UNIVERSIDADE FEDERAL FLUMINENSE
INSTITUTO DE GEOCIÊNCIAS
DEPARTAMENTO DE GEOLOGIA E GEOFÍSICA

FLAVIO COSTA DE MESQUITA

NOVO MÉTODO PARA IDENTIFICAÇÃO DE ESTRATIFICAÇÕES DE SAL UTILIZANDO *MACHINE LEARNING* SOBRE ATRIBUTOS SÍSMICOS

DISSERTAÇÃO DE MESTRADO

PROGRAMA DE PÓS-GRADUAÇÃO
DINÂMICA DOS OCEANOS E DA TERRA (DOT)

Niterói, RJ
2020

FLAVIO COSTA DE MESQUITA

NOVO MÉTODO PARA IDENTIFICAÇÃO DE ESTRATIFICAÇÕES DE SAL UTILIZANDO *MACHINE LEARNING* SOBRE ATRIBUTOS SÍSMICOS

Dissertação apresentada à Universidade Federal Fluminense como requisito parcial do Programa de Pós-Graduação em Dinâmica dos Oceanos e da Terra para a obtenção do título de Mestre em Ciências.

Área de concentração: Geologia e Geofísica.

Orientador
Prof. Marco Antonio Cetale Santos

Coorientador
Alexandre Rodrigo Maul

Niterói, RJ
2020

Ficha catalográfica automática - SDC/BIG
Gerada com informações fornecidas pelo autor

M578n Mesquita, Flavio Costa de
Novo método para identificação de estratificações de sal utilizando machine learning sobre atributos sísmicos / Flavio Costa de Mesquita ; Prof. Marco Antonio Cetale Santos, orientador ; Alexandre Rodrigo Maul, coorientador. Niterói, 2020.
100 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2020.

DOI: http://dx.doi.org/10.22409/PPGDOT .2020.m.07401179717

1. Machine Learning. 2. Estratificações de sal. 3. Geofísica marinha. 4. Pré sal. 5. Produção intelectual. I. Santos, Prof. Marco Antonio Cetale, orientador. II. Maul, Alexandre Rodrigo, coorientador. III. Universidade Federal Fluminense. Instituto de Geociências. IV. Título.

CDD -

Bibliotecário responsável: Sandra Lopes Coelho - CRB7/3389

FLAVIO COSTA DE MESQUITA

NOVO MÉTODO PARA IDENTIFICAÇÃO DE ESTRATIFICAÇÕES DE SAL UTILIZANDO *MACHINE LEARNING* SOBRE ATRIBUTOS SÍSMICOS

Dissertação apresentada à Universidade Federal Fluminense como requisito parcial do Programa de Pós-Graduação em Dinâmica dos Oceanos e da Terra para a obtenção do título de Mestre em Ciências.

Área de concentração: Geologia e Geofísica.

Aprovada em 03/11/2020 pela banca examinadora abaixo:

_____

Prof. Marco Antonio Cetale Santos, DSc. (Orientador)
UFF / DOT / GISIS

_____

Alexandre Rodrigo Maul, MSc., (Coorientador)
Petrobrás

_____

Prof. Alex Laier Bordignon, DSc.
UFF / IME

_____

Prof. Cleverson Guizan Silva, PhD.
UFF / DOT

_____

Marcos de Carvalho Machado, DSc.
Petrobrás

Niterói, RJ
2020

Agradecimentos

Agradeço aqueles que mesmo indiretamente me impulsionaram a buscar novas áreas de conhecimento, a me reinventar, me aperfeiçoar, a sair da minha zona de conforto para buscar novamente a sala de aula e com isso poder dividir meu conhecimento e aprender muito mais.

Aqueles que acreditaram em mim, que mesmo após tantos anos longe da academia consegui vencer este desafio, que é so o primeiro passo de uma longa carreira.

Ao meu orientador Marco Cetale pelos recursos fornecidos que me permitiram concluir este trabalho, pela oportunidade de participar do grupo de imageamento e inversão sísmica e pelas discussões e sugestões no decorrer do projeto .

Ao coorientador e amigo Alexandre Maul, pelas idéias e apoio sem os quais não teria conseguido seguir adiante.

Ao Prof Alex Bordignon pelas idéias iniciais que foram o norte deste projeto, bem como o apoio dado nas fases finais.

A Fundação Euclides da Cunha pela bolsa acadêmica, a ANP pela cessão dos dados para a realização do projeto, a Schlumberger pela licença acadêmica dos softwares de interpretação.

Aos amigos do GISIS que tanto me apoiaram e trocaram idéias e sugestões que foram muito importantes para meu desenvolvimento acadêmico.

E sobretudo a minha família por ter suportado minhas ausências para me dedicar a este mestrado.

"Não vos amoldeis às estruturas deste mundo,
mas transformai-vos pela renovação da mente,
a fim de distinguir qual é a vontade de Deus:
o que é bom, o que Lhe é agradável, o que é perfeito.
(Bíblia Sagrada, Romanos 12, 2)

Resumo

Para construir uma imagem sísmica, precisamos processar as informações das reflexões das interfaces das rochas. Essas reflexões ocorrem em função das diferenças das propriedades de impedância entre as rochas, que são calculadas como uma combinação de medidas de densidade e velocidade de compressão (inverso da vagarosidade). A halita, geralmente o mineral mais abundante na seção denominada sal, tem uma densidade média de cerca de 2,14 g/cm^3 e velocidade de compressão da ordem de 4.500 m/s. Em termos de estudos sísmicos, até o período recente, o modelo inicial para a seção de evaporitos era considerado aproximadamente constante refletindo as propriedades da halita, porém esta aproximação leva a erros no processo de migração em profundidade quando o sal se apresenta estratificado como é o caso da Bacia de Santos. Com a evolução dos algoritmos de migração sísmica e da capacidade computacional, percebeu-se a necessidade de modelar a seção de sal de forma mais complexa tornando-a menos homogênea, pois a formação de evaporitos (processo de evaporação) ocorre em estágios, de acordo com taxas de evaporação específicas, gerando as camadas observadas, também denominadas estratificações. Existem muitos tipos de minerais evaporitos na seção evaporítica das bacias de Santos e Campos, sendo os mais comuns halita, anidrita, gipsita, carnalita, taquidrita e silvita. Estudos realizados na análise de perfis mostram que nem todos esses tipos de minerais serão sísmicamente detectáveis por amplitude devido serem delgados e estarem abaixo da resolução sísmica. Assim, para facilitar a identificação dos estratos, os minerais evaporíticos na seção de sal vem sendo agrupados em três fácies principais: halita, sais de alta velocidade (anidrita e gipsita) e sais de baixa velocidade (carnalita, silvita e taquidrita), esses últimos são de especial interesse devido a sua alta solubilidade que pode vir a ocasionar problemas de circulação durante a perfuração podendo levar até mesmo ao abandono do poço. Devido a baixa quantidade de poços onde aparecem os chamados sais de baixa velocidade, além do próprio sinal foram também utilizados atributos sísmicos de forma que melhor representassem as camadas de sal esperadas de acordo com os perfis de poço. Muitos trabalhos foram feitos no sentido de inserir estas estratificações no modelo de velocidade, alguns trabalhos indicavam que mesmo a inserção de heterogeneidades randomicamente na camada de sal (o chamado “sal sujo”) já contribui na produção de melhores imagens no processo de migração. Neste trabalho apresentamos uma metodologia de construção de modelos de velocidades sísmicas onde é possível inserir essas estratificações baseado na identificação e separação das *wavelets* num modelo de *clusters* de uma sísmica migrada utilizando os poços perfurados na região como referencia e então utilizar esse modelo de *clusters* para comparar ponto a ponto o traço sísmico e gerar assim um novo modelo que considera as estratificações de sal e suas respectivas velocidades sísmicas. Para isso foram utilizadas técnicas de *machine learning* e os mais modernos algoritmos de clusterização, redução de dimensionalidade e atribuição.

Palavras-chave: sequência evaporítica, pré-sal, *machine learning*, modelagem.

Abstract

To build a seismic image, we need to process the reflection information from the rock interfaces. These reflections occur as a function of differences in impedance properties between rocks, which are calculated as a combination of density and compression velocity measurements (inverse of slowness). Halite, generally the most abundant mineral in the section called salt, has an average density of about 2.14 g/cm^3 and a compressional velocity of about 4,500 m/s. In terms of seismic studies, until the recent period, the initial model for the evaporite section was considered approximately constant reflecting the properties of halite, but this approximation leads to errors in the depth migration process when salt is stratified as is the case of the Santos Basin. With the evolution of seismic migration algorithms and computational capacity, it was realized the need to model the salt section in a more complex way making it less homogeneous, since the formation of evaporites (evaporation process) occurs in stages, according to specific evaporation rates, generating the observed layers, also called stratifications. There are many types of evaporite minerals in the evaporitic section in the Santos and Campos basins, the most common being halite, anhydrite, gypsum, carnallite, tachihydrite and silvite. Studies in well log analysis show that not all these minerals will be seismically detectable by amplitude, due to its thin layer it will be below the seismic resolution. Thus, to facilitate the identification of the strata, the evaporitic minerals in the salt section have been grouped into three main facies: halite, high-velocity salts (anhydrite and gypsum) and low-velocity salts (carnalite, sylvite and tachyhidrite). These last ones are of particular interest because of its high solubility which may cause circulation problems during drilling and may even lead to well abandonment. Due to the low number of wells where so-called low velocity salts appear, in addition to the signal itself, seismic attributes were also used to better represent the expected salt layers according to the well logs. Much work has been done to insert these stratifications into the velocity model, some studies have indicated that even randomly inserted heterogeneities into the salt layer (the so-called “dirty salt”) already helps to produce better images in the migration process. In this work we present a methodology for building seismic velocity models where it is possible to insert these stratifications based on the identification and separation of wavelets in a cluster model of a seismic already migrated using the wells drilled in the region as a reference and then use this cluster model to compare the seismic trace point by point and thus generate a new model that considers the salt stratifications and their respective seismic velocities. In order to achieve it we used machine learning techniques and the most modern clustering, dimensionality reduction and assignment algorithms.

Keywords: evaporitic sequence, pre-salt, machine learning, modelling.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DBSCAN	Density Based Spatial Clustering Application with Noise
HDBSCAN	Hierarchic Density Based Spatial Clustering Application with Noise
KNN	K Nearest Neighbors
g/cm3	gramas por centimetro cubico
m/s	metros por segundo
FWI	Full Waveform Inversion
HVS	High Velocity Salt
LVS	Low Velocity Salt
UMAP	Uniform *manifold* Aproximation and Projection
ANP	Agência Nacional do Petróleo
mm	milímetro
Ma	milhão de anos
km	quilometro
SOM	Self Organizing Map

Lista de símbolos

Γ	Letra grega Gama
Λ	Letra grega Lambda
ζ	Letra grega minúscula zeta
ϵ	Letra grega minúscula epsilon
$\in$	Pertence

Sumário

1 Introdução

As imagens sísmicas são obtidas pela propagação de ondas elásticas e apresentam relações com informações geológicas e estratigráficas. Os produtos resultantes dos levantamentos sísmicos e suas respectivas propriedades possibilitam informações não apenas sobre a disposição geométrica das camadas geológicas, mas também possibilitam estimativas de propriedades, tais como porosidade, velocidade sísmica e possivelmente inferir o tipo de fluído contido em seus poros. A interpretação destas propriedades é comumente feita utilizando atributos sísmicos aliados à correlação com as informações obtidas a partir de poços Chopra e Marfurt (2005).

A modelagem sísmica é essencialmente uma simulação de um campo de onda sísmico registrado, amplitudes sísmicas ou tempos de trânsito sísmicos. A entrada para a modelagem sísmica é uma representação da refletividade da Terra e um modelo de velocidade-profundidade. A migração sísmica é um processo de estimativa da refletividade da Terra a partir de um campo de ondas sísmicas registrado, usando um modelo de velocidade-profundidade. Portanto, a modelagem de campos de ondas sísmicas pode ser vista como o processo reverso da migração sísmica. Como tal, os algoritmos de migração sísmica e de modelagem de campos de ondas sísmicos são baseados na equação de onda, Yilmaz (2001).

Atualmente, existem muitas discussões a respeito da construção de modelos de velocidades mais acurados para propósitos de migração sísmica. A incorporação de feições geológicas complexas nem sempre é uma tarefa simples de se realizar. As imagens de subsal podem ser muito sensíveis à precisão do modelo de sal porque a má interpretação do sal pode facilmente levar a grandes erros de tempo e desvio significativo dos caminhos das ondas. Portanto, entre outros fatores, uma chave para o sucesso da geração de imagens de sub-sal é construir modelos geologicamente mais apurados de velocidade do sal.

Por décadas, a prática padrão para a construção de modelos de sal tem sido o uso de tomografia, às vezes combinada com inversão de onda completa (FWI) da onda de mergulho rasa, Yilmaz (2001), para construir primeiro o melhor modelo de velocidade de sedimento possível. Geralmente, são necessários testes de cenário para resolver a geometria do sal, especialmente para áreas complexas. Este procedimento não é apenas trabalhoso e demorado, mas também propenso a mais erros na interpretação. Por isso, no caso da sequência evaporítica da Bacia de Santos, embora os dados sísmicos e de poços revelem que a mesma encontra-se estratificada, tradicionalmente, os modelos de velocidades consideram-na como uma camada homogênea e isotrópica para fins de imageamento em profundidade.

1.1 Contribuições

Este trabalho apresenta uma nova metodologia interativa para identificação das sismofácies dentro da camada de sal, e uma posterior inserção dessas estratificações de sal com suas respectivas velocidades no modelo final de velocidades sísmicas do Campo de Búzios na Bacia de Santos, considerando a alternância dos diferentes tipos de rochas presentes na seção evaporítica, que influenciam, diretamente, na definição e caracterização sísmica dos reservatórios da seção Pré-Sal. Para isso foi usado um fluxo de trabalho de *machine learning*, onde os dados das *wavelets* em cada tipo de sal foram agrupadas em *clusters* também conhecidos como sismofácies e esse "modelo de *clusters*" foi usado como padrão para se comparar os dados (*wavelets*) da camada de sal de uma determinada área e com isso classificar cada sismofácie em um dos três grupos mencionados anteriormente, Halita, Sais de Alta Velocidade - HVS ou Sais de Baixa Velocidade - LVS, como agrupado por Maul, Santos e Silva (2018).

