992) propôs a combinação destes dois métodos num só chamado *Ramped half-and-half*. Neste método, metade da população inicial é construída utilizando-se o método *full* enquanto a outra metade é construída utilizando-se o método *grow*. A Figura 3.3 apresentam dois exemplos de árvores criadas pelos métodos *full*, à esquerda, e *grow*, à direita.

Figura 3.3 – Exemplos de árvores criadas pelos métodos *full*, a esquerda, e *grow*, a direita.

A definição do conjunto de primitivas indiretamente também define o espaço de busca a ser explorado pela PG. Ele inclui todos os programas que podem ser construídos pela composição dos elementos do conjunto de primitivas. Através da função de aptidão é possível saber quais elementos, ou regiões do espaço de busca, são melhores para a resolução do problema. A função de aptidão é o mecanismo através do qual é fornecida uma ligação de alto nível entre os requisitos do problema e o sistema de PG.

A aptidão pode ser medida de diversas maneiras, alguns exemplos são: a quantidade de erros entre a saída do programa e a saída correta desejada, a quantidade de tempo necessária para levar um sistema para um estado desejado, a taxa de acerto do programa em reconhecer padrões ou na classificação de objetos, a conformidade de uma estrutura de acordo com um projeto definido pelo usuário, dentre outros.

Assim como a maioria dos algoritmos evolucionários, operadores genéticos em PG são aplicados a indivíduos que são probabilisticamente selecionados baseados na aptidão. Isto é, melhores indivíduos são mais propensos de terem mais programas descendentes do que indivíduos inferiores. O método mais comumente empregado para a seleção de indivíduos em PG é a seleção por torneio. Na seleção por torneio um número de indivíduos é escolhido aleatoriamente da população. Eles são comparados entre si e o melhor de todos, baseado na aptidão, é selecionado para ser um dos genitores. Além da seleção por torneio, existem outros mecanismos utilizados em PG para a seleção de indivíduos como a seleção proporcional à aptidão e a amostragem universal estocástica (Goldberg, 1989).

Os sistemas baseados em PG diferenciam-se significativamente de outros algoritmos evolucionários no desenvolvimento de seus operadores de cruzamento e mutação. A forma mais comumente utilizada de cruzamento é o de subárvore. Dados dois genitores, o cruzamento de subárvore seleciona aleatoriamente um nó em cada árvore. Em seguida, os indivíduos da próxima geração são criados pela substituição da subárvore cuja raiz é o nó escolhido. A subárvore do primeiro genitor é substituída por uma cópia da subárvore do segundo genitor, enquanto a subárvore do segundo genitor é substituída por uma cópia da subárvore do

primeiro genitor. A Figura 3.4 apresenta um exemplo de cruzamento de subárvore.

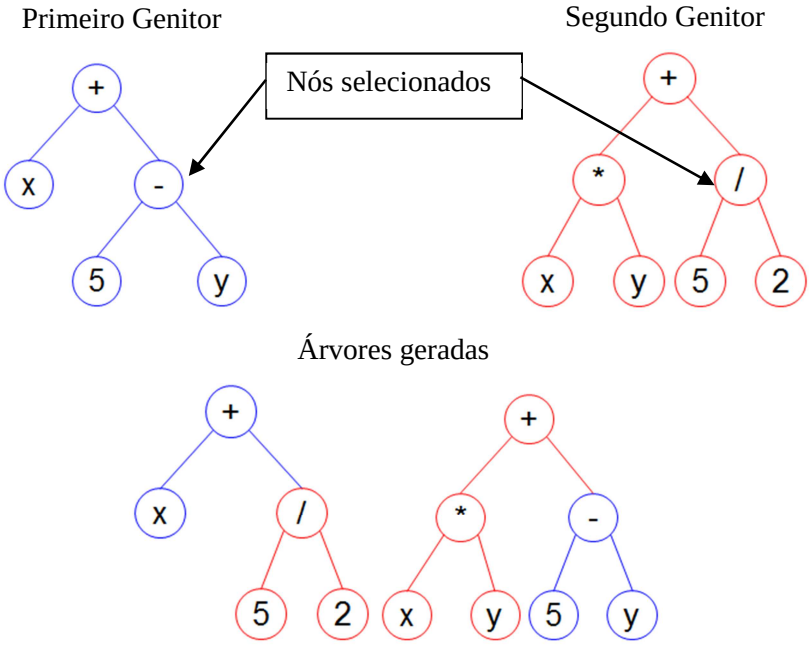


Figura 3.4 - Exemplo de um cruzamento de subárvore.

A forma mais comum de uso da mutação em PG é a chamada mutação de subárvore. Da mesma forma que o cruzamento, um nó é escolhido aleatoriamente sendo a sua subárvore substituída por uma nova subárvore gerada aleatoriamente. A mutação de subárvore geralmente é desenvolvida como um cruzamento entre um programa da população e um novo programa, aleatoriamente gerado para esta operação. A Figura 3.5 apresenta um exemplo de mutação de subárvore.

Outra forma geralmente utilizada é a chamada mutação pontual (*point mutation*), que é semelhante à mutação *bit-flip* utilizada em algoritmos genéticos (Goldberg, 1989). Na mutação pontual, um nó é aleatoriamente selecionado e o elemento do conjunto de primitivas representado por ele é substituído por outro elemento do conjunto de primitivas com o mesmo número de parâmetros. Se não há nenhuma outra primitiva com o mesmo número de parâmetros, o nó é mantido inalterado.

Na seção 3.1 serão abordados os aspectos relacionados ao uso de PG para classificação. Dentre eles serão citados o uso de PG no pré-processamento e na extração de modelos classificadores com diferentes formalismos.

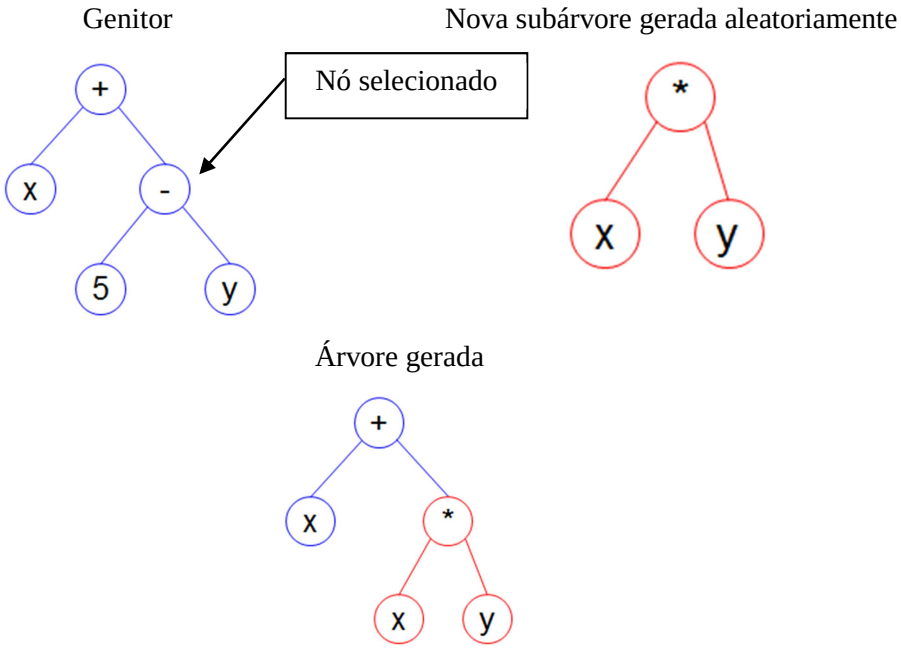


Figura 3.5 - Exemplo de uma mutação de subárvore.

3.1. Programação Genética para Classificação

Classificação é um dos problemas mais estudados na área de aprendizado de máquina e mineração de dados (Han *et al.*, 2011). O problema consiste em prever o valor de um atributo categórico, a classe, baseado nos valores de outros atributos. Um algoritmo de busca é utilizado na geração de um classificador, a partir de dados de um conjunto corretamente classificado chamado de conjunto de treinamento. Outro conjunto, também com dados corretamente classificados, conhecido como conjunto de teste é utilizado para medir a qualidade obtida pelo classificador ao avaliar sua capacidade de aprendizado quanto à generalização.

PG oferece um grande potencial para o problema de classificação (Espejo *et al.*, 2010). Primeiramente, PG pode ser aplicado na tarefa de pré-processamento dos atributos, tanto na seleção quanto na construção das características utilizadas na classificação. Essas tarefas são de grande importância, pois visam à melhoria no desempenho dos classificadores. Além disso, PG é uma técnica muito flexível que permite o uso de diversos padrões de representação, tais como as árvores de decisão, regras de classificação e funções discriminantes. A interpretação do classificador é outra característica muito favorecida pelo uso de PG, uma vez que é possível adotar formalismos de representação mais interpretáveis. Os

mecanismos disponíveis para limitação do tamanho do classificador resultante, também contribuem para aprimorar a interpretabilidade.

Um dos formalismos que mais privilegia a interpretabilidade são as árvores de decisão. Elas são uma das representações mais utilizadas para classificadores (Murthy, 1998). Como apresentado anteriormente neste capítulo, as árvores sintáticas são a forma de representação mais comum em PG, sendo assim, a utilização de PG para a evolução de árvores de decisão acaba se tornando uma abordagem muito utilizada. Frequentemente, cada indivíduo da população de PG codifica uma árvore de decisão. Ela pode possuir de zero a muitos nós internos e de uma a muitas folhas, sendo que os nós internos possuem pelo menos dois nós filhos. Todos os nós internos possuem divisões que testam o valor de uma expressão em função dos atributos. As ligações entre um nó interno e seus filhos são identificadas como diferentes resultados do teste executado. Cada folha é identificada como representante de uma das classes. A Figura 3.6 apresenta um exemplo de árvore de decisão para três classes, arenito (ARN), conglomerado (CGL) e folhelho (FLH), com sua respectiva representação como um indivíduo de PG.

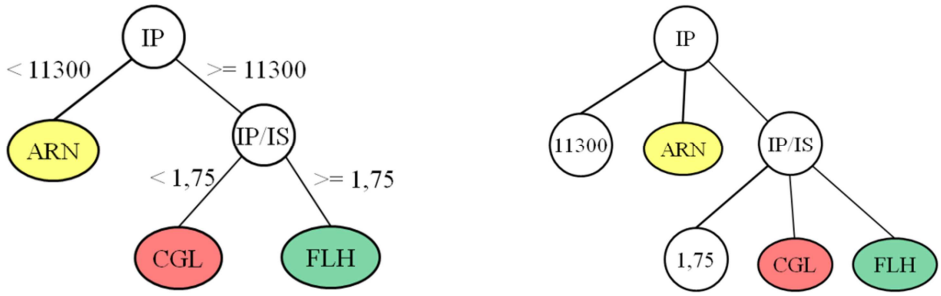
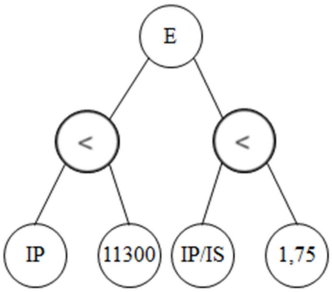


Figura 3.6 – À esquerda, exemplo de árvore de decisão para três classes, arenito (ARN), conglomerado (CGL) e folhelho (FLH). À direita sua representação como um indivíduo de PG.

As regras de decisão são uma maneira simples e facilmente interpretável para representar conhecimento (Han *et al.*, 2011)(Freitas, 2002). Uma regra tem duas partes, o antecedente e o consequente. O antecedente da regra contém uma combinação de condições para os atributos previsores. Geralmente, essas condições são formadas por uma conjunção do operador lógico “E”, porém qualquer operador lógico pode ser usado para conectar condições elementares. O consequente da regra contém o valor previsto para a classe. Desta forma, uma regra atribui uma classe, contida no consequente, a uma instância de dados se os

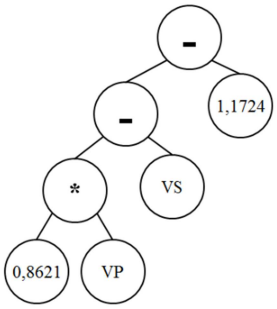
valores dos atributos satisfazem as condições expressas no antecedente. Sendo assim, um classificador é representado por um conjunto de regras. A Figura 3.7 apresenta um indivíduo de PG representando sua respectiva regra de classificação.



$$Se\ IP < 11300\ E\ \ IP/IS < 1{,}75$$
$$\rightarrow classe = arenito$$

Figura 3.7 – À esquerda, indivíduo de PG representando a regra de classificação à direita.

Outro formalismo para a representação de classificadores são as funções discriminantes. Uma função é uma expressão matemática em que diferentes tipos de operadores são aplicados aos atributos de um banco de dados, que deve ser classificado. Desta forma, um único valor é computado a partir das operações empregadas nos valores dos atributos. O valor calculado pela função indica a classe prevista. Geralmente, este procedimento é alcançado através do uso de um valor limítrofe (*threshold*) ou um conjunto de valores limítrofes. Para problemas com apenas duas classes, uma única função discriminante é necessária. Se o valor de saída é maior que o valor limítrofe, então o exemplo é atribuído a uma classe, caso contrário, ele é atribuído à outra classe. Normalmente, o valor mais utilizado como limítrofe é zero, assim um valor de saída positivo indica uma classe e um valor negativo indica a outra classe. A Figura 3.8 apresenta um indivíduo de PG representando sua respectiva função discriminante.



$$0{,}8621 * VP - VS - 1{,}1724$$

Figura 3.8 – À esquerda, indivíduo de PG representando a função discriminante à direita.