Novos algoritmos foram utilizados, que substituem de forma melhor e mais eficaz os usados atualmente na identificação de sismofácies que pode ser usado em qualquer ambiente deposicional somente alterando os dados de entrada e modelo utilizado.

1.2 Objetivos

O principal objetivo deste trabalho é identificar as sismofácies das estratificações de sal que compõem a chamada formação Ariri no Campo de Búzios, Bacia de Santos. Essas estratificações surgiram devido ao processo de formação dos evaporitos acontecer em estágios, de acordo com taxas de evaporação específicas, formando a Anidrita, a Halita, a Carnalita e a Taquidrita nesta sequência respectivamente, já a taxa de solubilidade se dá justamente ao contrario desta sequência.

Devido a esta alta solubilidade os sais de baixa velocidade (Carnalita e Taquidrita) se tornam um desafio para a indústria durante a fase de perfuração do sal, levando a perda de circulação e até mesmo uma possível perda do poço, por isso um objetivo secundário deste trabalho é a identificação destes sais na imagem migrada e com isso ser usado como uma ferramenta a mais para os engenheiros de perfuração poderem alterar os parâmetros antes de entrar na camada de sal.

Para chegarmos a estes objetivos um fluxo de trabalho de *machine learning* composto por várias etapas foi desenvolvido e será discutido nos próximos capítulos.

1.2.1 Organização do texto

Para uma melhor compreensão do método o texto foi organizado da seguinte forma:

a) no capítulo 2 dissertamos sobre a Bacia de Santos e a formação dos evaporitos;

b) no capítulo 3 falamos sobre os atributos sísmicos e a escolha dos mesmos;

c) no capítulo 4 depois falamos sobre o uso de *machine learning* na identificação de litofácies e o conjunto de dados utilizado;

d) no capítulo 4 seção 4.4 e seção 4.5 é apresentada uma fundamentação teórica sobre cada algoritmo utilizado, como funciona, parâmetros;

e) no capítulo 5 são apresentados os resultados das comparações entre métodos e os resultados obtidos com nossa metodologia;

f) por último no capítulo 6 apresentamos nossas conclusões e artigos apresentados e submetidos durante o desenvolvimento do projeto.

2 A Bacia de Santos

Em Outubro de 2018 a Bacia de Santos era a maior produtora brasileira, tanto de petróleo quanto de gás natural, com uma produção diária de, respectivamente, 1,303 milhões de barris de óleo e 59,24 milhões de metros cúbicos de gás, totalizando 1,675 milhões de barris de óleo equivalente, através de 10 campos petrolíferos, descobertos pela Petrobras, que é também a principal companhia operadora. Estes campos estão localizados desde águas rasas até águas ultra-profundas, desde 2007 a Petrobras descobriu importantes acumulações de petróleo e gás natural, em águas ultra-profundas e abaixo de uma espessa camada de sal. Em outubro de 2018, cinco campos de petróleo já produziam do pré-sal, Lula, Sapinhoá, Búzios, Lapa e Mero, com um produção diária total de 1,254 milhões de barris de óleo e 51,79 milhões de metros cúbicos de gás natural, ANP (2018).

2.1 Localização

A bacia de Santos abrange uma área de 350.000 km^2 que cobre os estados do Rio de Janeiro, São Paulo, Paraná e Santa Catarina, e é limitada a norte pelo Alto de Cabo Frio, ao sul pela plataforma de Pelotas e, na direção E-W, se estende desde a linha da costa continental a oeste até o platô de São Paulo a leste, Moreira et al. (2007). Em setembro de 2017 haviam mais de 600 poços já perfurados na bacia de Santos e aproximadamente 40 campanhas de aquisição sísmica perfazendo uma área de aproximadamente 127.000 km^2 , ANP (2018).

Do ponto de vista do potencial petrolífero, a Bacia de Santos constitui uma importante fronteira exploratória devido às descobertas relacionadas a grandes campos de gás nas décadas de 80 e 90. No entanto, a sua relevância mudou após a descoberta de grande quantidade de petróleo nas camadas pré-sal em 2007, após a descoberta do campo Tupi. Muitos esforços têm sido feitos para a identificação e caracterização dos campos *offshore* das camadas do pré-sal. A área de estudo se situa no polo Pré-Sal da Bacia de Santos, mais especificamente no Campo de Búzios. Esta área é especialmente caracterizada pela presença de diápiros salinos, que ora se apresentam homogêneos, ora estratificados, e estão situados acima dos reservatórios carbonáticos desta seção Pré-Sal. A seção evaporítica na Bacia de Santos, objeto deste estudo, possui significativas espessuras podendo chegar a mais de 2000 m, a extensão dos campos do Pré-Sal pode ser vista na Figura 1.

Figura 1 – Distribuição das rochas reservatórios do Pré-sal (em azul escuro) em relação às bacias sedimentares da margem continental sudeste brasileira. Tomado de Riccomini, Sant'Anna e Tassinari (2012).

2.2 Evolução tectono-sedimentar

De acordo com Moreira *et al.*, 2007, a evolução tectonoestratigráfica da Bacia de Santos é dividida em três supersequências: Rifte, Pós-Rifte e Drifte, ver Figura 2. Sendo o embasamento da bacia representado por rochas pré-cambrianas da Faixa Ribeira. O espaço de acomodação para a sedimentação foi gerado a partir da subsidência relacionada aos esforços distensivos que resultaram no rifteamento do Gondwana. Inicialmente, a deposição de sedimentos ocorreu em ambiente flúvio-lacustre, passando por estágio de bacia evaporítica, evoluindo para bacia de margem passiva, Chang et al. (2008).

A seguir vamos explicar cada uma das sequências mais importantes:

a) Fase Rifte: esta fase se estende do Hauteriviano ao Aptiano e compreende os sedimentos depositados durante o processo de ruptura do Gondwana. Essa supersequência está dividida em três sequências deposicionais, representadas pelas formações Camboriú, Piçarras e Itapema.

b) Fase Pós-Rifte: a supersequência pós-rifte foi depositada entre o Aptiano e início do Albiano, idades correspondentes ao andar Alagoas. As sequências descritas por Moreira et al. (2007), englobam as Formações Barra Velha e Ariri, depositadas em ambiente transicional entre continental e marinho raso bastante estressante. O topo da supersequência corresponde aos evaporitos da Formação Ariri, composto principalmente por halita e anidrita, ainda com presença de sais solúveis, tais como, taquidrita, carnalita e silvinita. Os evaporitos ocorreram no Neoaptiano, atingindo cerca de 2.000 metros de espessura, e tem como limite superior os sedimentos

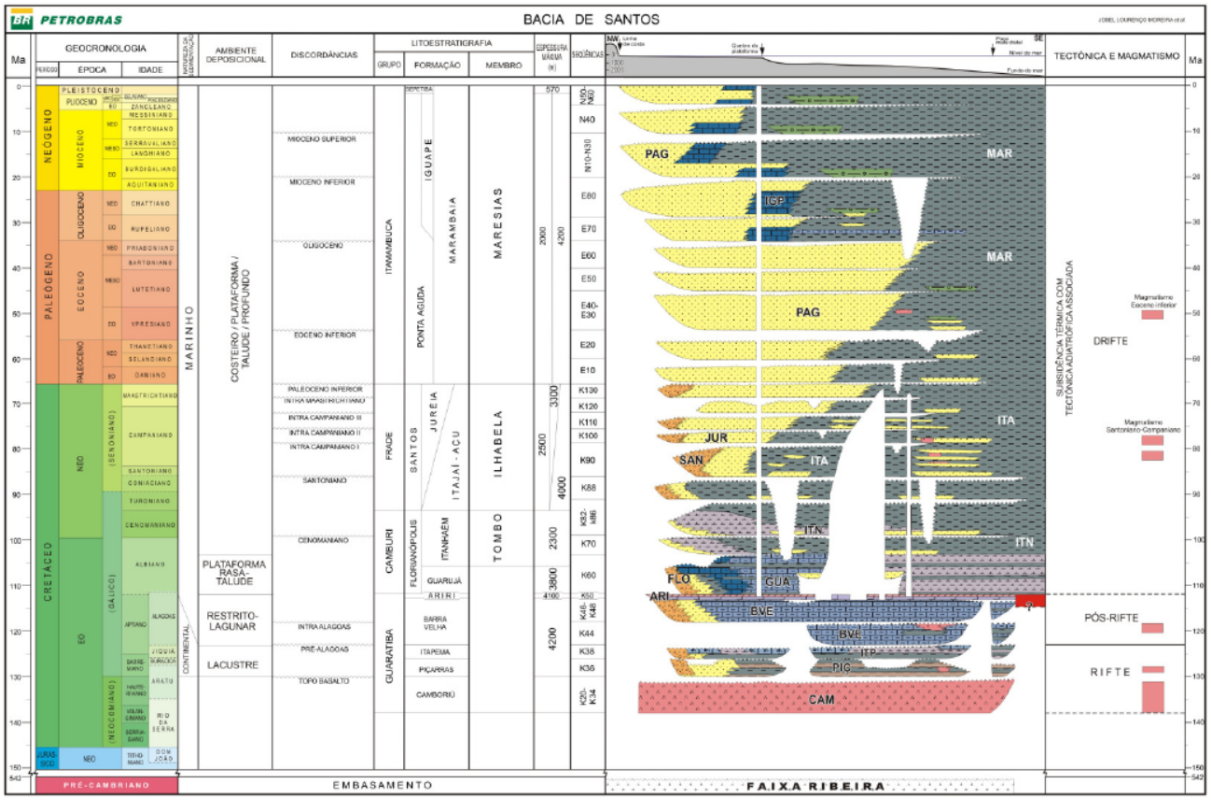

Figura 2 – Carta estratigráfica da Bacia de Santos, Moreira et al. (2007).

siliciclásticos/carbonáticos das Formações Florianópolis e Guarujá. É na Formação Ariri que se dá o foco deste trabalho, o qual dissertaremos mais adiante.

c) Fase Drifte: a última supersequência descrita por Moreira et al. (2007), foi denominada de fase drifte, depositada a partir do Albiano até o recente. Esta sedimentação de origem marinha está relacionada à subsidência termal e é representada pelos grupos Camburi, Frade e Itamambuca.

2.3 Os evaporitos da Formação Ariri

A fase rifte foi seguida pela fase transicional do Aptiano, representada pelos evaporitos, objeto deste estudo, definida como Formação Ariri. Acredita-se em uma aceleração acentuada da subsidência da bacia nesta área, concomitantemente à deposição dos evaporitos. Estes foram depositados sobre uma proeminente discordância causada por um soerguimento regional após a fase rifte, cuja continuidade pode ser seguida ao longo de toda margem oriental brasileira.

Do ponto de vista da evolução tectono-sedimentar da margem passiva do Atlântico Sul, os evaporitos estão localizados na zona de transição entre as fases sag e marinha, iniciada quando a subsidência já é governada pelo regime flexural termal, Gamboa

et al. (2008). Este extenso depósito evaporítico desenvolve-se desde a Bacia de Santos até a Bacia de Sergipe-Alagoas, no extremo norte. O comprimento total deste cinturão é da ordem de 1800 km e a sua porção sul, que possui mais de 700 km de largura, está intimamente associada ao arcabouço estrutural do embasamento representado pelo Platô de São Paulo, veja Figura 3, Gamboa e Rabinowitz (1984).

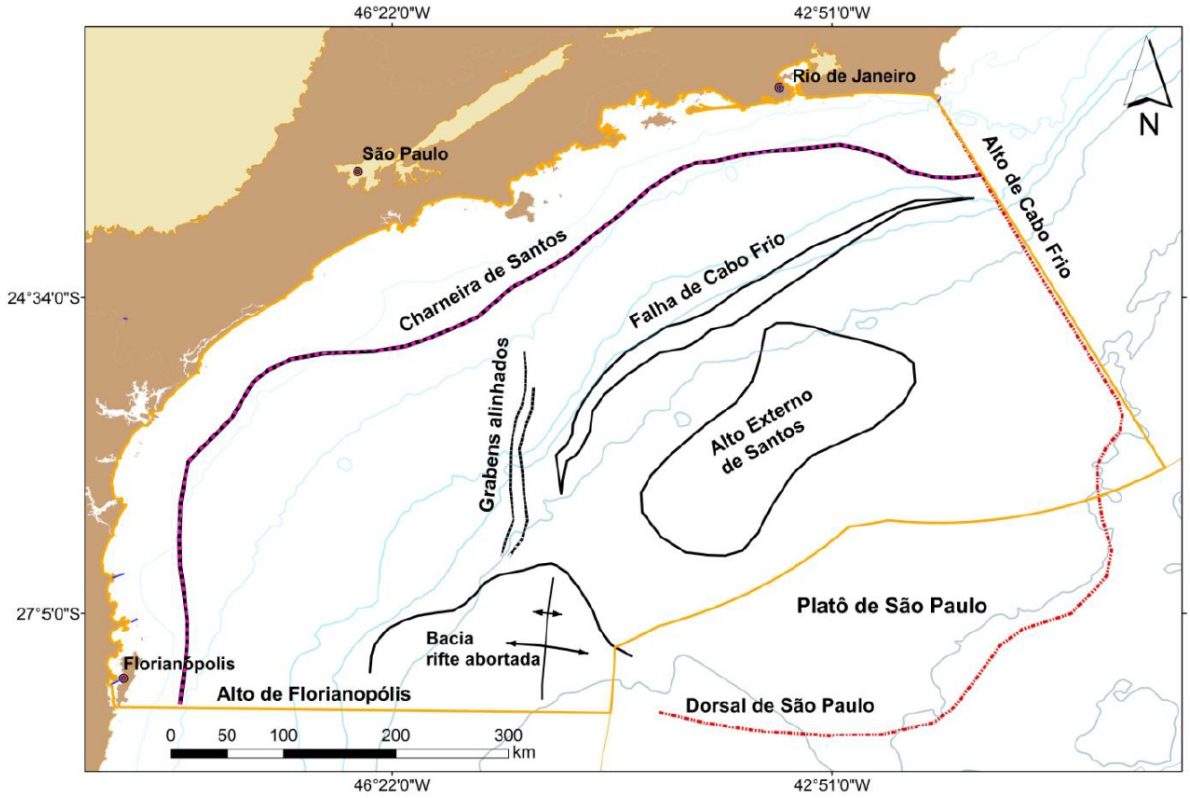

Figura 3 – Mapa de localização da Bacia de Santos com os principais elementos do seu arcabouço regional, editado de Garcia (2012).

2.3.1 Deposição dos evaporitos

Durante o Aptiano, a parte norte da bacia do Atlântico Sul estava localizada ao norte do trópico de Capricórnio, no meio do cinturão árido que contém a maioria dos desertos modernos do hemisfério sul: desertos de Kalahari e Namíbia na África, deserto de Atacama no Chile e o Deserto Australiano. As bacias de sal que faceiam uma a outra, entre o Alto do Rio Grande e o Golfo de Benin estão entre as maiores ao longo das margens oceânicas passivas do Fanerozoico. Elas se formaram durante o Aptiano (125-110 Ma) durante os estágios iniciais de abertura do norte do Atlântico Sul. A geometria, cinemática e ambiente temporal deste episodio de deposição de sal do inicio do Cretáceo é muito similar a do Mar Vermelho do meio pro final do Mioceno, veja Figura 4.

O período temporal e o cenário tectônico sumarizado por Karner e Gamboa (2007), é que após o impacto do *hot spot* de Tristão da Cunha na litosfera Africana-Sul

Americana, em torno de 143 Ma atrás, a Placa Sul-Americana começou a se separar da Africana a uma taxa de muitos mm/ano. Riftes estreitos de 50-80 km de largura, com segmentos se sobrepondo largamente, se formaram ao longo da placa recém formada, e foram preenchidos com sedimentos trazidos por rios fluindo de ambos os continentes. Lagos com centenas de metros de profundidade, alguns anóxicos, e vulcanismo basáltico, pontuaram a geologia de tais riftes no período Neocomiano – inicio do Barremiano (133-128 Ma).

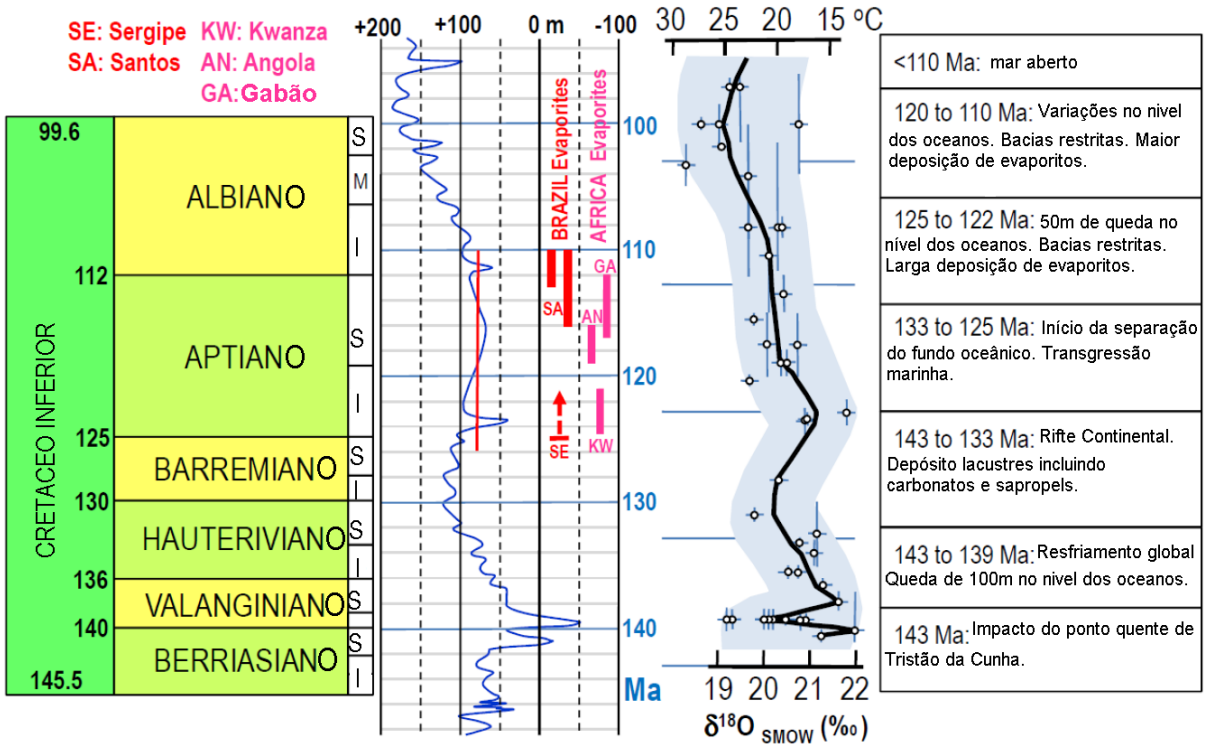

Figura 4 – Estratigrafia, eustasia, eras, temperatura e principais eventos do início do Cretáceo, Brasil e África Ocidental. Montaron e Tapponnier (2009).

A ruptura ocorreu entre o fim do Barremiano e o inicio do Aptiano (128-125 Ma), como o fundo oceânico começou a se abrir, a taxa de separação das placas aumentou 4 ou 5 vezes, para alguns cm/ano, e o espaço de acomodação entre a Africa e o Brasil rapidamente se alargou e aprofundou. Contudo, a bacia marinha complexa de 1700 km de comprimento, permaneceu isolada entre duas grandes "represas": a ondulação topográfica feita por vulcanismo no Sul (Alto de Walvis – Elevação do Rio Grande), e as zonas de fratura/transformação de Vema-Saint Paul na nascente do Golfo de Benin ao Norte.

Essas duas comportas vulcânico/tectônicas somente permitiam quantidades limitadas de água do mar fluírem para a bacia, na maioria através de fissuras tectônicas através de basaltos ao sul. Um ambiente deposicional tão restrito persistiu por 9 Ma, durante este período, a rápida evaporação da água do mar criou a maioria das espessas camadas de depósitos de evaporitos, cobrindo a crosta continental preenchida com sedimentos da fase rifte e pré – rifte, numa larga (300 – 500 km) e profunda (1 – 2 km) bacia, ainda

complexamente compartimentalizada por degraus e sobreposições. Condições perenes de mar-aberto foram restabelecidas no início do Albiano (112 – 110 Ma), devido a maior reorganização (simplificação) do alto do Meio-Atlântico ao norte do *hot spot*.

Na Figura 5 podemos ver as varias etapas de deposição:

Etapa 1: inicio do rifte, água dos rios invade a depressão formada ao redor de 134-130 Ma;

Etapa 2: em 123 Ma no início do Aptiano, o nível do oceano cai em 50 m. As bacias são isoladas das águas abertas do oceano. Formam-se Carbonatos;

Etapa 3: subida do nível do mar, alagamento;

Etapa 4: queda do nível do mar, água continua entrando por fissuras nos diques de basalto;

Etapa 5: a taxa de evaporação é maior que a ingestão de água de rios, chuvas e nascentes de água do mar, salinidade da água aumenta gradualmente e o nível da água da bacia cai até que a área superficial da água se estabilize em um valor crítico Ac (Área Crítica);

Etapa 6: a área da superfície da água não muda mais significativamente. A água continua evaporando com uma ingestão média de água do mar relativamente estável até que o sal comece a precipitar. A salinidade é mantida na concentração de saturação de sal;

Etapa 7: a espessura da camada de sal aumenta gradualmente até atingir o nível crítico da superfície de água, Ac;

Etapa 8: retorno às condições marinhas abertas. Os carbonatos de cálcio e magnésio precipitam primeiro, seguidos pela gipsita, que se transforma em anidrita sob condições particulares de temperatura e pressão, seguidos pela halita quando a densidade da água atinge $\rho_{sat} = 1214\ kg/m^3$ a 25 °C e, finalmente, sais complexos, incluindo carnalita, bischofita e taquidrita, nesta ordem.

De acordo com Montaron e Tapponnier (2009), a taxa média de deposição de halita foi de 2-3 cm/ano, assumindo inclinações na margem do rifte em torno de 4%. Essa taxa pode se aplicar aos 600 m inferiores da sequência de evaporitos da bacia de Santos que contém apenas anidrita e halita. Pode levar apenas 24000 anos para depositar esses 600 m mais baixos. No entanto, acima desse nível, há pelo menos nove sequências contendo sais complexos e essas podem levar muito mais tempo para serem depositadas. Além disso, a maioria dessas complexas camadas de sal incluem taquidrita, pois essa precipita no estágio terminal da bacia de sal quando a salmoura tem uma densidade acima de 1300 kg/m3 e a taxa de evaporação é quase zero. Em geral, a taquidrita não pode precipitar porque a água perdida durante o dia é reabsorvida da umidade atmosférica durante a noite.

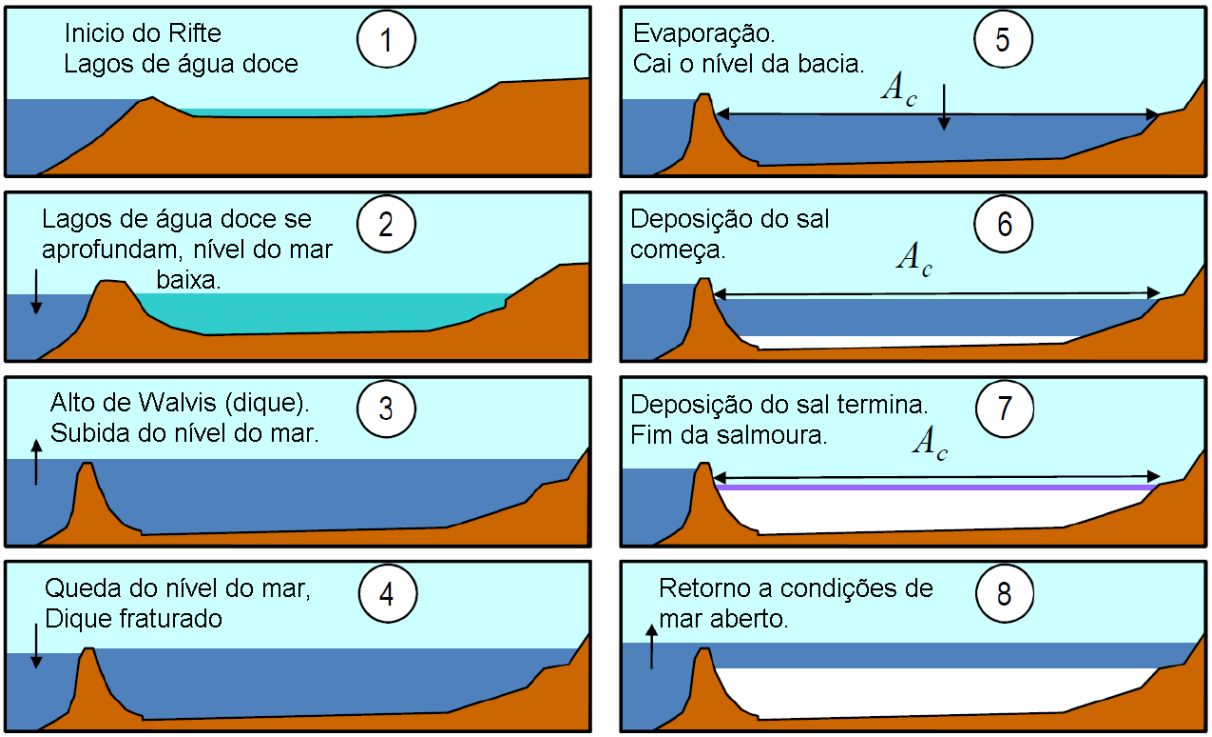

Figura 5 – Nesse cenário considera-se a deposição do sal homogênea, editado de Montaron e Tapponnier (2009).

Camadas de taquidrita são marcadores claros de episódios climáticos extremamente áridos que podem ser correlacionados com ciclos de força orbital de Milankovitch (por exemplo, 22000 anos). A observação detalhada dos núcleos de evaporito pode ajudar a determinar o tempo total de deposição (por exemplo, marcadores de depósito eólico).

De acordo com Guerra e Underhill (2012), a presença do sal influencia fortemente a prospecção de hidrocarbonetos em uma bacia sedimentar, devido sua baixa permeabilidade, baixa resistência e alta condutividade termal. Além de ser meramente uma rocha selante, sua mobilidade termina criando estruturas de trapeamento no processo de migração do petróleo. Mesmo pequenas camadas de sal impactam todo o funcionamento do sistema petrolífero de uma bacia.

Na Bacia de Santos a seção evaporítica não é homogênea, pelo contrário, existem diferentes padrões na camada de sal, o domínio estratificado consiste provavelmente de evaporitos de composições diferentes intercalados, incluindo anidrita, taquidrita, carnalita, silvita e halita, já no domínio basal existe a presença de uma rocha mais homogênea, composta provavelmente por halita pura, Freitas (2006).

Gamboa et al. (2008), definiram quatro sequências evaporíticas principais, da base para o topo: (1) um depósito composto predominantemente por halita, denominado "halita inferior"; (2) uma sucessão estratificada de anidrita, halita e sais complexos; (3) um outro pacote com predomínio de halita ("halita superior") e; (4) um pacote superior

também composto por sais estratificados, porém mais delgado, veja Figura 6. Estas sequências foram mapeadas por meio da sísmica ao longo do depocentro da bacia, e a sua composição foi reconhecida após a perfuração de poços pioneiros. A evolução dos ciclos desde a base até o topo denota sucessivas variações do nível de água dentro de um enorme mar alongado, até a progressiva depleção das salmouras, antecedendo à inundação definitiva da bacia, e o estabelecimento de condições de mar aberto no Albiano.