Além de atuar nas tarefas de pré-processamento e na evolução de modelos classificadores, PG também pode ser utilizado na melhoria do desempenho de

classificadores através do emprego de vários deles, em vez de apenas um. Esta é a ideia básica dos comitês de classificadores (*ensemble classifiers*). Duas questões principais envolvendo o uso de comitês podem ser solucionadas através do uso de PG. A primeira é a geração dos diversos classificadores e a segunda é como combinar esses classificadores.

A Figura 3.9 apresenta as questões relacionadas à classificação onde PG pode ser aplicado.

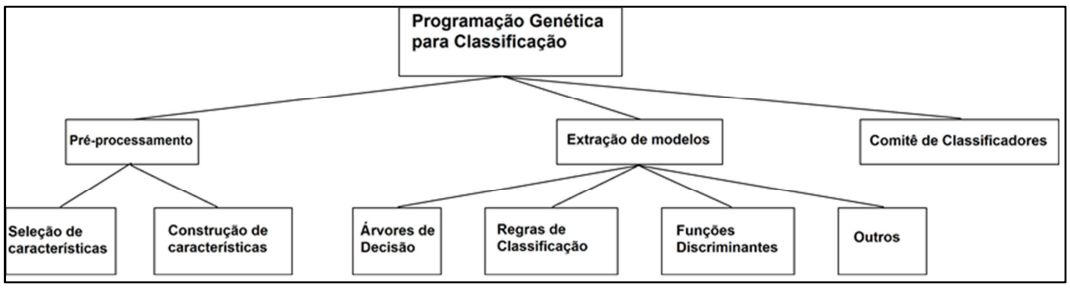


Figura 3.9 - Possibilidades de uso de PG para classificação (Espejo et al., 2010).

3.2. Seleção de características

As técnicas de geração de modelos para classificação geralmente não são aplicadas à base de dados em seu formato original. Existe uma grande variedade de técnicas de pré-processamento disponíveis, que têm por objetivo preparar os dados em seu potencial máximo (Espejo *et al.*, 2010). Algumas das questões envolvidas nessas técnicas dizem respeito à remoção de dados ruidosos ou fora de escala (*outliers*), estratégias para lidar com dados ausentes, seleção e construção de características[3], balanceamento de dados, normalização, dentre outros.

Todas as técnicas de pré-processamento têm como entrada o conjunto original de características C_0 e um conjunto de treinamento. O objetivo é criar outro conjunto de características C, derivado de C_0, que maximize algum critério, e que seja tão bom quanto C_0 em relação a este critério (Liu & Motoda, 1998). As abordagens que seguem este esquema podem ser agrupadas em três categorias: métodos de seleção de características, métodos de ponderação de características e métodos de construção de características (Espejo *et al.*, 2010).

[3] No caso do problema abordado no presente trabalho, as características citadas seriam os atributos elásticos.

Nos métodos de seleção de características o conjunto de características resultantes C é um subconjunto do conjunto original C_0, isto é, $C \subseteq C_0$. Os métodos de ponderação de características atribuem pesos a cada atributo, refletindo a importância relativa de cada um em relação aos outros. Nos métodos de construção de características novos atributos são criados através de algum mecanismo, como, por exemplo, pela utilização de expressões funcionais aplicadas aos valores das características originais.

Quando se utiliza a PG para a geração de um classificador, geralmente isto implica num processo de seleção de características que é inerente à evolução dos classificadores. Em PG, indivíduos de diferentes tamanhos são evoluídos. Esses indivíduos, como visto anteriormente, possuem uma estrutura na forma de árvore onde nós internos são funções e as folhas são constantes e variáveis. Estas variáveis correspondem às características do banco de dados. Como os indivíduos podem ter tamanhos variados, nem todas as características necessariamente irão aparecer em um indivíduo. É mais provável que apenas algumas das características estejam presentes em cada indivíduo. Portanto, implicitamente o mecanismo de seleção de características faz parte do processo evolucionário (Espejo *et al.*, 2010).

Pelo fato da seleção de características fazer parte implicitamente do processo evolucionário, poucos trabalhos foram publicados com a finalidade específica do uso de PG para esta finalidade. No trabalho de Davis *et al.* (2006), PG é empregada para classificação e a seleção de características. Cada indivíduo é uma árvore que codifica um classificador representado por uma função discriminante. A taxa de acerto da classificação é utilizada como função de aptidão. Os autores utilizaram a capacidade implícita de seleção de características de PG realizando a classificação em duas etapas: na primeira, um determinado número de execuções é realizado, cada uma resultando num melhor classificador; na segunda etapa, o sistema é novamente executado utilizando apenas as características que mais frequentemente apareceram nos melhores indivíduos da etapa anterior.

3.3. Evolução de modelos classificadores

A ideia básica para a aplicação de PG na geração de classificadores consiste em fazer com que cada indivíduo represente um classificador, ou uma parte de um classificador. Definindo-se uma função de aptidão para avaliar a sua qualidade, o esperado é que o processo de evolução leve a um classificador final de alta qualidade. A maioria dos trabalhos publicados, relacionados à aplicação de PG para classificação, concentra-se na sua utilização para extrair um modelo para o classificador. Esses modelos, como citados anteriormente, podem ser representados por formalismos mais facilmente interpretáveis, como as árvores de decisão, até as funções discriminantes que possuem uma estrutura mais complexa.

Quando uma elevada taxa de acerto é o principal objetivo do classificador, o modelo geralmente utilizado na maioria dos trabalhos são as funções discriminantes (Espejo *et al*., 2010). Funções discriminantes são também frequentemente empregadas em aplicações relacionadas à classificação de objetos em imagens e de reconhecimento de padrões. Para problemas de classificação com apenas duas classes, uma única função discriminante é necessária. Para problemas com mais de duas classes diferentes abordagens são propostas.

Para problemas de classificação com mais de duas classes, duas abordagens básicas podem ser encontradas nos trabalhos publicados. A primeira é considerar um problema de classificação de n classes como n problemas de classificação binária. Assim, n funções discriminantes são necessárias. Se o valor da saída é maior que um valor limítrofe, normalmente zero, o padrão é designado para uma determinada classe, caso contrário, ele é designado como não pertencendo a essa classe. Essa abordagem é geralmente denominada decomposição binária (*binary decomposition).* Técnicas de resolução de conflitos são frequentemente utilizadas nesta abordagem, caso mais de uma função forneçam resultados conflitantes.

A outra opção é utilizar apenas uma única função para distinguir todas as classes. Neste caso, $n-1$ valores limítrofes são necessários para definir n intervalos, sendo cada um designado para uma determinada classe. Assim, a classe prevista dependerá do intervalo ao qual o valor de saída pertencerá. Essa abordagem é geralmente denominada seleção por faixa (*range selection).* Diferentes abordagens podem ser encontradas a partir da ideia básica de uma

única função discriminante com múltiplos valores limítrofes. A mais simples é a determinação estática dos limites das classes (*static range selection*), que consiste em fixar os valores limítrofes através de pontos manualmente escolhidos. Outra abordagem, tendo como exemplo o trabalho de Zhang & Smart (2004), é a determinação dinâmica dos limites durante o processo evolucionário (*dynamic range selection*).

Um dos primeiros trabalhos abordando a classificação de múltiplas classes utilizando funções discriminantes evoluídas através de PG foi publicado por Kishore *et al.* (2000). Neste trabalho um problema de *n* classes é convertido em *n* problemas binários. O sistema é executado uma vez para cada classe, sendo evoluída uma função discriminante com um único valor limítrofe para a classe correspondente. A taxa de acerto da classificação é utilizada como função de aptidão. A resolução de conflitos é baseada na precisão de cada função além de regras heurísticas.

Um sistema para a evolução de funções discriminantes é descrito no trabalho de Brameier & Banzhaf (2001). Nesta abordagem cada indivíduo da população, em vez de uma árvore, é um código-fonte representando um programa na linguagem C, que mantém um vetor de saída contendo uma posição para cada classe a ser prevista. O programa calcula *n* valores de saída, um para cada classe. A classe prevista é aquela que apresenta o maior valor de saída. A função de aptidão combina o erro médio quadrático (*MSE*) com o erro médio de classificação (*MCE*).

Uma abordagem de classificação de múltiplas classes baseada em PG, em que é utilizada uma visão integrada de todas as classes, é proposta no trabalho de Muni *et al.* (2004). Neste sistema é possível construir um classificador completo em uma única execução. Cada indivíduo é uma estrutura composta por *n* árvores, sendo *n* o número de classes. Cada uma dessas *n* árvores codifica uma função discriminante para uma determinada classe. A função de aptidão é calculada como a taxa de acerto da classificação. Novamente, a resolução de conflitos é baseada na precisão de cada função, além de regras heurísticas.

Na proposta apresentada por Zhang & Smart (2006), cada indivíduo representa uma função discriminante com múltiplos valores limítrofes. Porém, em vez de utilizar apenas o melhor indivíduo evoluído da população, esta abordagem utiliza um conjunto dos melhores programas evoluídos para realizar a

classificação. Assume-se que o comportamento de um programa classificador é modelado utilizando múltiplas distribuições gaussianas, cada uma correspondendo a uma determinada classe. A função de aptidão é baseada na sobreposição da distribuição gaussiana de cada classe.

A classificação de múltiplas classes, por meio de um conjunto de funções discriminantes com um único valor limítrofe, é utilizada no trabalho de Chen & Lu (2007). Um conjunto de funções é evoluído, o que pode resultar em diversas funções para uma determinada classe. Cada uma dessas funções discrimina uma classe do resto. Quando uma instância deve ser classificada, cada uma das funções dá a sua saída e a previsão final é obtida através de um esquema de votação. A área sob a curva *Receiver Operating Characteristic* (ROC) é usada como função de aptidão.

No trabalho de Zhang & Wong (2008), a classificação de múltiplas classes é abordada pela evolução de uma única função discriminante com múltiplos valores limítrofes. Esta função distingue as n classes através de $n-1$ valores limítrofes. Estes valores limítrofes determinam n intervalos, sendo cada um deles atribuído a uma determinada classe. Desta maneira, a classe prevista irá depender do intervalo ao qual o valor de saída da função pertencerá. A taxa de acerto da classificação é utilizada como função de aptidão.

No trabalho proposto por Oltean & Diosan (2009) um único melhor indivíduo é evoluído como o classificador final. Cada classe possui um valor numérico associado. O padrão de dados é classificado como pertencente à classe que possui o valor mais próximo ao valor de saída calculado pela função discriminante evoluída. Nesta proposta é utilizado um vetor com índices de nós anexados a cada indivíduo para que este possa gerar múltiplas saídas, cada qual correspondendo a uma classe. Além disso, o conjunto de funções utilizado depende do tipo das variáveis de entrada. A função de aptidão é calculada como a taxa de acerto na classificação.

A Tabela 3.1 apresenta uma lista dos trabalhos aqui relatados contendo a abordagem utilizada, o conjunto de funções e a função de aptidão.

Tabela 3.1 - Listagem dos trabalhos de programação genética para classificação.

Trabalho	Abordagem	Conjunto de Funções	Função de Aptidão
Kishore *et al.* (2000)	decomposição binária	$(+,-,x,\div)$	Taxa de acertos
Chen & Lu (2007)	decomposição binária	$(+,-,x,\div)$	Área sob a curva ROC
Muni *et al.* (2004)	decomposição binária (múltiplos genes)	$(+,-,x,\div)$	Taxa de acertos
Zhang & Wong (2008)	seleção por faixas (estáticas)	$(+,-,x,\div,if)$	Taxa de acertos
Zhang & Smart (2006)	seleção por faixas (dinâmicas)	$(+,-,x,\div,if)$	Sobreposição da distribuição gaussiana entre classes
Oltean & Diosan (2009)	decomposição binária (múltiplas saídas)	$F_{Binário}$=(todos os operadores binários) $F_{Inteiro}=(+,-,x,\%,/)$ $F_{Real}=(+,-,x,/)$	Taxa de acertos
Brameier & Banzhaf (2001)	Linear Genetic Programming (múltiplas saídas)	$(+,-,x,\div,>,\leq, sin,cos,sqrt,exp,log)$	$MSE + w.MCE$

3.4. Comparação entre as abordagens

Primeiramente, pode-se constatar que a maioria dos trabalhos utiliza a abordagem de decomposição binária. A grande vantagem desta técnica é o seu fácil desenvolvimento. Porém, possui a desvantagem de apresentar o maior custo computacional uma vez que o sistema deve ser executado uma vez para cada classe. Somente no caso da utilização de múltiplos genes esta desvantagem é parcialmente eliminada, uma vez que a quantidade de operações neste caso é maior.

Conflitos ocorrem quando há mais de uma função discriminante com a saída positiva em relação a um padrão de dados. Diversos esquemas são propostos que vão desde medidas de precisão de cada função discriminante, passando por regras heurísticas retiradas da base de treinamento até esquemas baseados em votação. Outra abordagem apresentada é a utilização direta do valor real de saída de cada função discriminante. Neste caso duas opções foram apresentadas. Na primeira, a

classe selecionada é aquela que possui o valor mais próximo ao valor de saída calculado pela função discriminante, enquanto, na segunda, a classe selecionada é aquela cuja função resultou no maior valor de saída.