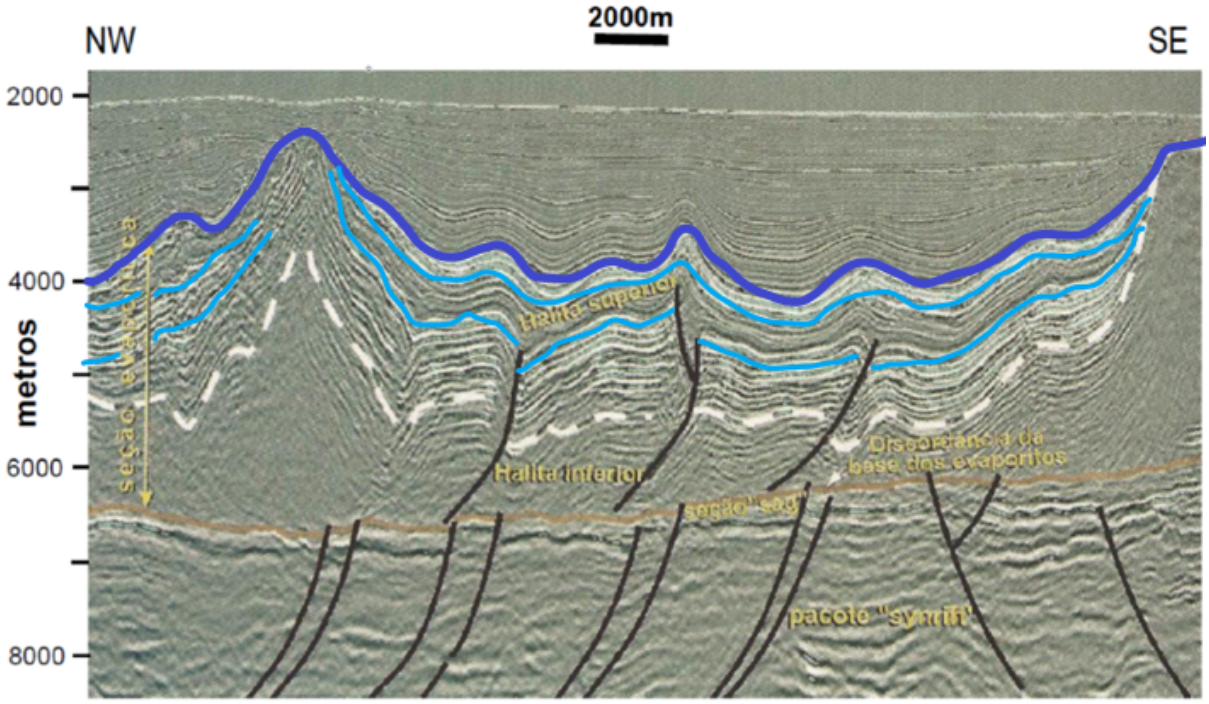

Figura 6 – Seção sísmica em profundidade mostrando os principais pacotes evaporíticos. As intercalações azuis e brancas mostram os altos contrastes de impedância existentes nos pacotes de evaporitos estratificados. Fonte: Gamboa et al. (2008).

De acordo com Fiduk e Rowan (2012), e Jackson et al. (2014), embora a Bacia de Santos seja composta predominantemente por halita (em geral, acima de 80%), a fim de definir a base percentual dos demais evaporitos (anidrita, carnalita, taquidrita), Jackson et al. (2014) dividiram a seção em quatro unidades (A1, A2, A3 e A4) por suas variadas densidades e padrões de refletividade, veja Figura 7, e observaram que unidades mais refletivas e estratificadas indicam maior ocorrência de sais de alta velocidade - HVS (anidrita, por exemplo) ou baixa velocidade - LVS (carnalita, por exemplo), como agrupado por Maul, Santos e Silva (2018).

A1 é considerada como sendo rica em halita (77-98%), pobremente refletiva, com baixa densidade e padrão sísmico caótico-fraco; já A2 possui alta refletividade, com fortes amplitudes sísmicas e alta densidade, devido à presença de anidrita intercalada a uma menor ocorrência de halita (67-86%); A3 é caracterizada por sua baixa refletividade, predominância de halita (69-94%) e densidade similar à de A1; e A4 é uma unidade

fortemente refletiva, com menor proporção de halita (31-94%), indicando a ocorrência de evaporitos de baixa densidade (carnalita, por exemplo).

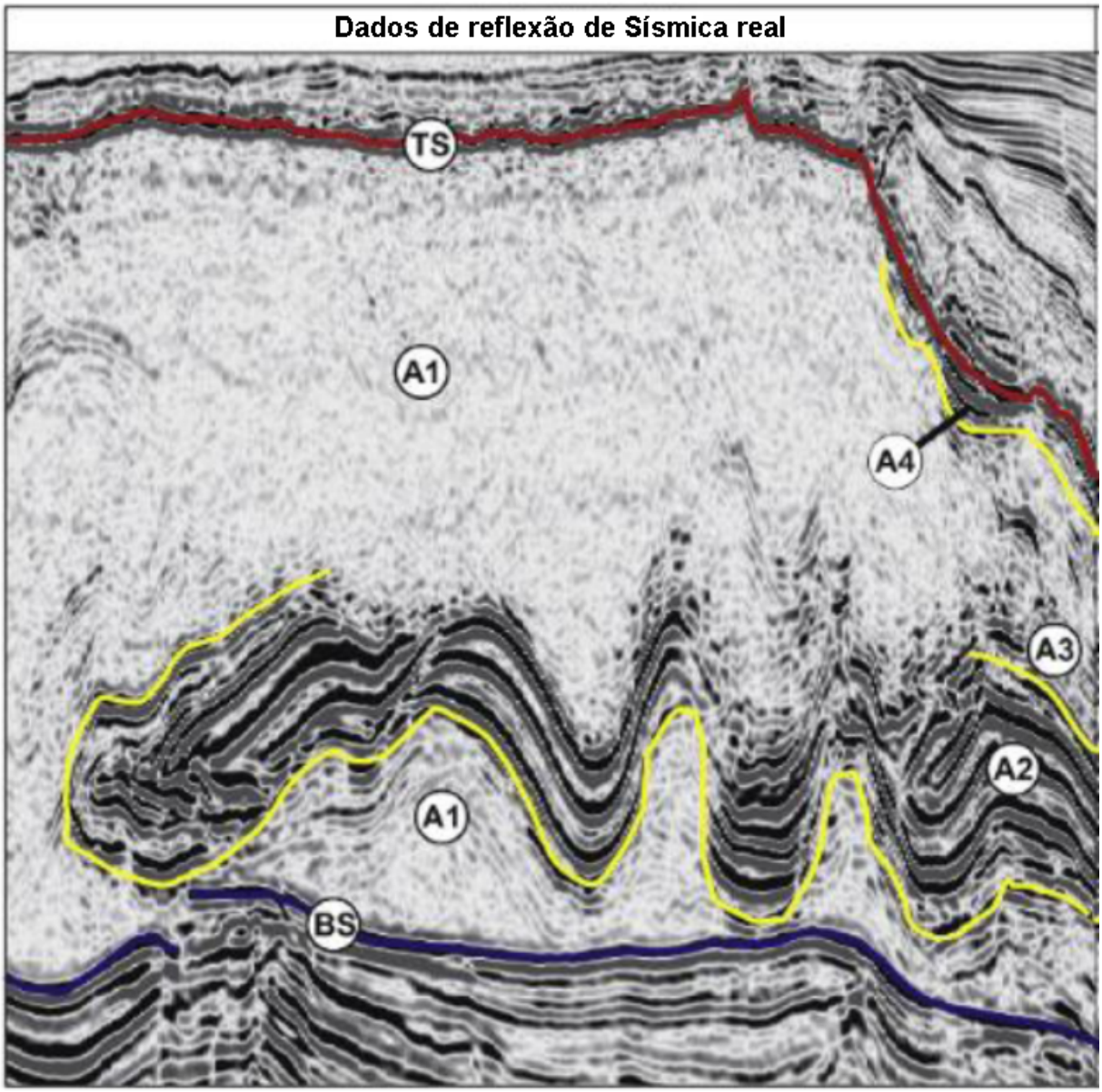

Figura 7 – Quatro unidades intra-sal definidas a partir de dados de poços (A1, A2, A3, A4), na Fm. Ariri, Bacia de Santos, por Jackson et al. (2014). O topo e a base do sal são definidos como refletores fortemente positivo e fortemente negativo, respectivamente (TS e BS).

As unidades intra-sal diferem entre si em composição e propriedades acústicas, e conforme observado, a sucessão estratigráfica que define as heterogeneidades da sequência evaporítica caracteriza reflexões sísmicas com alto contraste de impedância e comportamentos sísmicos bastante peculiares, o que as torna um potencial alvo de estudos de modelagem sísmica.

3 Atributos Sísmicos

A exploração, o desenvolvimento e a produção de campos de hidrocarbonetos depende de informações de subsuperfícies que podem ser acessados de formas diretas e indiretas. A principal forma direta conhecida é a perfuração de poços, que tem um custo muito elevado. Esta, por sua vez, depende muito de informações indiretas (métodos sísmicos) que, apesar de muito menos precisos, diminuem, e muito, o risco na perfuração de poços.

No campo dos métodos indiretos, ou métodos geofísicos, o método sísmico é um dos principais utilizados nestas etapas de exploração, de desenvolvimento e de produção de campos de hidrocarbonetos, especialmente em projetos *offshore*, em especial os projetos do polo Pré-Sal da Bacia de Santos, que é considerado como de águas ultra profundas, e com reservatórios profundidos (lâminas d'água acima de 1500 m, e reservatórios situados abaixo de 5000 m de profundidade). Assim, tecnologias têm sido desenvolvidas para melhorar a acurácia das informações contidas nestes dados, dentre as quais está inserido o tema desta dissertação.

Os métodos sísmicos são baseados em medições do intervalo de tempo entre o início de uma onda sísmica (elástica) e sua chegada aos detectores. A onda sísmica pode ser gerada por uma explosão, um peso derrubado, um vibrador mecânico, uma bolha de ar de alta pressão injetada na água ou outras fontes e pode ser detectada por um geofone em terra ou por um hidrofone na água.

Essa onda que também é chamada de *wavelet*, propagasse através das camadas litológicas e ao convolucionar com a refletividade característica em cada interface devido as diferenças de impedância acústica existentes em cada camada, juntamente com o ruído, forma o traço sísmico, veja Figura 8, e o conjunto de traços sísmicos variando no espaço e no tempo, após processamento, formam uma imagem sísmica. Essa *wavelet* ao viajar pelas camadas litológicas, sofre uma atenuação de amplitude, variação de fase, e filtragem das frequências mais altas, com isso a resolução em grandes profundidades perde um pouco de acurácia.

Sendo o método sísmico um dos principais métodos utilizados na exploração e produção de hidrocarbonetos, tecnologias têm sido desenvolvidas para melhorar a acurácia das informações contidas nestes dados. No âmbito da interpretação sísmica, uma das ferramentas mais utilizadas para auxiliar na visualização de aspectos qualitativos são os atributos sísmicos. Nascidos nos anos 50, os atributos ganharam real força no mercado por volta dos anos 70, devido ao avanço computacional. De acordo com diferentes autores o atributo sísmico pode ser descrito de diferentes maneiras:

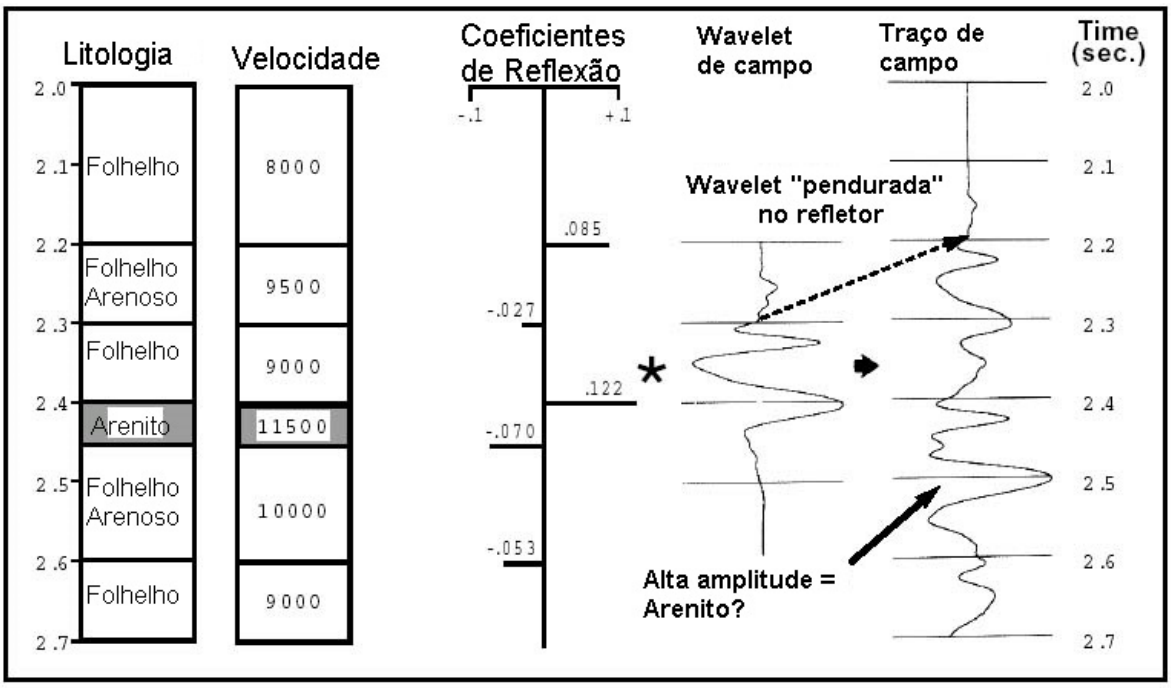

Figura 8 – Os limites litológicos definem uma série de coeficientes de reflexão, que quando convoluídos (*) com a *wavelet* de campo resultam no traço bruto de campo. Adaptado de Henry, S.(1997).

Taner, Koehler e Sheriff (1979): O traço sísmico convencional é tido como um componente real (f(t)) de um traço sísmico complexo (F(t)), que por sua vez pode ser considerado um vetor. A parte imaginária do traço complexo (g(t)) é obtida matematicamente através da Transformada de Hilbert, a partir do dado real. Sendo assim, separar o traço em parte imaginária e real favorece a obtenção de informações envolvendo amplitude e fase, chamadas atributos.