Nos trabalhos que utilizam a abordagem de seleção por faixas, onde não há a necessidade de resolução de conflitos, uma única função discriminante com múltiplos valores limítrofes é utilizada. O ponto chave nesta abordagem é a técnica de definição desses limites. Uma das técnicas utilizadas é a chamada *program classification map* (Zhang *et al.*, 2003) em que a definição dos limites é feita pelo usuário. Outra abordagem, em que os limites são modificados dinamicamente junto com o processo evolucionário, modela a saída dos programas como múltiplas distribuições gaussianas, uma para cada classe. Assim, a classe prevista de um padrão de dados é aquela na qual o valor de saída do programa atinge a maior probabilidade na respectiva distribuição gaussiana.

No trabalho de Loveard & Ciesielski (2001), além dos dois métodos citados anteriormente (decomposição binária e seleção por faixa), outros dois métodos foram abordados para a classificação de múltiplas classes utilizando PG*: class enumeration* e *evidence accumulation*. O método *class enumeration* constrói programas similares, em suas estruturas sintáticas, ao de uma árvore de decisão. O método *evidence accumulation* permite que diversos ramos do programa contribuam para a decisão de escolher uma determinada classe. Além da árvore do programa, cada indivíduo também contém uma área de armazenamento de dados denominado vetor de certeza. O trabalho foi desenvolvido utilizando-se três bases de dados com múltiplas classes: *Pixel*, *Thyroid* e *Vehicle* (Bache & Lichman, 2013). A principal conclusão do trabalho foi que os métodos decomposição binária e seleção por faixas dinâmicas foram os que obtiveram as maiores taxas de acerto. Além disso, o método seleção por faixas dinâmicas foi o mais consistente (menor desvio padrão dos erros) quando tempos de treinamento comparáveis são utilizados.

As duas abordagens para classificação de múltiplas classes, a construção de uma única função de classificação ou n classificadores binários, são comparados no trabalho de Teredesai & Govindaraju (2004). O problema abordado é o de reconhecimento de dígitos manualmente escritos. Quando uma única função capaz de discriminar todas as classes é evoluída, esta função gera diretamente o valor numérico da classe prevista, uma vez que cada classe é um dígito inteiro. A outra

opção consiste em executar o sistema tantas vezes quanto forem o número de classes, cada execução evoluindo uma única função discriminante para uma determinada classe. Em ambos os casos, a função de aptidão é baseada na taxa de acerto da classificação. O trabalho enumera algumas conclusões, dentre elas pode-se citar:

- A reorganização do conjunto de treinamento desempenha um papel importante em termos de uma convergência mais rápida na execução do sistema;
- A classificação baseada em PG utilizando vetores com múltiplas características é convergente;
- Decompor a classificação de n classes em n classificadores binários é a melhor abordagem quando se utiliza PG;
- A classificação exige que boas funções de aptidão devam ser projetadas no futuro, para conduzir o processo de treinamento de forma mais eficaz.

Três abordagens baseadas em PG são propostas no trabalho de Zhang & Nandi (2007) para resolver problemas de classificação de múltiplas classes em detecção de falhas em rolamentos. A primeira abordagem utilizada, denominada *single-GP scheme*, evolui uma única função discriminante e é semelhante à seleção por faixas estáticas. A segunda, denominada *independent-GP scheme* é semelhante ao método de decomposição binária, e a última, denominada *bundled-GP scheme*, é uma alternativa ao método de decomposição binária na qual todos os PGs binários, um para cada classe, são evoluídos conjuntamente com o objetivo de utilizar uma menor quantidade de características. Estas abordagens são comparadas com as redes neurais *multi-layer perceptron* (MLP) e as *support vector machines* (SVM), ambas utilizando algoritmo genético (GA) para selecionar o conjunto ótimo de características. As conclusões do trabalho são as seguintes:

- Tanto a abordagem *bundled-GP* quanto a *independent-GP* igualaram ou superaram na taxa de acerto as abordagens baseadas em GA;
- A abordagem *bundled-GP* atingiu, ligeiramente, uma menor taxa de acerto do que a *independent-GP*, porém selecionou menos características;
- A abordagem s*ingle-GP* é a melhor em relação à utilização de menos características, porém, ela não foi a melhor estratégia, pois apresentou as piores taxas de acertos;

- Ambas as abordagens de decomposição binária (*independent-GP* e *bundled-GP*) se mostraram melhores opções do que a estratégia *single-GP*.

4
Estudo de Caso

Neste capítulo, será apresentado um estudo de caso abordando a classificação de atributos elásticos em poços com o objetivo de discriminar a litologia. Primeiramente, serão apresentados o escopo dos experimentos e a contextualização dos dados utilizados. Em seguida, serão apresentadas as técnicas de classificação empregadas baseadas em PG, além das configurações e parâmetros utilizados. Estas técnicas serão utilizadas para classificar a base de dados original, assim como, a base de dados estendida através da técnica de substituição de fluidos de Gassmann.

4.1.
Escopo dos experimentos e contextualização dos dados

Para este estudo de caso foram selecionados dados de três poços de uma mesma bacia da costa brasileira. Esses poços foram identificados como poço A, B e C. Todos os três poços atravessam um reservatório que apresenta acumulações de gás com porosidade média abaixo de 12% e permeabilidade média abaixo de $10^{-3}mD$, conhecidos como "*tight gas sands*". A identificação das seções do reservatório nos poços foi feita com base na porosidade obtida através dos perfis de densidade, sônico compressional e porosidade neutrônica, bem como no comportamento das leituras da ferramenta de ressonância magnética. Os perfis também mostraram a ocorrência de boa porosidade e de volume de fluido livre. Apesar de ser constatada a presença de gás, a análise de amostras laterais e testemunhos mostrou que o reservatório possuía baixa permeabilidade. A Figura 4.1 apresenta as seções do reservatório nos três poços utilizados.

Através da análise dos dados destes poços, também foram definidos valores de corte sobre algumas propriedades petrofísicas. Estes valores de corte auxiliaram na localização das regiões de interesse no reservatório, assim como na localização das acumulações de gás. As propriedades petrofísicas cujos valores de corte foram definidos são: o volume de argila (Vsh), a porosidade ($\emptyset$) e a

saturação de água (SW). Para o volume de argila foi definido o valor de corte de 40%, para a porosidade e a saturação de água foram definidos os valores de corte de 5% e 65%, respectivamente. Assim, as regiões dos poços de maior interesse foram aquelas em que os perfis mostraram estar abaixo dos valores de corte do volume de argila e da saturação de água, e acima do valor de corte da porosidade.

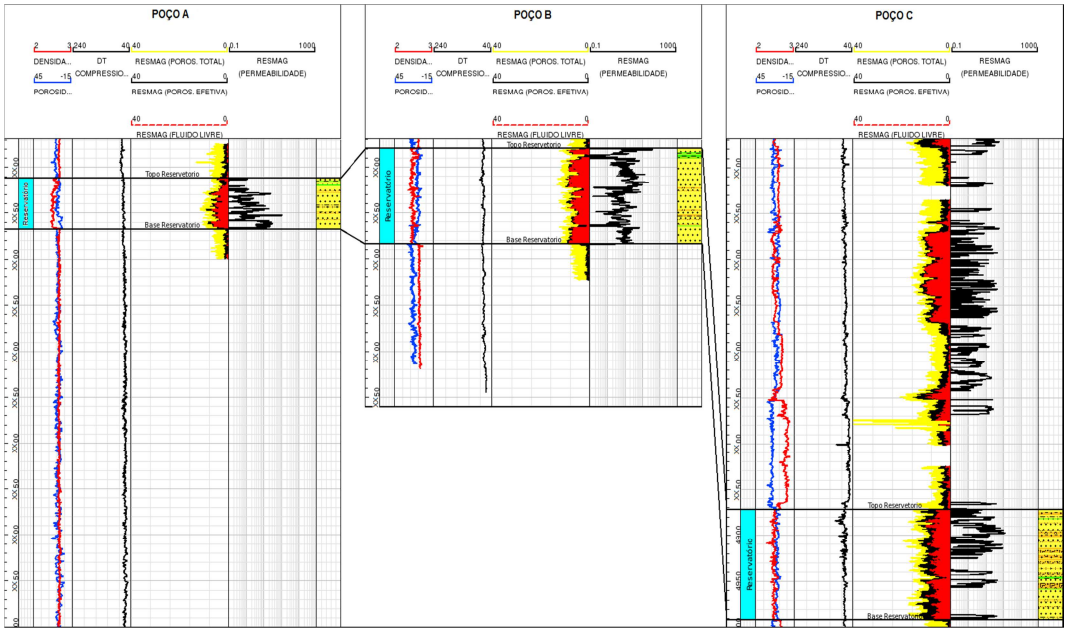


Figura 4.1 - Seções do reservatório identificadas nos poços A, B e C com os perfis Rhob, NPhi, DTC, TCMR, CMRP, CMFF e KSDR além da litologia.

No estudo de caso deste trabalho foi utilizado o próprio tipo de rocha como fácies de interesse para serem previstas através de atributos elásticos. A discriminação da litologia presente em um reservatório é uma das tarefas primordiais que auxiliam na construção do modelo geológico. Baseado na descrição de amostras laterais e de testemunhos, o intérprete identificou três tipos de rochas principais: conglomerado (CGL), arenito (ARN) e folhelho (FLH). O conglomerado e o arenito são rochas reservatório que podem conter acúmulo de gás e permeabilidade suficiente para a produção do mesmo. Já o folhelho não é uma rocha reservatório, devido a sua baixíssima permeabilidade. No entanto a sua identificação também é importante para a construção do modelo geológico.

Baseado no estudo integrado dos perfis de poços, das amostras laterais e dos testemunhos, o intérprete identificou ao longo das seções do reservatório nos poços qual era o tipo de rocha que estava presente dentre as três citadas anteriormente. Os dados também mostraram que onde foi identificado o conglomerado, apesar de ocorrer em menor quantidade, era observada uma maior

permeabilidade. Por se tratar de reservatórios do tipo "*tight gas sands*" esta característica relacionada a melhora da permeabilidade observada nos conglomerados fez dele um melhor candidato na busca pelos "*sweet-spots*" do reservatório. A Figura 4.2 apresenta uma seção com os três poços utilizados alinhados pelo topo do reservatório, onde são apresentados os principais perfis utilizados, além da identificação das zonas de gás (vermelho) e de água (azul) dado pelo valor de corte no perfil de saturação de água (SW).

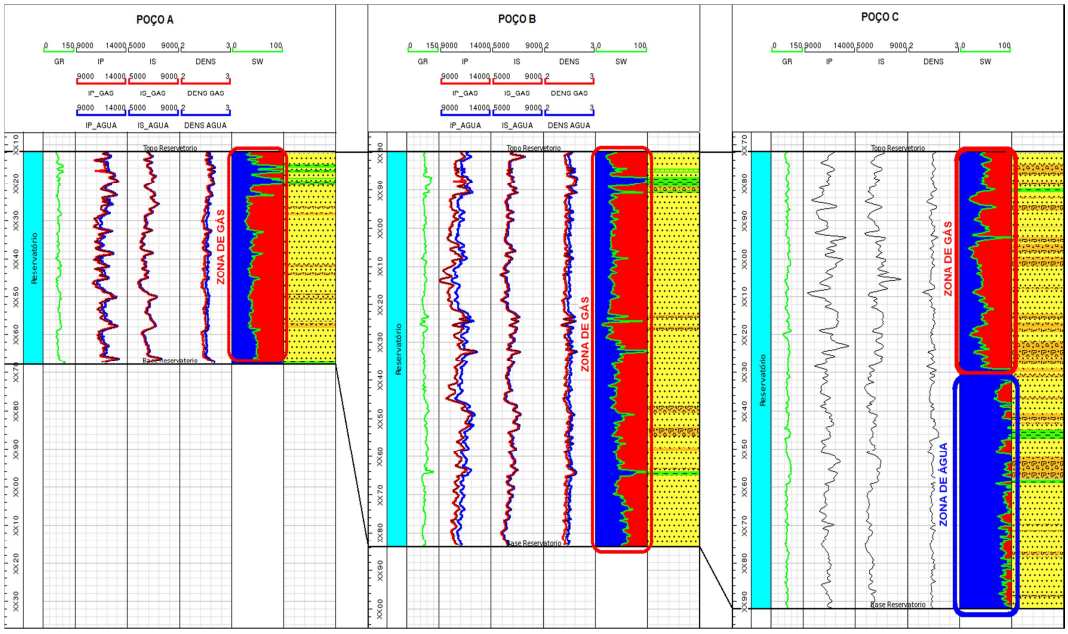


Figura 4.2 – Seção dos poços alinhados pelo topo do reservatório com os perfis GR, IP, IS, Rhob e SW além da litologia. Notar a identificação das zonas de gás (vermelho) e água (azul) no perfil de saturação de água (SW).