Chopra e Marfurt (2007): Atributos sísmicos são ferramentas utilizadas para inferir a geologia em dados sísmicos de reflexão. Dessa forma, os melhores atributos são extremamente sensíveis às características da geologia ou dos reservatórios aos quais estão sendo associados de maneira que seus objetivos são quantificar a amplitude e as características morfológicas capturadas pelo dado sísmico através de cálculos determinísticos geralmente efetuados por um software.

Barnes (2016): Atributos sísmicos atuam como filtros que removem uma característica do dado sísmico em prol de revelar outra. Atributos quantificam e descrevem dados sísmicos e podem ser divididos em categorias de acordo com seu significado, ver Figura 9.

Geralmente as áreas de exploração e produção não possuem cobertura regular de poços, ou mesmo os dados de poços não amostram todas as fácies existentes. Além disso, as campanhas de perfuração são planejadas para identificar a existência do reservatório

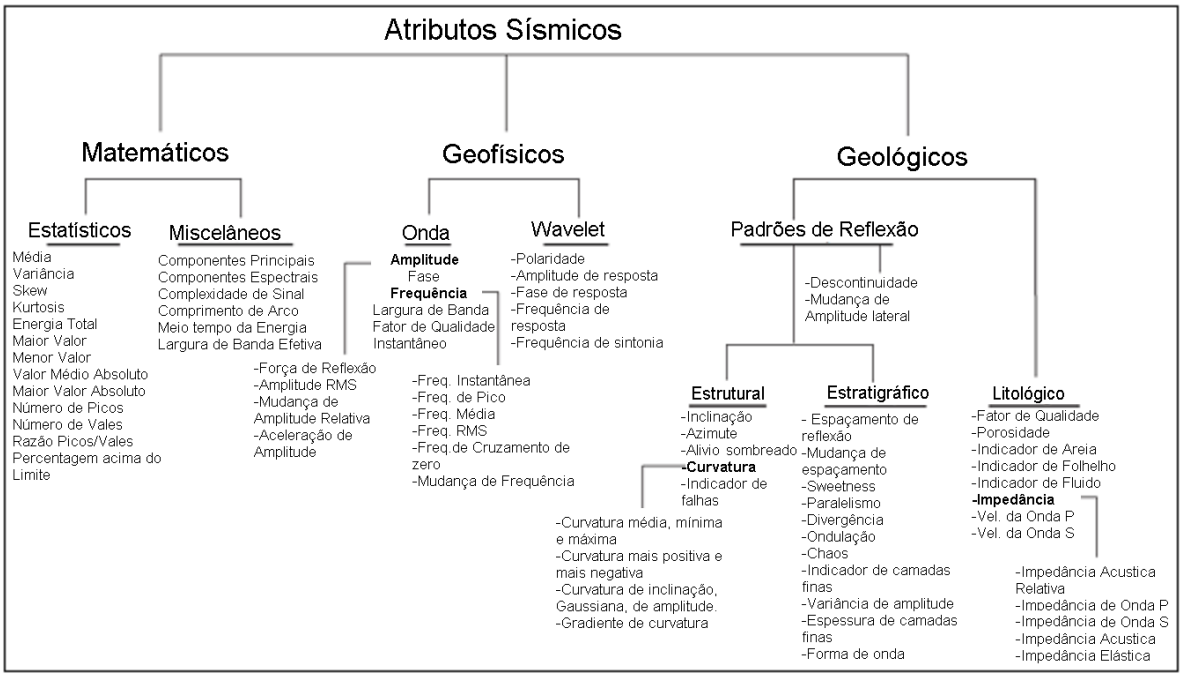

Figura 9 – Coletânea de Atributos sísmicos mais significativos, editado de Barnes (2016).

e presença de hidrocarboneto, terminam assim não coletando amostras de fácies em uma malha regular. Na ausência de poços, ou ainda para estender a abrangência das informações advindas destes, os atributos sísmicos são utilizados pois permitem analisar e identificar sismofácies de uma maneira robusta. No entanto, a utilização de diversos atributos conjuntamente não é simples nem mesmo corriqueira. Cada atributo sísmico possibilita o realce de uma característica específica, por exemplo litologia, indicação de fluído nos poros, etc, Brown (1996).

3.0.1 Resposta sísmica dos evaporitos

Numa sequencia de deposição evaporítica, a distinção dos sais utilizando dados de amplitude é ambigua e difícil. Tanto a anidrita quanto a carnalita apresentam resposta sísmica de pico positiva para negativa, contudo há uma variação de fase de 180 graus. Portanto é difícil saber se estamos entrando numa carnalita ou saindo de uma anidrita, ver Figura 10.

Devido a essa característica, o uso da amplitude sísmica somente, nos leva a situações de ambiguidade e de difícil identificação das sismofácies, como demonstrado por Teixeira et al. (2020). Então para uma melhor definição das camadas delgadas de sal a melhor solução é utilizar a inversão sísmica. Porém como o intuito desse estudo não era realizar um trabalho de inversão, decidimos gerar o atributo de volume de impedância acústica relativa, e fazer desse o atributo principal, ver Figura 11.

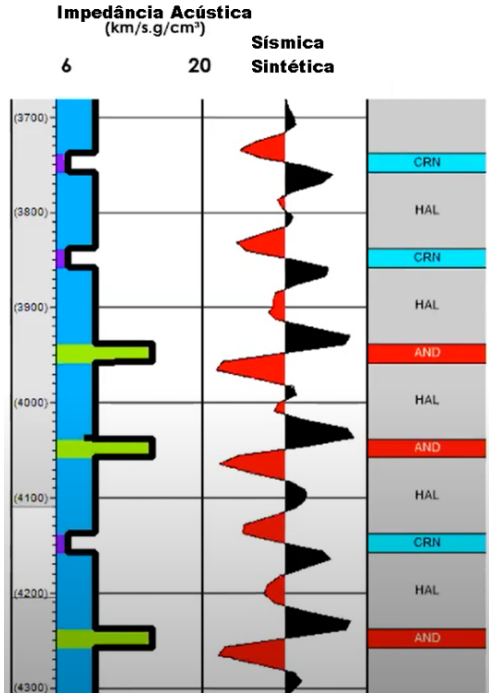

Figura 10 – Anidrita: entrada positiva, saída negativa; Carnalita: entrada negativa, saída positiva. Editado de Teixeira (2020)

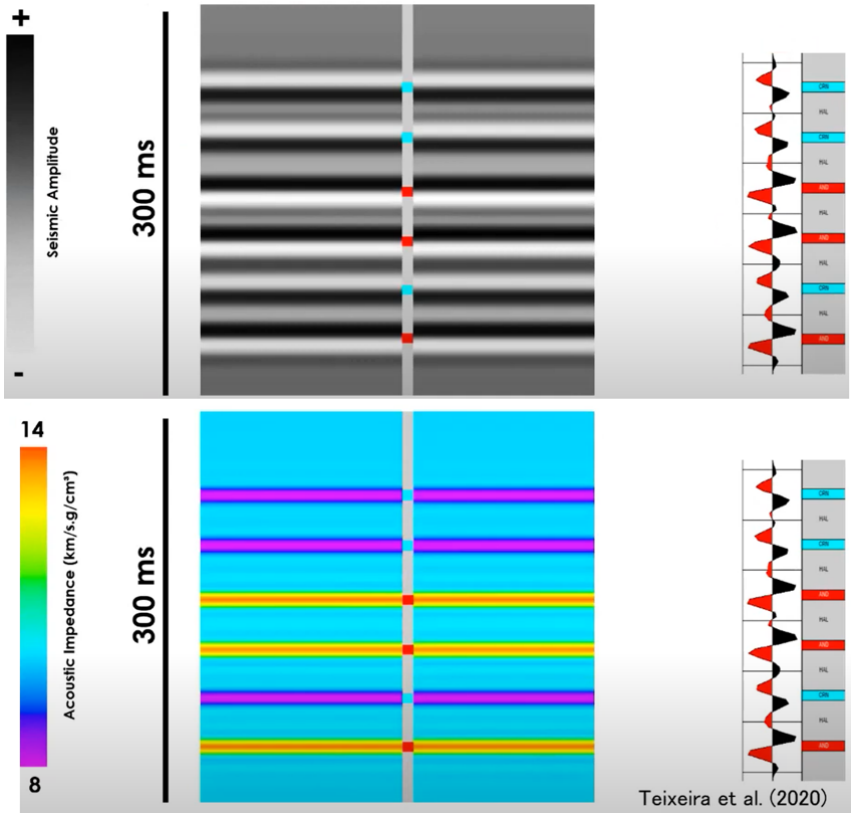

Figura 11 – Comparação da resposta sísmica do atributo de amplitude com o atributo de impedância acústica. Editado de Teixeira (2020)

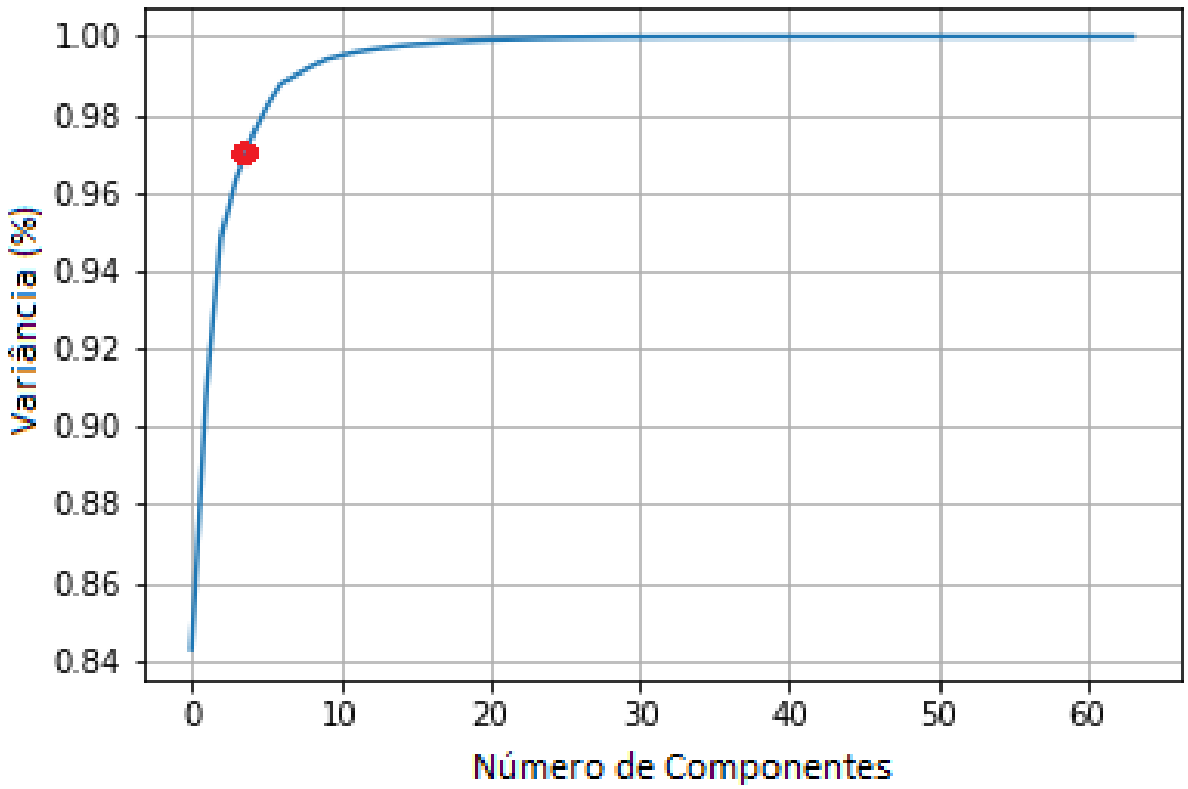

Figura 12 – Variância Explicada do Conjunto de Dados, com quatro atributos temos 97% da variância.

3.0.2 Seleção dos Atributos

Um problema crucial em qualquer análise multiatributo é a seleção e o número de atributos sísmicos a serem usados. Kalkomey (1997) mostrou que a probabilidade de observar uma correlação espúria aumenta à medida que o número de pontos de controle diminui e também à medida que aumenta o número de atributos sísmicos usados. Com base no conhecimento e na experiência, selecionamos oito atributos que melhor destacavam as camadas que estamos focando neste trabalho, o LVS. Em seguida, fizemos uma análise de variância para saber a quantidade razoável de atributos a usar, ver Figura 12.

De acordo com o gráfico da Figura 12, com apenas 4 atributos temos 97% da variância do conjunto de dados. Portanto, para evitar redundância e um possível *overfitting* do modelo, sabendo que só precisamos de 4 atributos, eliminaremos os que tiverem alta correlação, e para isso fizemos uma matriz de correlação, o resultado pode ser visto na Tabela 1.

Sabendo que nosso principal atributo seria a impedância acústica relativa (RAI), os melhores atributos neste caso seriam: 1^a derivada, Evidência de bordas, Amplitude RMS, além da própria RAI. Porém de acordo com a forma proposta por Hampson, Schuelke

	Contraste de Amplitude	Impedância Acustica Relativa	Frequência Instantânea	1a Derivada	TECVA	Amplitude	Amplitude RMS	Evidência de Bordas
Contraste de Amplitude	1	0,865	0,004	0,018	0,482	0,175	0,173	0,034
Impedância Acustica Relativa		1	0,988	0,208	0,812	0,585	0,584	0,016
Frequência Instantânea			1	0,109	0,991	0,149	0,197	0,056
1a Derivada				1	0,037	0,149	0,118	0,018
TECVA					1	0,122	0,080	0,482
Amplitude						1	0,674	0,023
Amplitude RMS							1	0,026
Evidência de Bordas								1

Tabela 1 – Correlação dos Atributos Propostos

e Quirein (2001) o melhor atributo a ser usado é aquele que destaca a reflexão que esta interessado quando comparado com o perfil do poço, e neste caso, apesar da 1^a derivada apresentar um bom ranking (baixa correlação com a RAI) ela foi descartada, devido a não representar bem as camadas de sal.