Os atributos elásticos utilizados nos experimentos foram as impedâncias elásticas compressional (IP) e cisalhante (IS). Estes atributos foram calculados através das equações (4.1) e (4.2) utilizando-se os perfis de poços sônico compressional (DT), sônico cisalhante (DTS) e densidade (ρ). Nas equações (4.1) e (4.2), o valor 304800 refere-se à conversão de unidades entre $ft/\mu s$ e m/s.

$$IP = \frac{304.800}{DT} \cdot \rho \tag{4.1}$$

$$IS = \frac{304.800}{DTS} \cdot \rho \tag{4.2}$$

Além de IP e IS, também foram utilizados dois outros atributos elásticos calculados a partir das impedâncias. O primeiro é a razão IP/IS e o segundo é a diferença $IP - IS$. Diferentes litologias geralmente apresentam diferentes

correlações entre velocidade compressional e velocidade cisalhante. Por exemplo, para os arenitos essa correlação é descrita por uma regressão linear, enquanto para os carbonatos essa correlação é quadrática. Atributos elásticos que refletem esta correlação, como a razão VP/VS ou IP/IS, reforçam estas diferenças, atuando eficientemente como discriminadores litológicos, enquanto atributos como $IP - IS$ são melhores discriminantes de fluido (Dillon, 2001). A utilização conjunta destes atributos permite uma interpretação mais quantitativa, reduzindo desta forma o risco exploratório e melhorando a caracterização de reservatórios.

4.2. Abordagens empregadas para a classificação

Três abordagens de classificação baseadas em programação genética foram desenvolvidas para realizar a identificação litológica a partir de atributos elásticos. As três abordagens empregadas foram baseadas nas seguintes propostas: Expressão Classificadora de Programação Genética (ECPG) (Kishore *et al.*, 2000), Programação Genética com Múltiplas Saídas (PGMS) (Brameier & Banzhaf, 2001) e Distribuição Gaussiana em Programação Genética (DGPG) (Zhang & Smart, 2006). A escolha foi feita baseada no tipo da abordagem que cada uma representava. A abordagem ECPG utiliza a técnica de decomposição binária, a abordagem DGPG adota a técnica de classificação por faixas de valores dinâmicos, enquanto a abordagem PGMS utiliza a abordagem de programação genética com múltiplas saídas onde um único indivíduo pode produzir múltiplas saídas.

Para efeito de comparação de resultados, também foram utilizados outras três técnicas de classificação, duas estatísticas, Bayes Ingênuo (BI) e Análise Discriminante Linear (DISC) e uma Rede Neural Perceptron de Múltiplas Camadas (RN). Todas as abordagens e experimentos foram efetuados no MATLAB R2008b (MATLAB, 2008), PC Windows 7 com processador Intel® Core™2 Duo, 4GB de RAM. Foi utilizada como parte do desenvolvimento das abordagens ECPG, PGMS e DGPG a biblioteca GPTIPS 1.0 (Searson, 2010) de programação genética, também no MATLAB.

Para medir o desempenho de cada abordagem, o poço C não foi utilizado como base de aprendizado, ficando separado para ser utilizado como base de teste.

Desta forma foi possível medir quantitativamente o desempenho de cada abordagem. A escolha do poço C como base de teste deveu-se ao fato deste poço ter atingido o contato gás-água do reservatório, como foi apresentado na Figura 4.2. Ou seja, neste poço, de acordo com o valor de corte definido para a saturação de água de 65%, foram observadas rochas saturadas tanto com gás como com água. No caso dos poços A e B, a quase totalidade das rochas estava saturada com gás. Esta situação foi interessante, pois permitiu avaliar a importância da análise de sensibilidade aos fluidos dos atributos elásticos através da substituição de fluidos de Gassmann. A Tabela 4.1 apresenta a tabulação do número de padrões por classe dos poços A e B, utilizados como base de aprendizado, e do poço C, que foi utilizado como base de teste.

Tabela 4.1 – Tabulação do número de padrões por classe dos poços A, B e C.

Litologia	Número de padrões Poços A e B	Número de padrões Poço C
CGL	65 (8,15%)	172 (28.67%)
ARN	687 (86,09%)	407 (67,83%)
FLH	46 (5,76%)	21 (3,50%)
TOTAL	798 (100,00%)	600 (100,00%)

Nas abordagens baseadas em PG e de redes neurais foi criado um conjunto de validação a partir dos dados dos poços A e B, no sentido de evitar o super treinamento (*overfitting*). Os padrões destes poços foram igualmente divididos (50%) entre os conjuntos de treinamento e validação. Este percentual igualitário entre a base de treinamento e validação foi sugerido em alguns trabalhos que aplicaram a programação genética para classificação (Kishore *et al.*, 2000)(Zhang & Smart, 2006)(Zhang & Wong, 2008). Nas abordagens baseadas em PG, o conjunto de validação foi utilizado ao final de cada geração para medir a aptidão do melhor indivíduo. Caso este possuísse uma aptidão melhor do que o melhor indivíduo da geração anterior, ele era selecionado como o melhor indivíduo do sistema.

Como pôde ser verificado na Tabela 4.1, o número de padrões para cada litologia encontra-se altamente desequilibrado. Devido a este fato, foi utilizada

uma replicação aleatória de dados para as classes com menos padrões com o objetivo de equilibrar os conjuntos de treinamento e validação. As abordagens estatísticas não utilizaram o conjunto de validação, porém utilizaram os mesmos conjuntos de treinamento e teste para efeito de comparação.

A Tabela 4.2 apresenta a tabulação dos conjuntos de treinamento, validação e teste utilizados nos experimentos.

Tabela 4.2 – Tabulação dos conjuntos de treinamento, validação e teste por classe utilizados nos experimentos.

Litologia	Conj. Treinamento	Conj. Validação	Conj. Teste
CGL	133 (33,33%)	133 (33,33%)	172 (28.67%)
ARN	133 (33,33%)	133 (33,33%)	407 (67,83%)
FLH	133 (33,33%)	133 (33,33%)	21 (3,50%)
TOTAL	399 (50,00% / 798)	399 (50,00% / 798)	600 (100,00%)

Outra técnica utilizada nas três abordagens baseadas em PG foi o chamado aprendizado incremental (*Incremental Learning ou step-wise learning*) (Kishore *et al.*, 2000)(Muni *et al.*, 2004). Nesta técnica, em vez de apresentar todos os padrões do conjunto de treinamento de uma só vez para a função de aptidão, só é utilizado um subconjunto desses dados denominado conjunto de avaliação. A cada determinado número de gerações, o tamanho do conjunto de avaliação é aumentado até que todos os padrões estejam contemplados. Dois parâmetros são necessários nesta técnica: a taxa de incremento do conjunto de avaliação e o número de gerações em que o incremento será aplicado.

Como estratégia de seleção foi utilizada a seleção por torneio. Nesta técnica um determinado número de indivíduos é aleatoriamente selecionado da população. O indivíduo com a melhor aptidão dentre os selecionados é retornado. Caso dois ou mais indivíduos possuíssem a mesma aptidão, a técnica de pressão lexicográfica foi utilizada como critério de desempate (Luke & Panait, 2002). Nesta técnica o indivíduo com menor número de nós era retornado. Caso um novo empate ocorresse, um indivíduo era aleatoriamente selecionado. Este e outros parâmetros utilizados nas três abordagens baseadas em PG são apresentados na Tabela 4.3.

Todas as configurações, parâmetros e técnicas descritos anteriormente foram igualmente utilizados nas três abordagens baseadas em PG, assim como nas abordagens de rede neural e estatísticas. Para a abordagem RN, a partir de vários experimentos preliminares, foi definida a seguinte configuração: uma camada escondida com sete neurônios (número de entradas + número de saídas) e a função de ativação tangente hiperbólica tanto na camada escondida como na camada de saída. Serão descritas a seguir as particularidades das três abordagens de classificação baseadas em PG.

Tabela 4.3 – Parâmetros e valores utilizados nas abordagens de PG para classificação.

Parâmetro	Valor	
Taxa de incremento do conjunto de avaliação	5%	
Número de gerações para aplicar o incremento	48	
Tamanho da população	250	
Número de gerações	1.000	
Estratégia de seleção	torneio	
Tamanho da seleção por torneio	7	
Desempate na seleção por torneio	Pressão lexicográfica	
Elitismo	5%	
Intervalo para a geração de constantes	[-100 a 100]	
Conjunto de funções	$(+, -, *, \div, \sqrt{x}, x^2, if \leq)$	
Taxa de mutação	10%	
Taxa de cruzamento	85%	
Taxa de reprodução direta	5%	
Método de criação dos indivíduos	*Ramped half-and-half*	
Profundidade máxima dos indivíduos	Inicial: 2	Final: 6
Número máximo de nós por indivíduo	Inicial: 4	Final: ∞
Profundidade máxima na mutação	Inicial: 2	Final: 6
Taxa de cruzamento alto nível (abordagem PGMS)	20%	
Taxa dos tipos de mutações	Mutação subárvore: 90%	
	Trocar um terminal por outro: 5%	
	Perturbação gaussiana em uma constante: 5% (dp: 0,1)	

4.2.1.
Expressão Classificadora de Programação Genética (ECPG)

Na abordagem ECPG, a classificação dos atributos elásticos foi feita através da técnica de decomposição binária. Desta forma, uma função discriminante foi evoluída para cada classe, exigindo que o sistema fosse executado três vezes. Foi utilizado um único valor limítrofe, zero, para separar os padrões da classe em evolução das demais. Ou seja, se o valor de saída da função discriminante fosse positivo, então o padrão era considerado como pertencente à classe em evolução, caso contrário o padrão era considerado como não pertencente a esta classe.

A abordagem de decomposição binária gera um desbalanceamento nos conjuntos de treinamento e validação. Os padrões que estavam igualmente divididos entre as três classes, $1/3$ para cada classe em cada conjunto, nesta abordagem foram reduzidos para apenas duas classes, resultando em $1/3$ dos padrões para a classe em evolução e $2/3$ para as outras. Para resolver este problema foi utilizado o formato de dados intercalado (Kishore *et al.*, 2000), onde os padrões da classe em evolução eram repetidos uma vez entre os padrões das outras duas classes. Por exemplo, se a função discriminante sendo evoluída correspondesse à classe conglomerado, os conjuntos de treinamento e validação ficariam com 266 padrões de conglomerado, 133 padrões de arenito e 133 padrões de folhelho.

A função de aptidão utilizada nesta abordagem foi a taxa de erros da classificação, ou seja, a razão entre o número de padrões incorretamente classificados e o número total de padrões utilizados na avaliação. O sistema foi configurado no sentido de minimizar este valor.

Na abordagem ECPG, quando mais de uma função discriminante aponta que um determinado padrão pertence a sua classe, uma técnica de resolução de conflitos é necessária. Como critério de resolução de conflito foi utilizada a métrica chamada *strength of association* (SA) (Kishore *et al.*, 2000). O SA indica o grau que cada função discriminante consegue reconhecer padrões pertencentes a sua classe e rejeitar padrões pertencentes às outras classes. Ele foi calculado ao final da execução do sistema utilizando-se o conjunto de treinamento. Todos os padrões eram apresentados a cada uma das três funções discriminantes. Se um padrão fosse ativado por uma função ele era contabilizado para a montagem da

matriz de contagem de classe, onde cada coluna corresponde a uma função e cada linha a uma classe. O SA era calculado a cada linha como a razão da diagonal principal da matriz pela soma total da linha. Em caso de conflito, a classe que apresentasse o maior SA era selecionada. A Tabela 4.4 apresenta um exemplo da matriz de contagem de classes para o cálculo do SA. A Tabela 4.5 apresenta o resultado do cálculo do SA de acordo com a Tabela 4.4, neste caso a sequência de prioridade das classes para a resolução dos conflitos seria: 1º arenito, 2º conglomerado e 3º folhelho.

Tabela 4.4 – Exemplo da matriz de contagem de classe para o cálculo do *strength of association* (SA).

Classe/Função	Função CGL	Função ARN	Função FLH
CGL	99	11	7
ARN	9	101	5
FLH	13	22	88

Tabela 4.5 – Resultado do cálculo do SA a partir da matriz de contagem de classe da Tabela 4.4.

Classe	CGL	ARN	FLH
SA	0,84	0,87	0,71

Além do SA, também foram utilizados na abordagem ECPG regras heurísticas para a resolução de conflitos. Essas regras heurísticas são produzidas a partir de padrões do conjunto de treinamento que foram incorretamente classificados pelo SA. Por exemplo, se os valores do SA na Tabela 4.5 fossem utilizados para resolver os conflitos entre as funções das classes conglomerado e arenito, todos os padrões seriam considerados da classe arenito, pois ela possui maior SA. Porém, caso fosse observado um determinado número de padrões no conjunto de treinamento, que apresentavam este conflito e pertencessem à classe conglomerado, a seguinte regra heurística era adotada: se houver conflitos entre padrões de arenito e conglomerado, então a classe selecionada é o conglomerado. O número de padrões necessários para gerar uma regra heurística foi de 10% do número de padrões da classe.

4.2.2.
Programação Genética com Múltiplas Saídas (PGMS)

Na abordagem PGMS, a classificação dos atributos elásticos foi baseada na técnica de programação genética com múltiplas saídas. Na proposta original (Brameier & Banzhaf, 2001) cada indivíduo era um código-fonte representando um programa na linguagem C. Cada programa possuía um vetor de saída contendo uma posição para cada classe a ser prevista. A classe prevista era aquela que apresentava o maior valor no vetor de saída. Para desenvolver esta abordagem no contexto onde cada indivíduo é uma árvore sintática, foram utilizados indivíduos da programação genética com múltiplos genes (PGMG) (Hinchliffe *et al.*, 1996). A utilização de indivíduos com múltiplas árvores, em detrimento de um único segmento linear como era a proposta original, foi motivada pela capacidade de interpretação direta das regras de classificação sob a forma de funções discriminantes.

Nesta abordagem, cada indivíduo possui várias árvores, cada uma representando um gene. A Figura 4.3 apresenta um indivíduo de programação genética com múltiplos genes com um total de três genes.

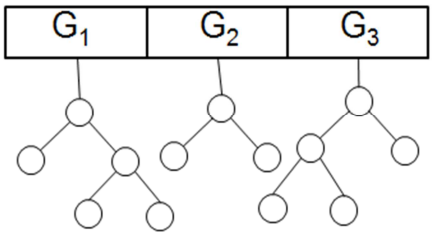


Figura 4.3 – Indivíduo da programação genética com múltiplos genes (3 genes).

Todos os indivíduos utilizados nesta abordagem possuíam exatamente três genes, cada um representando uma classe. Na verdade, um indivíduo da programação genética clássica pode ser visto como um caso específico de indivíduo da PGMG com apenas um gene. Todos os operadores genéticos aplicados aos indivíduos da programação genética clássica, também podem ser aplicados aos indivíduos da PGMG. A única diferença foi a utilização do operador de cruzamento de alto nível. Neste operador um conjunto de genes de um indivíduo é trocado com outro conjunto de genes de outro indivíduo. Para manter os indivíduos com o mesmo número de genes, os pontos de corte foram mantidos simétricos. A Figura 4.4 apresenta o operador de cruzamento de alto nível em dois

indivíduos da PGMG. Neste caso, o segundo gene foi trocado entre os indivíduos 1 e 2.

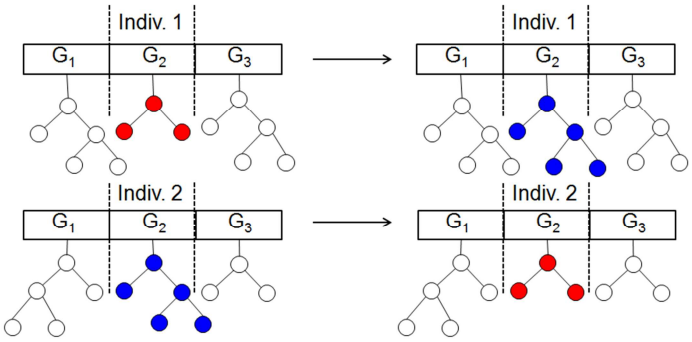


Figura 4.4 – Operador de cruzamento de alto nível em indivíduos da PGMG.

Nesta abordagem, um padrão era corretamente classificado quando o gene correspondente a sua classe retornava um valor maior que todos os outros genes (*winner-takes-all*). Na proposta original, a função de aptidão era calculada pela equação (4.3), ou seja, a soma entre a taxa de erros da classificação (MCE) e o erro médio quadrático (MSE). O MSE era calculado entre o valor de saída de cada função discriminante e o valor 1, caso a função correspondesse à classe correta, e -1, caso a classe fosse a incorreta. A intenção do uso do MSE era dar um caráter mais contínuo à função de aptidão.

$$F(p) = MCE + MSE \tag{4.3}$$

Durante os experimentos, observou-se que o MSE tornava-se o critério preponderante na função de aptidão devido aos seus valores elevados em relação ao MCE. Na proposta original este problema provavelmente não ocorreu, pois os valores dos atributos foram normalizados antes de serem utilizados pelo sistema. Devido a estes fatores, neste estudo de caso o MSE foi retirado da função de aptidão, permanecendo apenas a taxa de erros da classificação. O MSE foi introduzido na seleção por torneio como critério de desempate antes da pressão lexicográfica.

4.2.3. Distribuição Gaussiana em Programação Genética (DGPG)

Na abordagem DGPG, a classificação dos atributos elásticos foi baseada na distribuição gaussiana dos valores de saída de cada classe. Desta maneira, uma

única função discriminante é evoluída para classificar todas as classes. Diferentemente das propostas que utilizam valores limítrofes fixos para cada classe, na abordagem DGPG esses limites são evoluídos juntamente com a função discriminante (Zhang & Smart, 2006).

Para o cálculo da função de aptidão o seguinte procedimento era realizado. Primeiramente, a função discriminante era utilizada para calcular o valor de saída para todos os padrões do conjunto de treinamento. Esses valores eram agrupados de acordo com as suas respectivas classes. Em seguida, eram calculados a média e o desvio padrão para cada agrupamento. Assim, através da média e do desvio padrão eram definidas uma distribuição gaussiana para cada classe. A Figura 4.5 apresenta o gráfico com as distribuições gaussianas para o conglomerado, arenito e folhelho através da função discriminante $IS - 862{,}1 * IP + 1172{,}4$ e os dados dos poços A e B. No gráfico da Figura 4.5 os valores de saída da função discriminante foram normalizados.

Após a criação das distribuições gaussianas, era calculada a distância ponderada entre as distribuições através da equação (4.4). Nesta equação, m_1 e m_2 são as médias enquanto dp_1 e dp_2 são os desvios padrão da distribuição de duas classes. Essa distância era normalizada pela equação (4.5) para que seu valor, no limite, variasse entre zero e um. Como o problema do estudo de caso possuía três classes, a função de aptidão foi calculada pelo somatório das distâncias ponderadas duas a duas, totalizando três fatores. A equação (4.6) apresenta a equação da função de aptidão.

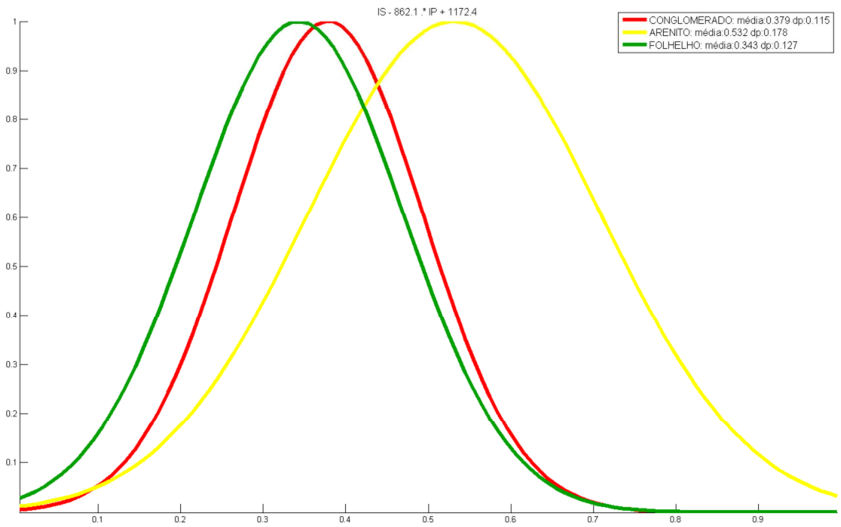


Figura 4.5 – Distribuições gaussianas de cada classe para a função discriminante $\boldsymbol{IS - 862{,}1 * IP + 1172{,}4}$. Os valores de saída da função foram normalizados.

$$d = 2 * \frac{|m_1 - m_2|}{dp_1 + dp_2} \tag{4.4}$$

$$d_n = \frac{1}{1 + d} \tag{4.5}$$

$$f(p) = \sum_{i=1}^{3} d_n \tag{4.6}$$

O objetivo da função de aptidão era maximizar a distância ponderada entre as distribuições. Funções discriminantes que levavam a uma maior distância entre as médias das distribuições e a menores desvios padrões eram favorecidas na evolução do sistema. Com a normalização feita pela equação (4.5), o caso ótimo da função de aptidão seria o valor zero, quando as distâncias tenderem ao infinito. Por isso o sistema foi configurado com o objetivo de minimizar a função de aptidão. O pior caso será quando as distribuições possuírem a mesma média ou possuírem valores de desvios padrões tendendo ao infinito, neste caso a função de aptidão terá o seu valor máximo de uma unidade.

Para medir a qual classe um dado padrão pertence, foram utilizados os cinco melhores indivíduos evoluídos da população em vez de utilizar apenas o melhor indivíduo. Assim, a probabilidade de um dado padrão pertencer a uma determinada classe c foi calculada pela equação (4.7), onde P é a função densidade de probabilidade da distribuição gaussiana (equação (4.8)), s_i é a saída da função discriminante i para o padrão a ser classificado, m_i e dp_i são respectivamente a média e o desvio padrão de cada função discriminante para cada classe c. Baseado na equação (4.7), a classe que possuía a maior probabilidade era selecionada como a classe do padrão avaliado.

$$Prob_c = \prod_{i=1}^{5} P(m_i, dp_i, s_i) \tag{4.7}$$

$$P(m, dp, s) = \frac{exp\left(\frac{-(s-m)^2}{2 * dp^2}\right)}{dp * \sqrt{2\pi}} \tag{4.8}$$

4.3. Apresentação e análise dos resultados

Nesta seção, serão apresentados e analisados os resultados encontrados pela classificação dos atributos elásticos para a discriminação litológica utilizando-se as abordagens ECPG, PGMS, DGPG, BI, DISC e RN. Primeiramente, serão apresentados os resultados alcançados utilizando-se apenas os valores dos atributos elásticos ($IP, IS, IP/IS \ e \ IP - IS$) calculados diretamente a partir das medidas dos perfis dos poços A e B. Em seguida, serão apresentados os resultados alcançados com os valores dos atributos elásticos estendidos através da substituição de fluidos de Gassmann em duas situações: a primeira da condição *in situ* do reservatório para a saturação de 100% de água; a segunda da condição *in situ* para a saturação de 100% de gás.

Como critério de comparação foi utilizado o percentual de acerto na classificação das litologias do poço C, ou seja, o número de padrões corretamente classificados divididos pelo total de padrões. Para as abordagens ECPG, PGMS, DGPG e RN um total de 100 execuções foram realizadas todas com a mesma semente para a geração de números aleatórios. Assim, foi possível medir dentre as 100 execuções o maior percentual de acerto, a média do percentual de acerto assim como o seu desvio padrão. Este procedimento foi adotado, pois assim como os pesos iniciais de uma rede neural são selecionados aleatoriamente, a população inicial na programação genética também depende de escolhas feitas aleatoriamente a partir do conjunto de primitivas. Esta situação inicial pode afetar os resultados alcançados.

A Tabela 4.6 apresenta os resultados da classificação a partir dos valores originais dos atributos elásticos para discriminar as litologias no poço C. É possível observar que os melhores classificadores produzidos pelas abordagens baseadas em PG conseguiram atingir os melhores percentuais de acertos. A abordagem DGPG foi a que obteve o maior percentual de acerto (72,00%), seguido pela abordagem PGMS (69,33%) e pela abordagem ECPG (67,83%). Todas elas atingiram resultados superiores a abordagem DISC (56,83%), que é a utilizada pelo intérprete. Um dos fatores que levaram a um melhor desempenho da abordagem DGPG foi a utilização dos cinco melhores indivíduos encontrados ao final da execução do sistema, em vez de apenas o melhor.

Tabela 4.6 – Resultados da classificação a partir dos valores originais dos atributos elásticos para a litologia do poço C.

Abordagem	% maior acerto	% médio acerto	desvio padrão
DISC	56,83	56,83	-
DGPG	72,00	58,93	8,93
ECPG	67,83	33,20	15,03
PGMS	69,33	35,99	17,10
RN	58,67	51,31	5,43
BI	55,50	55,50	-

Em relação ao percentual de acerto médio dos 100 experimentos, pode-se observar que a abordagem DGPG também obteve o melhor desempenho (58,93%), além de apresentar o segundo menor desvio padrão (8,93%). As abordagens estatísticas (DISC e BI) não apresentam desvio padrão, pois atingem o mesmo resultado em todas as execuções. Novamente, o uso dos melhores cinco indivíduos na abordagem DGPG contribuiu para que a taxa de acertos permanecesse elevada entre os vários experimentos. O uso da distribuição gaussiana na função de aptidão também se mostrou muito efetivo neste tipo de problema, pois os atributos elásticos sempre se encontram influenciados por mais de uma propriedade de rocha simultaneamente, o que leva a uma grande sobreposição de valores.

Por outro lado, as abordagens ECPG e PGMS foram as que alcançaram as piores médias, respectivamente 33,20% e 35,99%. Além disso, essas duas abordagens foram as que apresentaram os maiores desvios padrões (15,03% e 17,10%). O que pode explicar este baixo resultado médio das abordagens ECPG e PGMS é que ambas utilizam a taxa de erros da classificação como função de aptidão. Foi observado nos experimentos, que nas gerações iniciais dessas duas abordagens, a grande maioria dos indivíduos possuía exatamente a mesma taxa de erros da classificação. Nestes casos, a seleção por torneio ficava a cargo do critério de desempate, a pressão lexicográfica no caso do ECPG e o MSE no caso da abordagem PGMS. O uso excessivo desses critérios não levava a um bom classificador, por não estarem diretamente ligados ao problema. A Figura 4.6 apresenta um gráfico do comportamento típico da função de aptidão na

abordagem ECPG. Os valores em azul representam a aptidão do melhor indivíduo no conjunto de treinamento, os valores em preto representam a aptidão deste mesmo indivíduo no conjunto de validação. É possível observar a estacionaridade ao longo de muitas gerações.