Como de acordo com a variância explicada do conjunto de dados necessitamos pelo menos quatro atributos, decidimos então utilizar também a amplitude original no lugar da 1^a derivada. Abaixo falaremos um pouco mais sobre cada um:

a) Amplitude: a amplitude sísmica de reflexão é um atributo relacionado às propriedades físicas da subsuperfície em função da refletividade nas interfaces acústicas, ver Figura 13;

$$A = A(\rho, V_p, V_s, \theta, t)$$

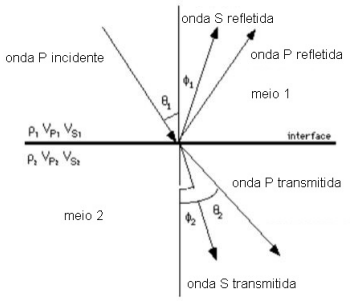

Figura 13 – Amplitude como função da densidade ρ, Velocidade Compressional Vp, Velocidade Cisalhante Vs, angulo de incidência θ, e tempo t.

b) Amplitude *Root Mean Square* (RMS): é um atributo básico de ampli-

tude, uma medida estatística da magnitude de variação na amplitude ao longo de um conjunto de dados. Muito utilizado em caracterização de reservatórios, como nos trabalhos de Sarhan (2017) e Hossain (2020). Geralmente, variações mais altas da impedância acústica (associadas a variações na litologia empilhada) resultam em valores RMS mais altos. É calculado através de uma janela cônica deslizante de N amostras como a raiz quadrada da soma de todos os valores dos traços x ao quadrado, onde w e n são os valores da janela, conforme apresentado na Equação 3.1;

$$x_{rms} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} w_n x_n^2} \tag{3.1}$$

c) Impedância Acústica Relativa (RAI): a impedância é um método estratigráfico, e é o produto da densidade e velocidade sísmica, que varia entre as diferentes camadas rochosas e geralmente simbolizado como Z. A RAI tenta estimar esta impedância (relativa) obtida da amplitude, gerando uma pseudo propriedade de camada (que seria gerada num processo de inversão, de fato).Esse atributo mostra contrastes acústicos aparentes, indica limites de sequências estratigráficas, superfícies de inconformidade e descontinuidades, também pode indicar porosidade ou conteúdo de fluido no reservatório.

Define o contraste de densidade encontrado na interface entre duas litologias distintas e é calculado fazendo-se a integral do traço e passando o resultado por um filtro passa-alta para reduzir o ruído de baixa frequência potencialmente introduzido, Connolly (1999). Neste caso, uma frequência de corte de 10 Hz foi utilizada, pois era a frequência dominante da *wavelet* na camada de sal. Pode ser expressa como na equação 3.2:

$$< Z_n >=< Z_0 > \exp\left(2\sum_{j=0}^{n} R_j \Delta t\right) \tag{3.2}$$

Onde $< Z >$ indica a impedância média sobre uma camada, R é a refletividade, j é a amostra de tempo variando de 0 a n, e t é o tempo.

d) Evidência de Borda (Método Estrutural): é um método de aprimoramento estatístico de borda usado para delinear falhas e bordas do corpo de sal dentro dos dados sísmicos. O algoritmo está relacionado à transformação Radon e Hough mas usa uma integral para detectar arestas dentro de uma imagem e é limitado a uma janela definida pelo usuário. Em matemática, a transformação Radon é a transformação integral que leva uma função f definida no plano para uma função Rf definida no espaço (bidimensional)

de linhas no plano, cujo valor em uma linha específica é igual à integral de linha da função nessa linha.

A transformação Hough é uma técnica de extração de características usada na análise de imagens, visão computacional e processamento digital de imagens. O objetivo da técnica é encontrar instâncias imperfeitas de objetos dentro de uma determinada classe de formas por um procedimento de votação. Esse procedimento de votação é realizado em um espaço de atributos, a partir do qual os candidatos a objetos são obtidos como máximos locais em um espaço chamado acumulador que é explicitamente construído pelo algoritmo para calcular a transformação Hough. O atributo de evidência de borda funciona pesquisando localmente em todas as direções para segmentos de linha em que os valores na linha diferem significativamente dos valores circundantes, Aqrawi e Boe (2011).

Os 4 atributos usados podem ser vistos na Figura 14, onde mostramos os 4 atributos extraídos sobre uma mesma *inline*.

Embora algoritmos de computador sejam capazes de extrair padrões (aqueles resultantes de um filtro por exemplo), um computador na verdade não consegue ”ver” padrões da mesma maneira que um intérprete humano conseguiria. Com essa informação, podemos quantificar diferentes componentes do padrão de amplitude sísmica em um único *voxel* usando um conjunto de atributos sísmicos.

Um *voxel* representa uma única amostra, ou dado pontual, em uma malha regular tridimensional. Este dado pontual pode consistir em um único aspecto da informação, tal como a opacidade, ou múltiplos aspectos, tais como cor, além do aspecto visual. Esse ”vetor” de atributos sísmicos pode então ser interpretado em cortes verticais, de tempo ou de horizontes, usando visualização 3D ou ferramentas de corenderização de multiatributos.

Idealmente, corenderizar 3 atributos contra um mapa de cores Vermelho-Verde-Azul (do inglês RGB) ou Matiz-Iluminação-Saturação (do inglês HLS), resulta numa cor específica correlacionando a deformação estrutural, característica da fácie de interesse, etc. Porém, quando há mais de 3 atributos, o intérprete pode pintar áreas de interesse candidatas na combinação dos primeiros 3 atributos e então modificar a área pintada através de combinações subsequentes, ou o intérprete pode combinar quatro ou mais atributos diretamente usando um algoritmo de *machine learning*, como veremos no próximo capítulo.

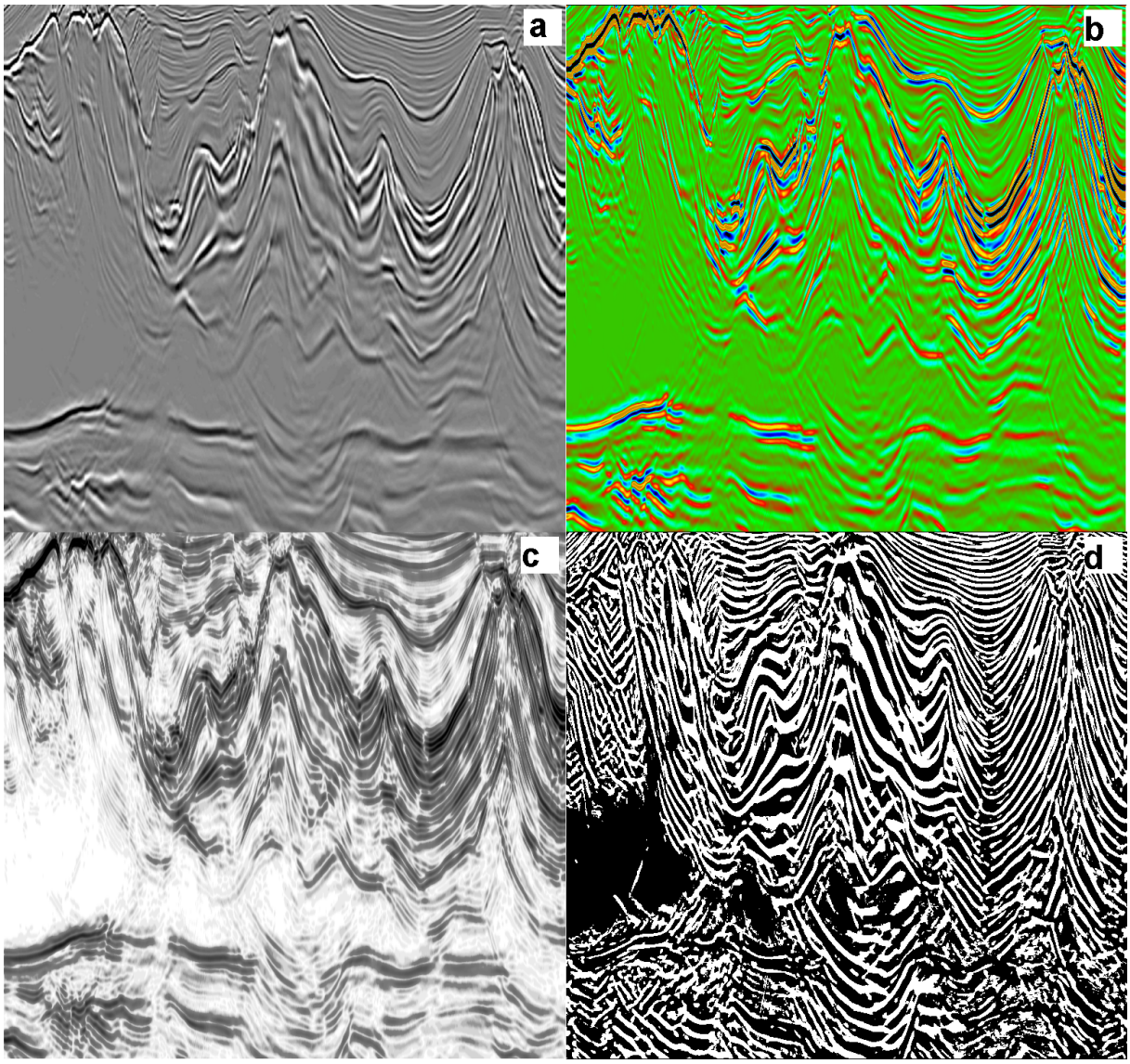

Figura 14 – Atributos usados nesta classificação de sismofácies. a) Amplitude; b)Impedância Acústica Relativa; c) Amplitude RMS; d) Evidência de Borda.

4 Machine Learning na Identificação de Litofácies

Um dos objetivos da caracterização de reservatórios é mapear quantitativamente a distribuição espacial de suas heterogeinidades e propriedades relacionadas. Similarmente a um reservatório, podemos caracterizar a camada de sal utilizando técnicas equivalentes. Com a disponibilidade de dados sísmicos 3D é possível usar redes neurais para descobrir relacionamentos entre atributos sísmicos e parâmetros de reservatórios. Da mesma forma podemos utilizar as redes neurais para encontrar litofácies, considerando que a camada de sal é formada por sais de diferentes velocidades, cada interface gerará uma reflexão, a qual pode ser vista na sísmica. Atributos também são utilizados para ressaltar certas características e para aumentar a quantidade de dados disponíveis sobre determinada área.

Uma tentativa de reconhecimento automatizado de padrões levou a forma das redes neurais, Russell et al. (1997), em que um conjunto de padrões de entrada está relacionado à saída por uma transformação que é codificada nos pesos da rede. A classificação de dados, que é bastante comum em todas as áreas, inclusive a geociências, vem sendo executada por diferentes autores já a algum tempo em diferentes contextos, temos como exemplo Carneiro et al. (2012); Dumay e Fournier (1988); Wu et al. (2019); Lai et al. (2016).

Todos esses métodos, embora tenham facilitado o trabalho do intérprete na identificação de fácies sísmicas desejadas em grandes volumes sísmicos 3D, não substitui o intérprete, pois o mesmo ainda precisa fornecer os dados iniciais de treinamento para cada levantamento, verificar a exatitude do resultado, etc. Muitos trabalhos tem se utilizado de aprendizado de maquina para este fim, uma das técnicas mais bem sucedidas e utilizadas é o Mapa Auto Organizável, do inglês *Self Organizing Maps* (SOM), seguido de uma rede neural supervisionada, como o trabalho de Moqbel e Wang (2011).

Um SOM com aprendizado não supervisionado, realiza uma projeção de atributos não lineares (não no sentido de atributos sísmicos), de um espaço multi-dimensional para um espaço bi dimensional, além da possibilidade de clusterizar, classificar e visualizar os dados. Neste trabalho ao invés de utilizar o SOM, separo todos estes passos para um melhor controle num fluxo de trabalho de *machine learning* e utilizo as técnicas mais avançadas neste processo, tais como UMAP e HDBSCAN. Neste capítulo dissertarei sobre cada uma destas técnicas e sua aplicação no fluxo de trabalho.

4.1 Conjunto de Dados e Aplicativos Utilizados

Para a execução do projeto, dados sísmicos e de poços adquiridos para reservatórios do Pré-Sal da Bacia de Santos foram autorizados pela ANP (Agência Nacional do Petróleo) e disponibilizados pela Petróleo Brasileiro S.A. (Petrobras). O conjunto de dados sísmicos é composto por um volume sísmico (área de aproximadamente 300 km^2), oriundo do processamento sísmico PSDM (*Post-Stack Depth Migration*), fornecido no domínio da profundidade (dimensão vertical total de 12 km). Já o conjunto de dados 1D contemplou 30 poços existentes na mesma área, incluindo perfis litológicos, perfis elétricos (*Caliper*, Raios Gama, Resistividade, Sônico, Densidade, Neutrão, etc.) e dados direcionais.

A distribuição em área dos volumes sísmicos e dos poços mencionados pode ser visualizada na Figura 15. Para o desenvolvimento do projeto foi disponibilizado pela Schlumberger um pacote de licenças da plataforma Petrel 2016®. Algoritmos foram desenvolvidos utilizando a linguagem de código aberto Python 3.6 para carregamento dos dados e todas as outras etapas de criação do modelo e classificação.