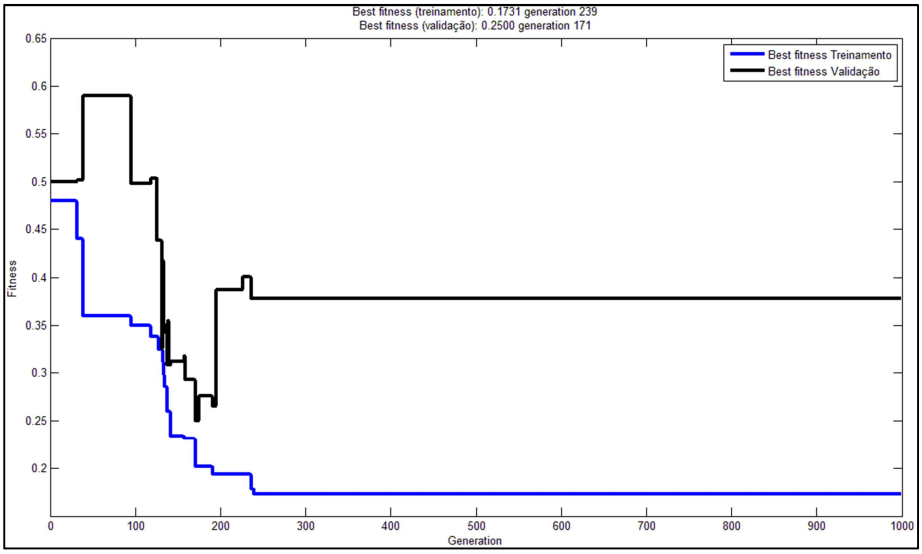


Figura 4.6 – Gráfico mostrando o comportamento da função de aptidão ao longo das gerações da abordagem ECPG.

Em uma segunda rodada de experimentos, a base de dados original foi estendida através da substituição de fluidos de Gassmann. O objetivo foi a geração de dados teóricos para os atributos elásticos, que simulassem outras condições em relação à saturação de fluidos que não foram encontradas nos poços A e B. Na Figura 4.2 é possível observar, através do perfil de saturação de água, que os poços A e B possuem rochas majoritariamente saturadas de gás, não havendo amostras de rochas reservatório saturadas majoritariamente por água. Assim, dois procedimentos foram adotados: o primeiro foi o cálculo a partir da condição *in situ* do reservatório para uma saturação de 100% de água; o segundo também se deu a partir da condição *in situ* do reservatório, porém, neste caso, o objetivo era simular a condição de saturação em 100% de gás.

Primeiramente, as velocidades compressional (VP) e cisalhante (VP) foram calculadas a partir dos perfis sônico compressional (DT) e sônico cisalhante (DTS) pelas equações (4.9) e (4.10). Em seguida, foram calculados os módulos bulk (K_{sat1}) e cisalhante (μ_{sat1}) a partir das velocidades e do perfil de densidade.

$$VP = \frac{304800}{DT} \tag{4.9}$$

$$VS = \frac{304800}{DTS} \tag{4.10}$$

Para calcular os novos módulos bulk (K_{sat2}) das rochas 100% saturadas por água ou por gás, a equação de Gassmann foi utilizada com as seguintes constantes obtidas para este reservatório: $K_{mineral} = 40GPa$, $K_{água} = 2{,}7GPa$ e $K_{gás} = 0{,}07GPa$. Além das constantes, também foi utilizado o perfil de porosidade na equação de Gassmann. O fluido encontrado nos poços A e B, apesar de apresentar maior saturação de gás, é uma mistura entre água e gás. Por esse motivo, o módulo bulk do fluido *in situ* $(K_{fluido1})$ depende dos módulos bulk da água e do gás, além de suas respectivas saturações. Ele foi calculado através da equação (4.11). O mesmo se aplica à densidade do fluido *in situ* $(\rho_{fluido1})$ que também depende das densidades da água e do gás, além de suas respectivas saturações. Esta densidade do fluido no reservatório foi calculada através da equação (4.12).

$$K_{fluido1} = \frac{1}{\frac{SW}{K_{água}} + \frac{(1-SW)}{K_{gás}}} \tag{4.11}$$

$$\rho_{fluido1} = SW.\rho_{água} + (1-SW).\rho_{gás} \tag{4.12}$$

Após o cálculo dos novos módulos bulk através da equação de Gassmann, foram calculadas as novas densidades da rocha (ρ_2). Estas novas densidades dependem da densidade original da rocha medida pelo perfil de densidade, além da densidade do fluido *in situ*, do fluido final (100% água ou 100% gás) e da porosidade. Após o cálculo dos novos módulos bulk e das novas densidades, foi possível calcular as novas velocidades compressional (VP_2) e cisalhante (VS_2). O módulo cisalhante permaneceu inalterado visto que ele não é sensível ao fluido. Após o calculo das novas velocidades, foram calculadas os novos valores para as impedâncias compressional (IP) e cisalhante (IS) através das equações (4.13) e (4.14).

$$IP_2 = VP_2 \,.\, \rho_2 \tag{4.13}$$

$$IS_2 = VS_2 \,.\, \rho_2 \tag{4.14}$$

A Figura 4.7 apresenta dois *cross-plots* entre IP e IS, mostrando os valores para cada litologia na condição *in situ* e após a substituição de fluidos de Gassmann para a saturação de 100% água (acima) e para a saturação de 100% gás (abaixo). É importante salientar que o procedimento só foi aplicado aos conglomerados e arenitos, permanecendo inalterados os valores para os folhelhos. É possível verificar no gráfico de $IP\ x\ IS$ para a saturação de 100% água o forte aumento nos valores da impedância compressional (IP). Este aumento de IP se dá devido ao aumento no valor do módulo bulk, assim como o aumento da densidade devido à presença de água. A impedância cisalhante (IS) apresenta um pequeno aumento, unicamente devido ao aumento da densidade visto que o módulo cisalhante não é sensível aos fluidos. No gráfico de $IP\ x\ IS$ que apresenta os valores para a saturação de 100% gás houve uma pequena queda nos valores da impedância compressional (IP) e cisalhante (IS). Esse resultado era esperado, pois nos poços A e B as rochas já estavam saturadas majoritariamente por gás. A mudança para a situação de saturação de 100% gás altera pouco a condição *in situ*.

A partir da geração desses novos valores calculados teoricamente, duas novas bases foram criadas. A primeira uniu os valores originais das impedâncias na condição *in situ* com os valores na saturação de 100% água (BD ÁGUA). A segunda uniu os valores originais na condição *in situ* com os valores na saturação de 100% água, além dos valores na condição 100% gás (BD ÁGUA e GÁS). Para a geração destas novas bases, foi utilizado o procedimento descrito no fluxo de trabalho estatístico e de física de rochas de Avseth *et al.* (2005). Nesta abordagem, os atributos elásticos teoricamente calculados são adicionados aos dados originais através da simulação de Monte Carlo. Estas simulações de Monte Carlo devem ser projetadas a partir da distribuição de cada fácies. No presente estudo de caso, as fácies são cada uma das três litologias a serem discriminadas: conglomerado, arenito e folhelho.

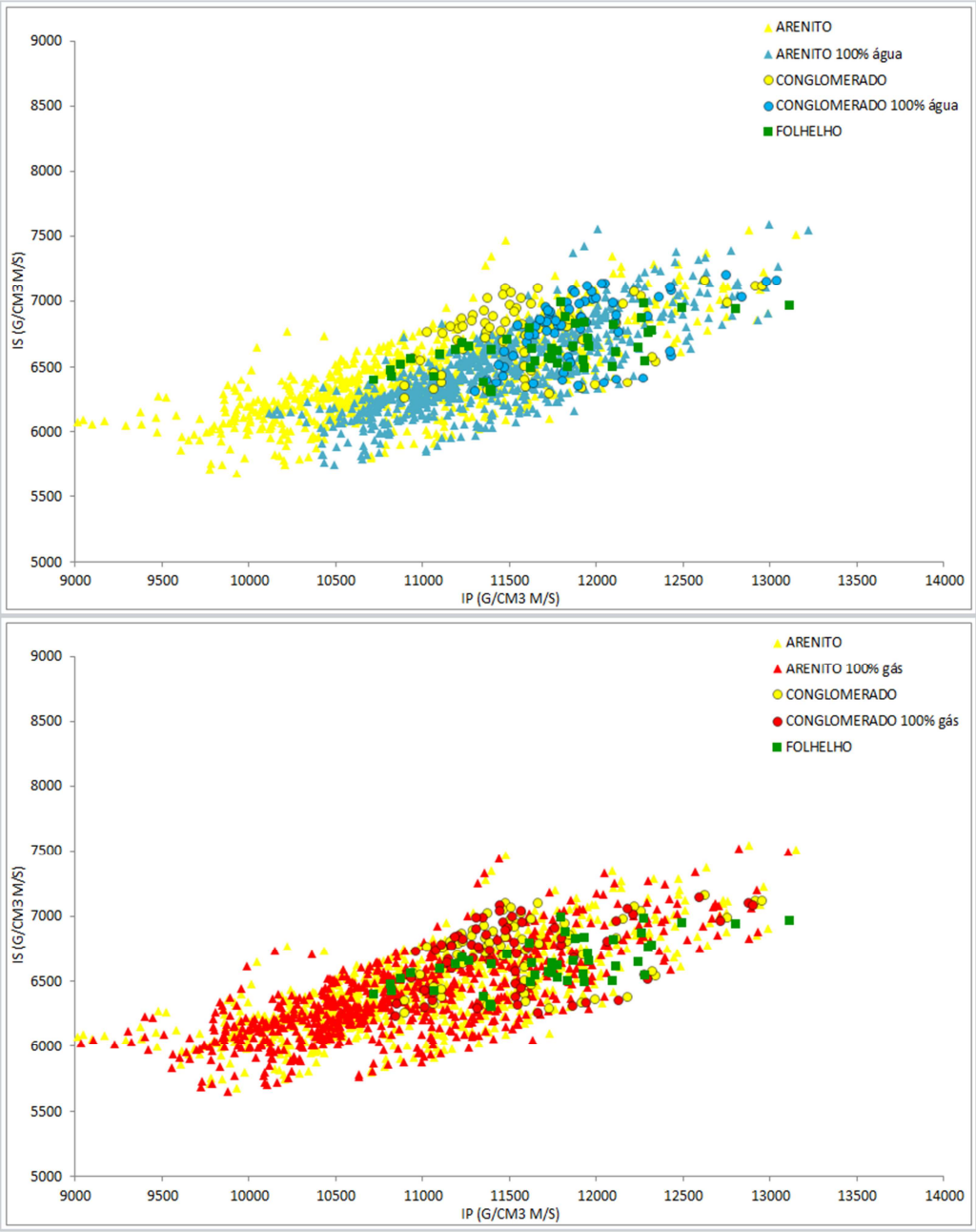


Figura 4.7 – C*ross-plots* entre ***IP*** e ***IS*** mostrando os valores para cada litologia na condição *in situ* (acima e abaixo), para a saturação de 100% de água (acima) e para a saturação de 100% gás (abaixo).

O procedimento para a criação das novas bases pela simulação de Monte Carlo se deu da seguinte maneira. Primeiramente, foram calculadas as funções de distribuição acumulada a partir da distribuição conjunta de *IP* e *IS* da base original para cada litologia. Em seguida, o número de padrões que seriam adicionados à base original era estipulado. Neste caso, foi definido o mesmo número de padrões que a base original possuía, ou seja, 798 padrões. Esta quantidade adicional de padrões foi proporcionalmente dividida entre as três

litologias de acordo com os percentuais da base original: 8,15% para o conglomerado, 86,09% para os arenitos e 5,76% para os folhelhos. Para cada quantidade de novos padrões o procedimento descrito a seguir foi repetido. Um número aleatório entre 0 e 1 era gerado, possibilitando a seleção de um par de valores de *IP* e *IS* na base original, de acordo com a função de distribuição acumulada de cada litologia. Para cada par de valores de *IP* e *IS* selecionado, o novo valor calculado pela substituição de fluidos de Gassmann era inserido na nova base. A Figura 4.8 apresenta os gráficos das funções de distribuição acumulada de *IP* para cada litologia. Cada ponto azul nos gráficos corresponde a um novo padrão inserido nas bases de dados estendida pela substituição de fluidos (*IS* foi retirado apenas para facilitar a visualização).

A Tabela 4.7 apresenta a tabulação das novas bases BD ÁGUA e BD ÁGUA e GÁS. Embora o número de padrões escolhido fosse igual ao da base original (798), qualquer quantidade poderia ter sido escolhida. Após a criação das novas bases pela simulação de Monte Carlo, os novos conjuntos de treinamento e validação foram criados, novamente pela replicação aleatória de dados para as classes com menos padrões. A Tabela 4.8 e a Tabela 4.9 apresentam as tabulações dos novos conjuntos de treinamento e validação. Os dados do poço C permaneceram sendo utilizados como conjunto de teste.

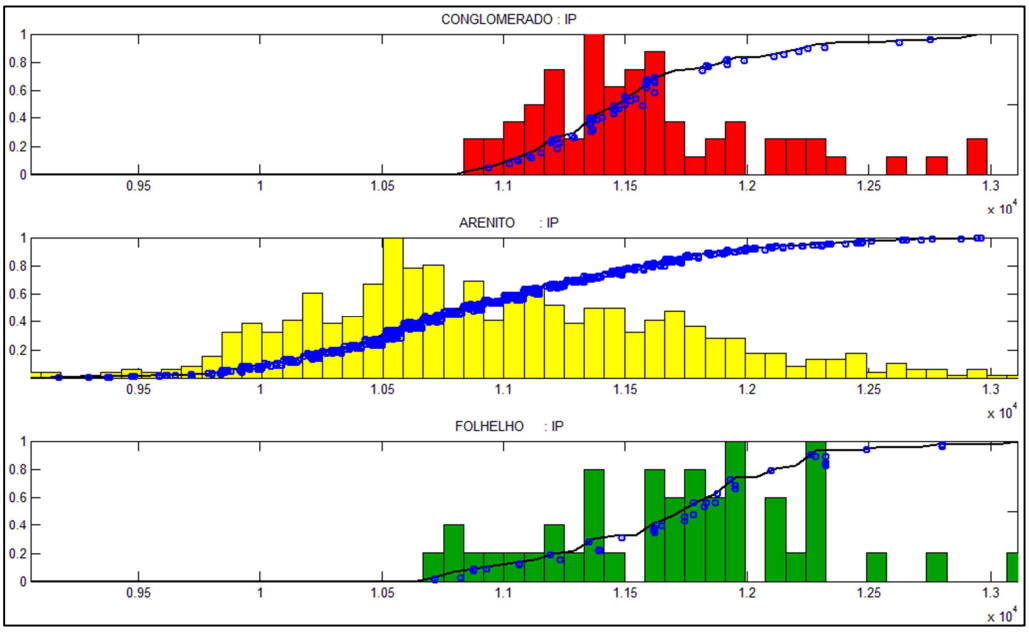


Figura 4.8 – Gráficos das funções de distribuição acumulada de ***IP*** para cada litologia. Cada ponto azul nos gráficos corresponde a um novo padrão inserido nas bases de dados estendida pela substituição de fluidos.

Tabela 4.7 - Tabulação do número de padrões por classe nas bases BD ÁGUA e BD ÁGUA e GÁS.

Litologia	Número de padrões BD ÁGUA	Número de padrões BD ÁGUA e GÁS
CGL	130 (8,15%)	195 (8,15%)
ARN	1374 (86,09%)	2061 (86,09%)
FLH	92 (5,76%)	138 (5,76%)
TOTAL	1596 (100,00%)	2394 (100,00%)

Tabela 4.8 - Tabulação dos conjuntos de treinamento, validação e teste por classe para a BD ÁGUA.

Litologia	Conj. Treinamento	Conj. Validação	Conj. Teste
CGL	266 (33,33%)	266 (33,33%)	172 (28.67%)
ARN	266 (33,33%)	266 (33,33%)	407 (67,83%)
FLH	266 (33,33%)	266 (33,33%)	21 (3,50%)
TOTAL	798 (50,00% / 1596)	798 (50,00% / 1596)	600 (100,00%)

Tabela 4.9 - Tabulação dos conjuntos de treinamento, validação e teste por classe para a BD ÁGUA e GÁS.

Litologia	Conj. Treinamento	Conj. Validação	Conj. Teste
CGL	399 (33,33%)	399 (33,33%)	172 (28.67%)
ARN	399 (33,33%)	399 (33,33%)	407 (67,83%)
FLH	399 (33,33%)	399 (33,33%)	21 (3,50%)
TOTAL	1197 (50,00%/2394)	1197 (50,00%/2394)	600 (100,00%)

A Tabela 4.10 apresenta os resultados da classificação a partir dos atributos elásticos da base BD ÁGUA para discriminar as litologias do poço C. Assim como na base original, os melhores classificadores produzidos pelas abordagens baseadas em PG conseguiram atingir os melhores percentuais de acertos. Neste caso, a abordagem que obteve o melhor percentual de acerto foi a PGMS (75,33%), seguida da abordagem ECPG (72,50%) e pela abordagem DGPG

(71,50%). Todas as abordagens conseguiram aumentar os seus percentuais de acertos utilizando-se a base BD ÁGUA. A única exceção foi a abordagem DGPG que manteve praticamente o mesmo percentual de acerto da base original (72,00% e 71,50%). As abordagens estatísticas (DISC e BI) e a abordagem RN foram as que alcançaram os maiores incrementos no percentual de acertos da base original para a base BD ÁGUA. Pode-se inferir que, talvez, estas abordagens dependam de uma maior quantidade de dados para atingir um melhor desempenho, enquanto as abordagens baseadas em PG conseguem obter melhores percentuais de acerto com uma menor quantidade de dados.

Tabela 4.10 - Resultados da classificação a partir dos atributos elásticos da base BD ÁGUA para as litologias do poço C.

Abordagem	% maior acerto	% médio acerto	desvio padrão
DISC	64,67	64,67	0,00
DGPG	71,50	61,58	10,74
ECPG	72,50	55,12	13,10
PGMS	75,33	54,92	12,81
RN	69,50	62,81	1,86
BI	66,17	66,17	0,00

Em relação ao percentual de acerto médio dos 100 experimentos, pode-se observar que a abordagem DGPG atingiu um resultado comparável ao da abordagem RN (61,58% e 62,81%). Porém, diferentemente da base original, a abordagem DGPG na média não conseguiu alcançar o percentual de acerto das abordagens estatísticas. Uma possível explicação para este fato pode ser o procedimento utilizado para a construção da base BD ÁGUA. Como a simulação de monte carlo seleciona padrões baseado na distribuição e nas probabilidades dos dados, as abordagens estatísticas podem ser mais beneficiadas, pois realizam a classificação baseada na análise dessas distribuições e nas probabilidades.

A Tabela 4.11 apresenta os resultados da classificação a partir dos atributos elásticos da base BD ÁGUA e GÁS para discriminar as litologias do poço C. Assim como na base original e na base BD ÁGUA, os melhores classificadores produzidos pelas abordagens baseadas em PG conseguiram atingir os melhores

percentuais de acertos. Neste caso, a abordagem que obteve o melhor percentual de acerto foi a DGPG (72,83%), seguida da abordagem PGMS (72,00%) e pela abordagem ECPG (68,00%). Neste caso, apesar de todas as abordagens terem apresentado melhora nos percentuais de acerto em relação à base original, o desempenho ficou pior em relação à base BD ÁGUA. A única exceção foi a abordagem DGPG que manteve praticamente o mesmo percentual de acerto nas três bases de dados (72,00%, 71,50% e 72,83%).

Tabela 4.11 - Resultados da classificação a partir dos atributos elásticos da base BD ÁGUA e GÁS para as litologias do poço C.

Abordagem	% maior acerto	% médio acerto	desvio padrão
DISC	62,67	62,67	0,00
DGPG	72,83	68,41	4,24
ECPG	68,00	57,65	10,77
PGMS	72,00	57,72	13,06
RN	67,83	63,76	1,24
BI	64,67	64,67	0,00

A diferença entre a base BD ÁGUA e a base BD ÁGUA e GÁS foram os valores calculados para as impedâncias, a partir da substituição de fluidos de Gassmann, para as rochas saturadas por 100% de gás. Porém, como visto anteriormente, os poços A e B já se encontravam majoritariamente saturados por gás. Por isso, esses novos valores calculados se diferenciaram muito pouco dos valores originais. Então, uma possível explicação para este pior desempenho da base BD ÁGUA e GÁS em relação à base BD ÁGUA, é que estes novos valores podem ter sido encarados como dados ruidosos e ao tentar se adaptar a esses novos dados, as diferentes abordagens perderam em generalização. A única abordagem que não foi afetada por esta possível inserção de dados ruidosos foi a DGPG.

Apesar dos melhores classificadores produzidos pelas abordagens PGMS e ECPG na base BD ÁGUA e GÁS, não terem alcançado o desempenho dos melhores classificadores na base BD ÁGUA, o percentual médio de acertos alcançado entre os 100 experimentos foi superior. A abordagem DGPG foi a que

obteve o maior ganho, de 61,58% na base BD ÁGUA para 68,41% na base BD ÁGUA e GÁS. A abordagem DGPG, além de produzir o melhor classificador para a base BD ÁGUA e GÁS, também foi a melhor abordagem no percentual médio de acertos entre os 100 experimentos. A Tabela 4.12 apresenta todos os resultados alcançados a partir da classificação dos atributos elásticos nas três bases utilizadas. Os melhores resultados estão destacados em negrito.

Tabela 4.12 - Resultados da classificação a partir dos atributos elásticos nas três bases utilizadas. Em negrito são destacados os melhores resultados.

Abordagem	BD ORIGINAL		BD ÁGUA		BD ÁGUA e GÁS	
	% maior acerto	% médio acerto	% maior acerto	% médio acerto	% maior acerto	% médio acerto
DISC	56,83	56,83	64,67	64,67	62,67	62,67
DGPG	**72,00**	**58,93**	71,50	61,58	**72,83**	**68,41**
ECPG	67,83	33,20	72,50	55,12	68,00	57,65
PGMS	69,33	35,99	**75,33**	54,92	72,00	57,72
RN	58,67	51,31	69,50	62,81	67,83	63,76
BI	55,50	55,50	66,17	**66,17**	64,67	64,67

Quando se utiliza abordagens baseadas em PG, um importante ponto de atenção é o tempo de execução do sistema. A Tabela 4.13 apresenta o tempo médio gasto dentre os 100 experimentos realizados para cada uma das bases. É possível observar que a abordagem que apresentou os menores tempos médios de execução foi a DGPG. Este resultado era previsto, pois nesta abordagem apenas uma função discriminante é evoluída para todas as classes. Também é possível verificar que o tempo médio de execução das abordagens para as diferentes bases não apresentou incremento significativo, apesar do número de padrões ter duplicado e triplicado. A característica de cálculo por matrizes do MATLAB pode ter contribuído para esta manutenção no patamar dos tempos entre as bases.

Tabela 4.13 - Tempo médio de execução das abordagens baseadas em PG.

Abordagem	BD ORIGINAL	BD ÁGUA	BD ÁGUA e GÁS
DGPG	06min 33s	06min 21s	07min 02s
ECPG	15min 40s	16min 47s	18min 20s
PGMS	24min 30s	23min 49s	24min 35s

A abordagem PGMS, apesar de necessitar de uma única execução para evoluir as funções discriminantes de cada classe, foi a que apresentou os maiores tempos médios. A utilização de indivíduos da PGMG, levando a um aumento na complexidade dos operadores genéticos e da própria representação dos indivíduos, e o cálculo das saídas de todas as funções discriminantes numa mesma execução da função de aptidão, além do cálculo do MSE para todos os indivíduos, podem explicar este maior tempo médio de execução desta abordagem.

A abordagem ECPG, em que o sistema era executado uma vez para cada classe, surpreendentemente não apresentou os maiores tempos médios de execução. A sua simples função de aptidão, que calculava apenas a taxa de erros da classificação, pode ter contribuído para que ela não atingisse tempos superiores. Porém, esta abordagem é a que apresenta o maior potencial de crescimento do tempo médio de execução visto que ela depende do número de classes envolvidas no problema.

5
Conclusões e Trabalhos Futuros

Esta dissertação abordou a utilização da programação genética como modelo classificador de atributos elásticos para a discriminação litológica. Esta classificação foi empregada como parte integrante do fluxo de trabalho estatístico e de física de rochas (Avseth *et al.*, 2005). O princípio que norteou o uso da programação genética foi a sua habilidade de seleção automática de atributos, além de permitir a interpretação do classificador.

Foram apresentados alguns conceitos da teoria da elasticidade e da teoria de física de rochas, notadamente as áreas de análise de sensibilidade aos fluidos e as relações quantitativas entre a velocidade compressional e cisalhante. Também foram apresentadas duas abordagens utilizadas para a discriminação litológica através de atributos elásticos: a primeira foi a abordagem frequentemente utilizada pelo intérprete que utiliza de maneira independente métodos estatísticos e de física de rochas, enquanto a segunda é o fluxo de trabalho estatístico e de física de rochas que propõe a utilização integrada dos métodos citados.

Alguns conceitos relacionados à programação genética também foram revistos, além de alguns trabalhos que abordaram a utilização da programação genética para o problema de classificação. Foi constatado que duas abordagens principais são empregadas para problemas de classificação de mais de duas classes: a decomposição binária e a seleção por faixas de valores. Além destas, trabalhos envolvendo a abordagem de programação genética com múltiplas saídas também foram apresentados. Foi possível constatar que na abordagem de decomposição binária o ponto central é a técnica de resolução de conflitos. Já na abordagem de seleção por faixas de valores a questão principal é como os valores limítrofes devem evoluir junto com os indivíduos.

Por fim, foi apresentado um estudo de caso onde três abordagens de classificação baseadas em PG foram utilizadas. Elas foram comparadas a uma rede neural além de duas abordagens estatísticas de classificação. Neste estudo de caso foram utilizados dados de atributos elásticos de três poços de uma mesma

bacia brasileira. Foram discriminadas três litologias nesta classificação: conglomerado, arenito e folhelho. A quantidade de padrões dessas três litologias encontrava-se altamente desbalanceadas.

Desses três poços, dois estavam majoritariamente saturados com gás enquanto o terceiro possuía uma zona saturada de gás e outra saturada de água. Foram realizados três experimentos a partir dos atributos elásticos desses poços. Primeiro, foi realizada a classificação dos atributos elásticos na condição *in situ*, ou seja, com os valores originais derivados dos perfis de poços. Em seguida, os dados originais foram estendidos através do procedimento de substituição de fluidos de Gassmann.