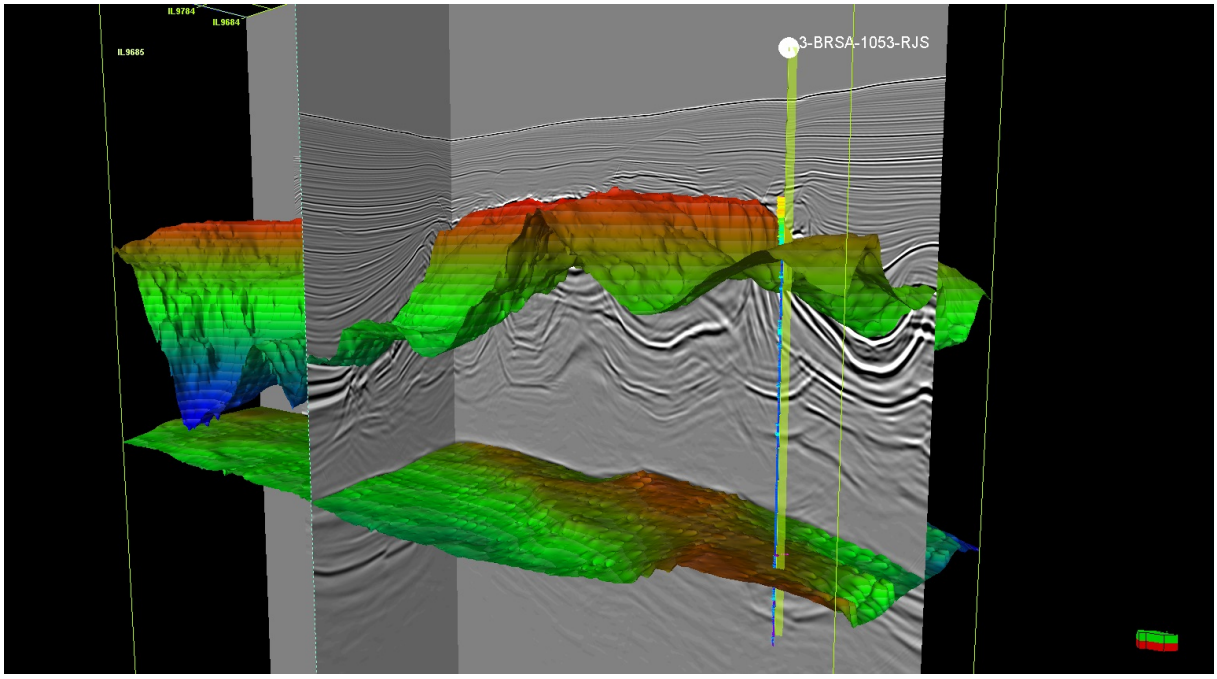

Figura 15 – Recorte do volume sísmico utilizado como entrada do modelo para classificação, delimitado pelos horizontes do topo e fundo do sal.

4.2 Aprendizado Supervisionado vs Não Supervisionado

O aprendizado supervisionado é usado quando existem dados de treinamento disponíveis que fornecem informações básicas sobre o padrão a ser descoberto, o chamado *Ground-Truth*. Os dados de treinamento podem ser perfis de poço, volumes de produção ou polígonos sísmicos fornecidos pelo intérprete, definindo fácies sísmicas específicas. No

machine learning supervisionado, não há modelo, e o objetivo é descobrir a relação entre um determinado par de dados de entrada e saída. Nesse sentido, o computador está aprendendo por emulação como integrar vários tipos de dados para produzir o resultado desejado.

O aprendizado não supervisionado ainda tem alguma supervisão, na medida em que o intérprete seleciona ou descarta os atributos que devem ser usados no fluxo de trabalho. O aprendizado não supervisionado pode usar algoritmos semelhantes aos usados no aprendizado supervisionado, mas, no caso não supervisionado, o computador tenta identificar um número finito de *clusters* que representam o comportamento da maioria dos dados.

Uma limitação do aprendizado supervisionado é que o controle do poço pode falhar em representar todas as fácies possíveis ou relações petrofísicas-sísmicas que são representadas pelos dados sísmicos. A limitação do aprendizado não supervisionado é bem diferente. Aqui, o intérprete deixa os dados falarem por si mesmos, pelo qual o software constrói um número de *clusters* que representam a resposta em massa do volume de dados sísmicos.

4.3 Operações com Janela Deslizante

Visto que tanto o aprendizado supervisionado como o não supervisionado tem suas vantagens e desvantagens e nenhum deles consegue resolver o problema por inteiro, resolvemos utilizar um método em dois estágios, porém ao invés de usar o SOM na fase não supervisionada usamos técnicas mais modernas como o UMAP e o HDBSCAN e na fase supervisionada utilizamos o KNN. Após selecionarmos os dados, *inlines/crosslines* sobre os poços que contém os sais que estamos buscando, fazemos um "corte" nos dados, selecionando os traços desde a base até o topo do sal, num raio de 50 traços para cada lado do poço, como na Figura 16.

Após a seleção dos dados usamos uma técnica chamada "janela deslizante", que também serve para fazer um "aumento de dados", como na Figura 17. Operações com janela deslizante são realizadas frequentemente em análise e processamento de imagens. Uma típica operação de Janela Deslizante, do inglês, *Sliding Window Operation* (SWO), aplica repetidamente um filtro de imagem em uma sub-janela pré definida que desliza progressivamente através da imagem alvo. Muitos filtros envolvem operações lógicas e matemáticas de alta complexidade, por exemplo, filtros de *rank* de ordem que envolvem sortear os valores dos *pixels* na janela em ordem ascendente. Filtros morfológicos que realizam operações como erosão, dilatação, abertura e fechamento usando uma janela móvel.

Em um filtro de imagem tipico, uma janela definida pelo usuário escaneia

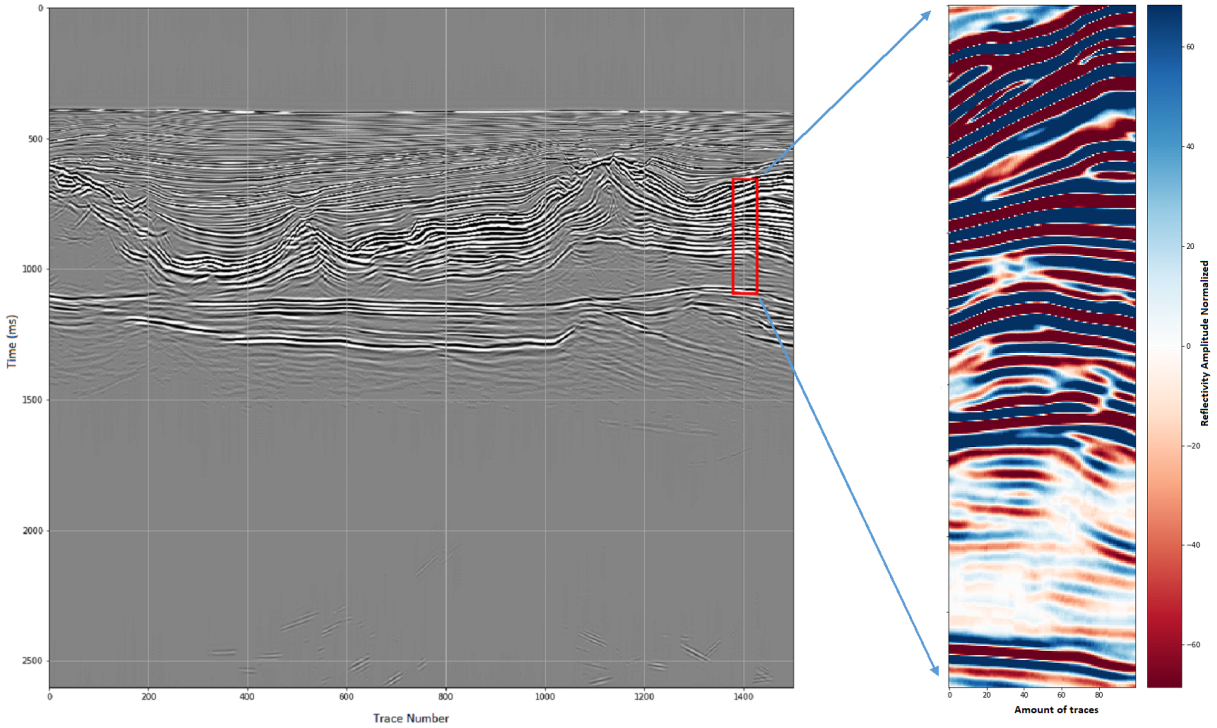

Figura 16 – Recorte de dados diretamente sobre a locação do poço, limitado entre o topo e a base do sal.

e rasteriza até que toda a imagem seja coberta. Denotamos uma imagem de tamanho $N \times M$ como $D_{n,m} : n = 1, .., N; m = 1, .., M$; e uma janela deslizante de tamanho $I \times J$ como $W_{i,j} : i = 1, .., I; j = 1, .., J$; e o conjunto de passos r dentro desta janela deslizante de $P(p1, .., T)$. Se o tempo de execução das operações $P(p1, .., T)$ é t, então o tempo de execução da janela deslizante é : $T = N \times M \times I \times J \times t$ Uma vez q selecionamos as *inlines/crosslines* e a região de interesse, fazemos a janela "deslizar" sobre a zona de interesse e acumular esses valores num vetor, que concatenado junto com outros de outros poços irá formar o conjunto de dados que servira como entrada para a rede neural.

Como a janela é do formato 8 x 8, terminamos com um vetor de entrada de dimensão 64. Essa janela pode ter qualquer tamanho desde 1 x 1 até o que for práticode de acordo com o dado sendo utilizado. Com isso, nesse caso, teremos que reduzir a dimensão para dois ou três para que possamos trabalhar, para isso utilizamos um algoritmo relativamente novo chamado Projeção por Variedade e Aproximaçao Uniforme, do inglês, *Uniform manifold Approximation and Projection* (UMAP), McInnes, Healy e Melville (2018).

4.4 UMAP

A redução de dimensão é um dos passos de pré-processamento largamente utilizado num fluxo de *machine learning*, e busca produzir uma representação em poucas

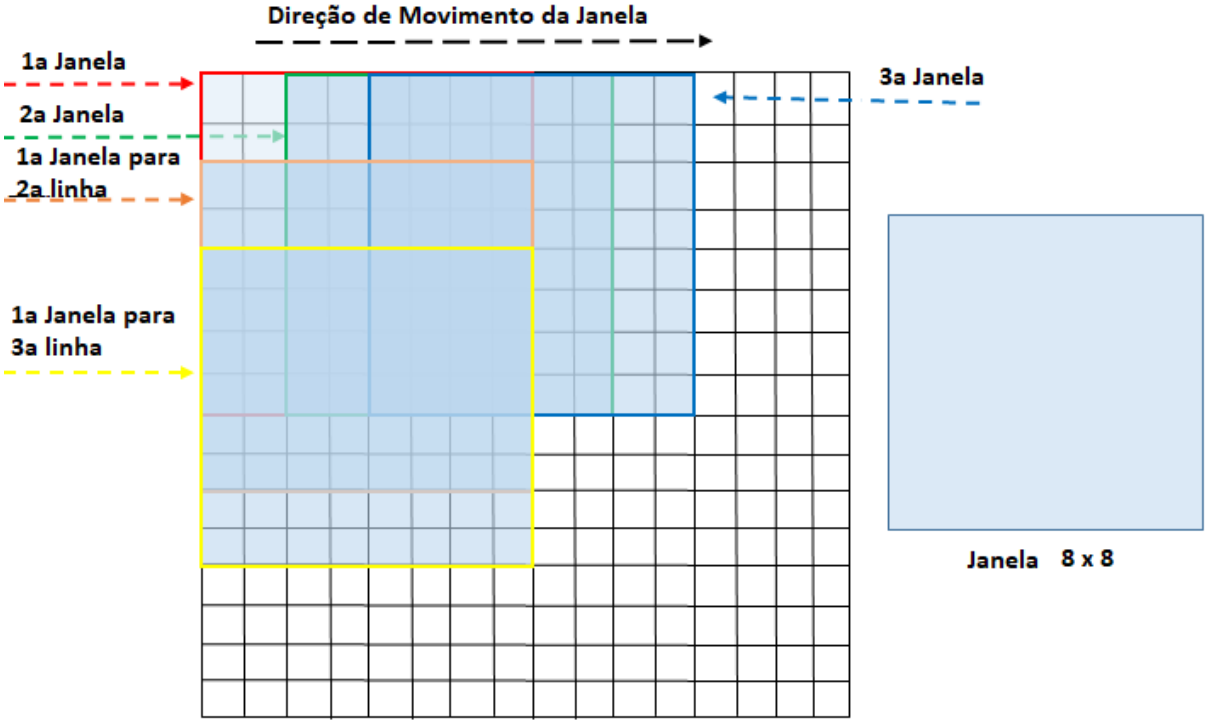

Figura 17 – Exemplo de janela deslizante 8x8.

dimensões de um conjunto de dados de grandes dimensões preservando a estrutura relevante. Algoritmos de redução de dimensão se dividem em duas categorias: aqueles que buscam preservar a estrutura de distância dentro do dado e aqueles que favorecem a preservação das distancias locais sobre as globais. o UMAP é uma técnica de aprendizado de *manifold* para redução de dimensionalidade, com uma estrutura teórica baseada em geometria Riemaniana e topologia algébrica. Também usada para visualização e com a vantagem de preservar a estrutura global dos dados.

4.4.1 Fundamentação Teórica do UMAP

Este algoritmo é fortemente baseado em teoria de *manifolds* e análise topológica de dados. Em um alto nível, UMAP usa aproximações de *manifold* locais e une suas representações de conjunto difusas locais para construir uma representação topológica dos dados em alta dimensão. Dada uma representação de baixa dimensão dos dados, um processo similar pode ser usado para construir um representação topológica equivalente. UMAP então otimiza o layout da representação dos dados no espaço de baixa dimensão,para minimizar a cross-entropia entre as duas representações topológicas. Uma das vantagens do UMAP com relação a outros algoritmos de redução de dimensionalidade é a não necessidade de um pré-condicionamento dos dados utilizando uma Análise de Componentes Principais por exemplo. A construção de representações topológicas difusas podem ser divididas em dois casos:

a) aproximar um *manifold* no qual se assume que o dado esta localizado;

b) construir uma representação de conjunto simplista difusa do *manifold* aproximado.

4.4.2 Distribuição Uniforme dos Dados no *manifold* e Aproximação Geodésica

O primeiro passo é aproximar o *manifold* onde se supõe que os dados estão. O *manifold* pode ser conhecido a priori (como simplesmente R^n) ou pode necessitar ser inferido dos dados. Computacionalmente, é somente tratável como um esqueleto que pode ser descrito em termos de construção e operação como um grafo ponderado, e isso situa o UMAP na classe de algoritmos de aprendizado de grafos baseados em K-vizinhos, tais como t-SNE, Isomap e Eigenmaps Laplacianos. Assim como outros algoritmos de grafos baseados em K-vizinhos é dividido em duas etapas:

a) na primeira parte, um particular grafo ponderado em K-vizinhos é construído.

b) na segunda parte, um layout de baixa dimensão deste grafo é calculado.