Como resultado, o estudo de caso confirmou que a utilização de métodologias híbridas obtiveram melhores resultados em todas as abordagens utilizadas. Ou seja, a utilização de dados teoricamente calculados através da física de rochas mostrou-se uma poderosa ferramenta para a obtenção de melhores resultados quando da classificação de atributos elásticos. Em relação às abordagens baseadas em PG, elas acabaram se mostrando uma escolha apropriada para este tipo de problema, pois permitem o uso de modelos que privilegiam a extração de conhecimento.

Além disso, as abordagens de classificação baseadas em PG conseguiram gerar classificadores que igualaram e até superaram o desempenho das outras abordagens, tanto com os dados originais como com as bases teoricamente estendidas. A abordagem de seleção por faixa (DGPG) foi a que atingiu os melhores resultados pelo fato do uso em conjunto de vários melhores indivíduos. Uma questão que afetou o desempenho das abordagens ECPG e PGMS foi a função de aptidão. Utilizar a taxa de acertos ou de erros da classificação, apesar de ser a forma mais direta de se relacionar a evolução de um sistema baseado em PG com o problema, mostrou-se não ser suficiente para levar a bons classificadores.

Outro resultado verificado é que as abordagens baseadas em PG conseguiram atingir resultados superiores com menor quantidade de dados. Isto é extremamente interessante do ponto de vista exploratório onde a quantidade de dados é diminuta. Este resultado confirma o poder das técnicas de inteligência computacional, no caso a programação genética, sobre as técnicas estatísticas quando uma pequena quantidade de dados está disponível.

Uma possível abordagem em trabalhos futuros é a utilização de funções de aptidão mais elaboradas, como atestou o trabalho de Teredesai & Govindaraju (2004). Elas são desejáveis no sentido de conduzir o processo evolucionário de forma mais eficaz no caso das abordagens ECPG e PGMS. Uma proposta a ser avaliada é a utilização da área sob a curva *Receiver Operating Characteristic* (ROC) como função de aptidão para as abordagens de decomposição binária.

Outra questão que deverá ser abordada em trabalhos futuros é como montar conjuntos de treinamento e validação em dados altamente desbalanceados. Esta situação é frequentemente encontrada na discriminação litológica de reservatórios de óleo e gás. Uma proposta envolvendo técnicas evolucionárias é apresentada no trabalho de GARCIA e*t al.* (2012).

Por fim, a utilização de operadores genéticos mais elaborados também será o escopo de trabalhos futuros. Dois exemplos desses operadores são os cruzamentos inteligentes, proposto no trabalho de ZHANG e*t al.* (2007), além dos operadores de cruzamento e mutação modificados, propostos no trabalho de MUNI *et al.* (2004).

REFERÊNCIAS BIBLIOGRÁFICAS

AVSETH, P., MUKERJI, T., MAVKO, G.; **Quantitative Seismic Interpretation: Applying Rock Physics Tools to Reduce Interpretation Risk**. Cambridge University Press. 2005.

BACHE, K., LICHMAN, M., UCI Machine Learning Repository, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013

BALDWIN, J.L., OTTE, D.N., WHEATLEY, C.L., Computer Emulation of Human Mental Processes: Application of Neural Network Simulators to Problems in Well Log Interpretation, 64th Annual Technical Conference and Exhibition of the Society of Petroleum Engineers, p. 481-493. 1989.

BIOT, M. A., Theory of propagation of elastic waves in a fluid saturated porous solid. I. Low frequency range and II. Higher-frequency range. J. Acoust. Soc. Am., 28, 168-191, 1956.

BRAMEIER, M., BANZHAF, W., A comparison of linear genetic programming and neural networks in medical data mining," IEEE Trans. Evol. Comput., vol. 5, no. 1, pp. 17–26, 2001.

BUSCH, J. M., FORTNEY, W. G., BERRY, L. N., Determination of lithology from well logs by statistical analysis: Society of Petroleum Engineers Formation Evaluation, v. 2, p. 412– 418, 1987.

CASTAGNA, J. P., BATZLE, M. L., EASTWOOD, R. O., Relationships between compressional-wave and shear-wave velocities in clastic rocks. **Geophysics**, 50, 571-581, 1985.

CASTAGNA, J. P., BATZLE, M. L., KAN, T. K., Rock physics: The link between rock properties and AVO response. **Investigations in Geophysics**, N. 8. Tulsa: Soc. Expl. Geophys., 135-171, 1993.

CHEN, Z., LU, S., A genetic programming approach for classification of textures based on wavelet analysis, in Proc. Int. Symp. Intell. Signal Process., Piscataway, NJ: IEEE, pp. 1–6, 2007.

COVER, T., THOMAS, J., **Elements of information Theory**. 2nd Ed. New York: Wiley. 2006.

DAVIS, R. A., CHARLTON, A. J., OEHLSCHLAGER, S., WILSON, J. C., Novel feature selection method for genetic programming using metabolomic 1H NMR data, Chemometrics Intell. Lab. Syst., vol. 81, no. 1, pp. 50–59, Mar. 2006.

DILLON, L. D., A contribuição da informação elástica no processo de caracterização de reservatórios, Tese de Doutorado, UFRJ, 2001.

DOVETON, J. H., **Geologic Log Analysis Using Computer Methods**, AAPG Computer Applications in Geology. 1994.

DOVETON, J. H., The Geological Application of Wireline Logs: A Keynote Perspective, AAPG Methods in Exploration No. 13, pp. 115–122, 2002.

ESPEJO, P. G., VENTURA, S., HERRERA, F., A Survey on the Application of Genetic Programming to Classification. IEEE Transactions on systems, man, and cybernetics—Part C: Applications and reviews, vol. 40, No. 2, March 2010.

FREITAS, A. A., Data Mining and Knowledge Discovery With Evolutionary Algorithms. Berlin, Germany: Springer-Verlag, 2002.

GARCIA, S., DERRAC, J., TRIGUERO, I., CARMONA, C. J., HERRERA, F., Evolutionary-based selection of generalized instances for imbalanced classification, Knowledge-Based Systems, Elsevier, Volume 25, Issue 1, Pages 3–12, 2012.

GASSMANN, F., Über die Elastizität poröser medien. Vier. Der Natur. Gesellschaft in Zürich, 96, 1-23. 1951.

GOLDBERG, D. E., Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, 1989.

GREENBERG, M. L., CASTAGNA, J. P., Shear-wave velocity estimation in porous rocks: Theoretical formulation, preliminary verification and applications. **Geophysics Prospecting**, 40, 195-209, 1992.

HAN, D., Effects of porosity and clay content on acoustic properties of sandstone and unconsolidated sediments. Ph.D. thesis, Stanford University, 1986.

HAN, J., KAMBER, M., PEI, J., **Data Mining - Concepts and Technique**. (The Morgan Kaufmann Series in Data Management Systems), 3rd ed. San Francisco, CA: Morgan Kaufmann, 2011.

HINCHLIFFE, M., HIDEN, H., MCKAY, B., WILLIS, M., THAM, M., BARTON, G., Modelling chemical process systems using a multi-gene genetic

programming algorithm. In: Koza, J. R., editor, Late Breaking Papers At The Genetic Programming Conference, p. 56-65, Stanford University, CA, USA, 1996.

International Commission on Stratigraphy (ICS) - Stratigraphic Guide. Disponível em: <http://www.stratigraphy.org/index.php/ics-stratigraphicguide>. Acessado em: 08 de Julho de 2014.

KISHORE, J. K., PATNAIK, L. M., MANI, V., AGRAWAL, V. K., Application of genetic programming for multicategory pattern classification, IEEE Trans. Evol. Comput., vol. 4, no. 3, pp. 242–258, 2000.

KOZA, J. R., **Genetic Programming: On the Programming of Computers by Means of Natural Selection**. MIT Press, Cambridge, MA, USA, ISBN 0-262-11170-5, 1992.

LEITE, V. R. C., Uma análise da classificacão de litologias utilizando SVM, MLP e métodos Ensemble, Dissertacão (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2012.

LIU, H., MOTODA, H., Feature Extraction, Construction and Selection: A Data Mining Perspectiv (The Springer International Series in Engineering and Computer Science Series, 453).. Berlin, Germany: Springer-Verlag, 1998.

LOVEARD, T., CIESIELSKI, V., Representing classification problems in genetic programming, in Proc. IEEE Congr. Evol. Comput.. vol.2, Seoul, South Korea: IEEE, pp. 1070–1077, 2001.

LUKE, S., PANAIT, L., Lexicographic Parsimony Pressure, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), 2002.

MATLAB, MATLAB R2008b, The Mathworks, Natick, MA, 2008.

MAVKO, G., MUKERJI, T., DVORKIN, J., **The Rock Physics Handbook: tools for seismic analysis in porous media**. Cambridge University Press. 1998.

MUNI, D. P., PAL, N. R., DAS, J., A novel approach to design classifiers using genetic programming, IEEE Trans. Evol. Comput., vol. 8, no. 2, pp. 183–196, 2004.

MURTHY, S. K., Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining Knowl. Discov., vol. 2, no. 4, pp. 345–389, 1998.

NIKRAVESH, M., AMINZADEH, F., ZADEH, L.A., **Soft Computing And Intelligent Data Analysis In Oil Exploration**. Elsevier Science B.V., Amsterdam, 1st Ed., p. 701, 2003.

OLTEAN, M., DIOSAN, L., An autonomous PG-based system for regression and classification problems, Appl. Soft Comput., vol. 9, no. 1, pp. 49–60, 2009.

PICKETT, G. R., Acoustic character logs and their applications in formation evaluation. J. Petrol. Technol., 15, 650-667, 1963.

POLI, R., LANGDON, W. B., MCPHEE, N. F., **A field guide to genetic programming**. Published via http://lulu.com and freely available at http://www.gpfield-guide.org.uk (With contributions by J. R. Koza), 2008.

ROGERS, S. J., FANG, J. H., KARR, C. L., STANLEY, D. A., Determination of lithology from well logs using a neural network: AAPG Bulletin, v. 76, no.5, p. 731-739, 1992.

SAGGAF, M. M., NEBRIJA, E. L., A fuzzy logic approach for the estimation of facies from wire-line logs. AAPG Bulletin, v. 87, no. 7 (July 2003), pp. 1223–1240, 2003.

SANTOS, R. O. V., VELLASCO, M. M. B. R., ARTOLA, F. A. V., DA FONTOURA, S. A. B., Neural Net Ensembles for Lithology Recognition, In: Windeatt, T., Roli, F., editors, Multiple Classifier Systems, volume 2709 Lecture Notes in Computer Science, p. 246–255. Springer Berlin / Heidelberg, 2003.

SCHLUMBERGER, Schlumberger Oilfield Glossary, Disponível em <http://www.glossary.oilfield.slb.com/Display.cfm?Term=lithology>. Acessado em 29 de julho de 2014.

SEARSON, D. P., LEAHY, D. E., WILLIS, M. J., GPTIPS: an open source genetic programming toolbox for multigene symbolic regression, Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS), Hong Kong, 17-19 March, 2010.

TAKAHASHI, I., MUKERJI, T. MAVKO, G., A strategy to select optimal seismic attributes for reservoir property estimation: Application of information theory. Soc. Expl. Geophys. 69th Ann. Mtg, Expanded Abstracts, 1584-1587. 1999.

TELFORD, W. M.; GELDART, L. P.; SHERIFF, R. E.; **Applied Geophysics**. Cambridge University Press. 2nd Ed. 1990.

TEREDESAI, A. M., GOVINDARAJU, V., Issues in evolving PG based classifiers for a pattern recognition task, in Proc. IEEE Congr. Evol. Comput., vol. 1, Portland, Oregon: IEEE, pp. 509–515, 2004.

U.S. Geological Survey (USGS), Earthquake Glossary, Disponível em <http://earthquake.usgs.gov/learn/glossary/?term=lithology>. Acessado em 29 de julho de 2014.

VILLAÇA, S. F., GARCIA, L. F. T., **Introdução à Teoria da Elasticidade**. Rio de Janeiro. COPPE/UFRJ. 4ª Edição. 2000.

ZHANG, L. NANDI, A. K., Fault classification using genetic programming, Mech. Syst. Signal Process., vol. 21, no. 3, pp. 1273–1284, 2007.

ZHANG, M., CIESIELSKI, V. B., ANDREAE, P., A Domain-Independent Window Approach to Multiclass Object Detection Using Genetic Programming, EURASIP Journal on Applied Signal Processing, 8, pp. 841–859, 2003.

ZHANG, M., GAO, X., LOU, W., A New Crossover Operator in Genetic Programming for Object Classification, IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, Vol. 37, No. 5, 2007.

ZHANG, M., SMART, W. D., Multiclass object classification using genetic programming, in Proc. Appl. Evol. Comput., (Lecture Notes in Computer Science Series, 3005). Coimbra, Portugal: Springer-Verlag, pp. 369–378. 2004.

ZHANG, M., SMART, W. D., Using Gaussian distribution to construct fitness functions in genetic programming for multiclass object classification, Pattern Recogn. Lett., vol. 27, no. 11, pp. 1266–1274, 2006.

ZHANG, M., WONG, P., Genetic programming for medical classification: a program simplification approach, Genet. Program. Evol. Mach., vol. 9, pp. 229–255, 2008.