4.4.2.1 Construção do Grafo

A primeira fase do UMAP pode ser pensada como a construção do grafo ponderado em K-vizinhos, onde K representa o numero de vizinhos, seja $X = x_1, ..., x_n$ o conjunto de dados de entrada, com uma métrica (ou medida de dissimilaridade) $d : X \times X \to R \geq 0$. Dado um hiper-parâmetro de entrada K , para cada x_i calcula-se o conjunto $\{x_{i1}, ..., x_{ik}\}$ dos K vizinhos de x_i sob a métrica d. Este cálculo pode ser realizado através de qualquer algoritmo de busca de vizinho mais próximo ou vizinho mais próximo aproximado, nesse caso UMAP usa o algoritmo de vizinho mais próximo descendente. Para cada x_i definiremos ρ_i e σ_i . Sejam:

$$\rho_i = min\{d(x_i, x_{i_j}) \mid 1 \leq j \leq k, d(x_i, x_{i_j}) > 0\} \tag{4.1}$$

e colocamos σ_i para ser o valor tal que:

$$\sum_{j=1}^{k} exp(\frac{-max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}) = log_2(k) \tag{4.2}$$

Podemos agora definir um grafo ponderado direcionado $\bar{G} = (V, E, \omega)$. Os vértices V de $\bar{G}$ são simplesmente o conjunto X. Podemos então formar as arestas direcionadas $E = \{(x_i, x_{ij}) | 1 \leq j \leq k, 1 \leq i \leq N\}$, e definimos a função peso ω definindo:

$$\omega((x_i, x_{i_j})) = exp(\frac{-max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}) \tag{4.3}$$

Seja A a matriz de adjacência ponderada de $\bar{G}$, e considere a matriz simétrica: $B = A + AT - A \circ AT$, onde $\circ$ é o produto Hadamard (ou ponto de sentido). O grafo de UMAP G é então um grafo ponderado não direcionado cuja matriz adjacente é dada por B.

4.4.2.2 Layout do Grafo

Na prática UMAP usa um algoritmo de layout de grafo de força direcionada em um espaço de baixa dimensão. Esse algoritmo utiliza um conjunto de forças atrativas aplicadas ao longo das arestas e um conjunto de forças repulsivas aplicadas entre os vértices. Qualquer algoritmo desse tipo requer uma descrição tanto das forças atrativas e repulsivas. O algoritmo procede aplicando iterativamente forças atrativas e repulsivas em cada aresta. Convergência é conseguida diminuindo-se lentamente as forças atrativas e repulsivas numa maneira similar aquela usada em anelamento simulado. Em UMAP a força atrativa entre dois vértices i e j em coordenadas y_i e y_j respectivamente, é determinado por:

$$\frac{-2ab\,\|y_i - y_j\|_2^{2(b-1)}}{1 + \|y_i - y_j\|_2^2} \times \omega((x_i, x_{i_j}))(y_i - y_j) \tag{4.4}$$

onde a e b são hiper-parâmetros.

Forças repulsivas são calculadas por amostragem para diminuir as restrições computacionais, então quando uma força atrativa é aplicada a uma aresta, um dos vértices daquela aresta é repelida por uma amostragem dos outros vértices. A força repulsiva é dada por :

$$\frac{b}{(\epsilon + \|y_i - y_j\|_2^2)(1 + \|y_i - y_j\|_2^2)} \times (1 - \omega((x_i, x_{i_j}))(y_i - y_j)) \tag{4.5}$$

onde ϵ é um numero pequeno para prevenir divisão por zero (0.001).

O algoritmo pode ser inicializado aleatoriamente mas na prática, já que o Laplaciano simétrico do grafo G é uma aproximação discreta do operador do *manifold* Laplace-Beltrami, podemos usar um layout espectral para inicializar o *embedding*. O que resulta numa convergência mais rápida e maior estabilidade do algoritmo.

4.4.3 Implementação e Hiper-Parâmetros

Implementação prática deste algoritmo requer o cálculo do K-vizinho mais próximo e otimização via gradiente descendente estocástico.

O algoritmo requer quatro hiper-parâmetros:

a) n, o numero de vizinhos a considerar quando aproximando a métrica local;

b) d, a dimensão de *embedding* alvo;

c) min-dist, a separação desejada entre pontos próximos no espaço *embedding*;

d) n-epochs, o numero de épocas de treinamento a usar quando otimizando a representação de baixa dimensão.

Pode-se interpretar o número de vizinhos n como a escala local na qual considerar o *manifold* como aproximadamente plano, com a estimação do *manifold* sendo a média sobre os n vizinhos. Características do *manifold* que ocorrerem em menor escala que dentro dos pontos de n vizinhos mais próximos serão perdidos, enquanto que características de *manifold* de larga escala que não podem ser vistas ao juntar gráficos localmente planos na escala dos n vizinhos mais próximos podem não ser bem detectados. Então n representa algum grau de troca entre entre as características de *manifold* de larga escala e de granulação fina, valores menores garantirão que a estrutura detalhada do *manifold* seja capturada com precisão (a custa de perder a "visão geral" do *manifold*), enquanto valores maiores capturarão estruturas de *manifold* de larga escala, mas a custa de uma perda de estrutura de detalhes finos que será calculada em média nas aproximações locais. Com menores valores de n o *manifold* tende a ser quebrado em muitos pequenos componentes conectados.

Em contraste min-dist é um hiper-parâmetro que afeta diretamente a saída, pois controla a construção do conjunto simplicial difuso da representação de baixa dimensão. Age como a distancia até o vizinho mais próximo usado para garantir a conectividade local. Em essência determina quão próximos os pontos podem ser agrupados no espaço de baixa dimensão. Pequenos valores de min-dist resultarão em regiões densamente compactadas, mas que representarão mais fielmente a estrutura do *manifold*. Aumentar o valor de min-dist forçará o *embedding* a espalhar mais os pontos, ajudando na visualização, sendo portanto um parâmetro mais estético, como visto na Figura 18.

Após reduzirmos a dimensão dos dados temos que agrupa-los em *clusters*, para isso usamos um algoritmo também inovador do mesmo autor, HDBSCAN, sobre o qual falaremos a seguir.

4.5 HDBSCAN - Agrupamento Hierarquico Espacial de Aplicações com Ruído Baseado em Densidade

Do inglês, *Hierarchical Density Based Spatial Clustering of Applications with Noise*, é um conjunto de ferramentas para usar aprendizado não supervisionado para encontrar *clusters*, ou regiões densas de um conjunto de dados. Assim como o DBSCAN, inicialmente proposto por Ester et al. (1996) e posteriormente aprimorado por Campello, Moulavi e Sander (2013), o HDBSCAN pode ser considerado como uma evolução do

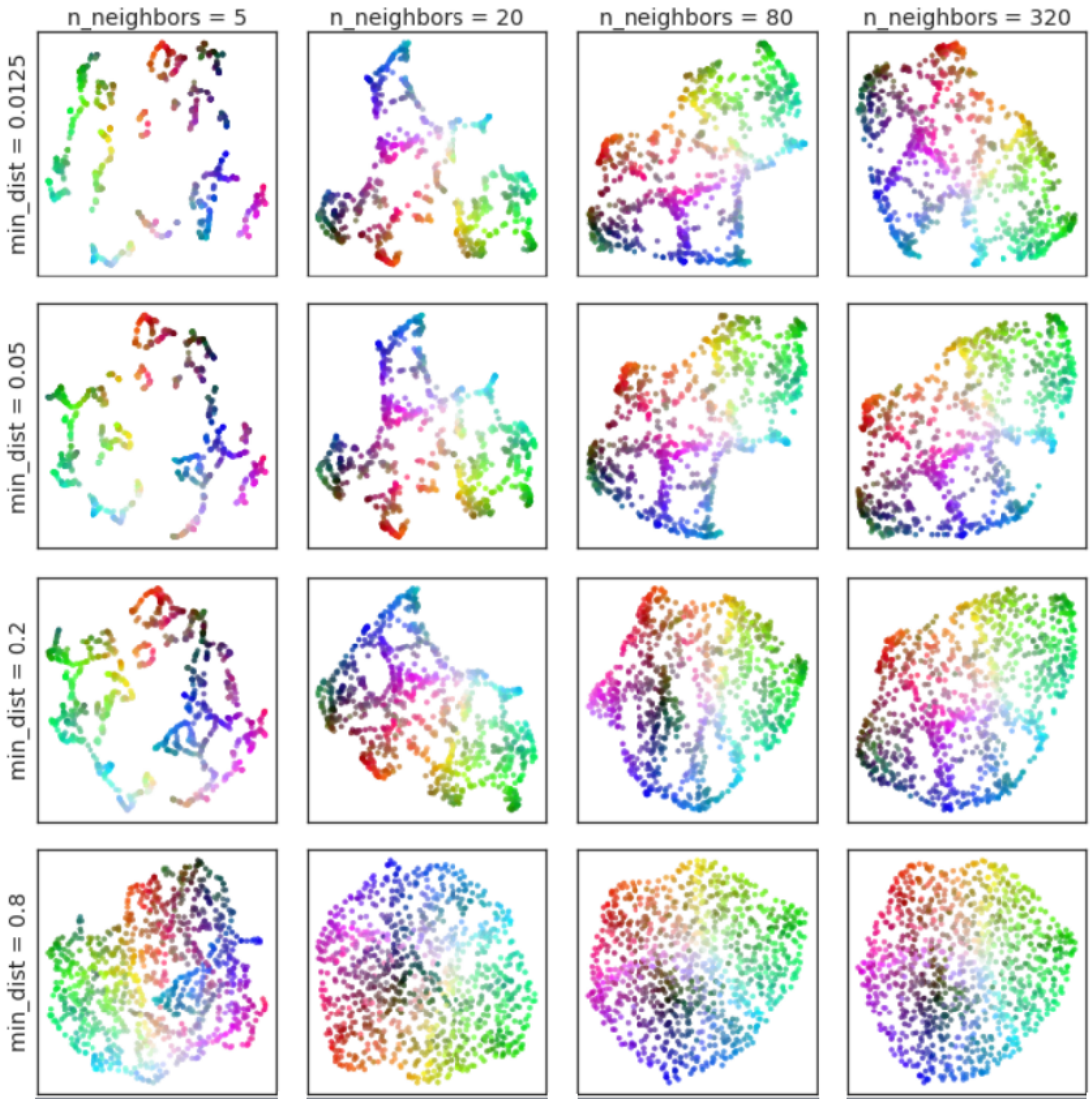

Figura 18 – Variação dos hiper-parâmetros n e min-dist resultam em diferentes *embeddings*. Os dados são amostras aleatórias de um cubo colorido 3-dimensional. Retirado de UMAP - McInnes, Healy e Melville (2018).

mesmo, visto que enquanto o DBSCAN necessita como parâmetros de entrada o tamanho mínimo do *cluster* e o valor de ϵ (epsilon), HDBSCAN somente necessita do tamanho mínimo do *cluster*. Nesse caso epsilon é o raio no qual os pontos vizinhos tem de estar para que o núcleo do *cluster* possa se formar, enquanto que no HDBSCAN o valor de ϵ é variável para cada *cluster*.

4.5.1 Clusterização em *machine learning*

É basicamente um tipo de método de aprendizado não supervisionado, no qual extraímos referências de conjuntos de dados que consistem em dados de entrada sem respostas rotuladas. Geralmente, é usado como um processo para encontrar estrutura significativa, processos subjacentes explicativos, recursos generativos e agrupamentos

inerentes a um conjunto de exemplos. Clusterizar é a tarefa de dividir a população ou os pontos de dados em vários grupos, de modo que os pontos de dados nos mesmos grupos sejam mais semelhantes a outros pontos de dados no mesmo grupo e diferentes dos pontos de dados em outros grupos. É basicamente uma coleção de objetos com base na semelhança e dissimilaridade entre eles.

4.5.1.1 Métodos de Clusterização

a) Métodos baseados em densidade: Esses métodos consideram os aglomerados como a região densa com alguma semelhança e diferente da região menos densa do espaço. Esses métodos têm boa precisão e capacidade de mesclar dois *clusters*. Exemplo: DBSCAN (clusterização espacial de aplicativos com ruído baseado em densidade) Ester et al. (1996), OPTICS (pontos de ordenação para identificar estrutura de *cluster*), Ankerst et al. (1999), etc.

b) Métodos hierárquicos: os *clusters* formados nesse método formam uma estrutura do tipo árvore com base na hierarquia. Novos *clusters* são formados usando o anteriormente formado. É dividido em duas categorias:

- Aglomerativo (abordagem ascendente)
- Divisivo (abordagem de cima para baixo)

c) Métodos de particionamento: esses métodos particionam os objetos em k *clusters* e cada partição forma um *cluster*. Este método é usado para otimizar uma função de similaridade de critério objetivo, como quando a distância é um exemplo de parâmetro principal, exemplo: K-Means, MacQueen (1967), CLARANS (Clusterizando Grandes Aplicações baseado em Pesquisa Aleatória), Ng e Han (2002), etc.

d) Métodos baseados em grade: nesse método, o espaço de dados é formulado em um número finito de células que formam uma estrutura semelhante a uma grade. Todas as operações de armazenamento em *cluster* realizadas nessas grades são rápidas e independentes do número de objetos de dados, como STING (Statistical Information Grid), Wang, Yang e Muntz (1997), CLIQUE (Clustering In Quest), Agrawal et al. (1998), etc.

Portanto, baseado nas afirmativas acima, o método de clusterização escolhido deve ser em função dos dados, neste caso quando temos:

a) *clusters* de forma arbitrária;

b) *clusters* com diferentes tamanhos e densidades;

c) Ruído.

HBSCAN é o método mais apropriado. Como podemos ver na Tabela 2, nossos dados são fortemente desbalanceados devido as diferentes proporções de sal encontradas na Bacia de Santos, que pode ser visto num levantamento feito por Maul, Santos e Silva (2018), em 182 poços perfurados.

Campo	#Poços	%LVS	%Halita	%HVS
1	20	8	83	8
2	29	9	82	9
3	17	12	77	12
4	3	13	71	16
5	5	3	84	13
6	7	3	80	17
7	72	8	81	11
8	25	4	88	8
9	4	6	81	13
Total Poços	182			
Média		7	81	12

Tabela 2 – Adaptado de Maul, Santos e Silva (2018). HVS - Proporções de sal na Bacia de Santos, HVS - Sais de Alta velocidade, LVS - Sais de Baixa Velocidade.

4.5.2 Como HDBSCAN funciona

HDBSCAN estende as funcionalidades do DBSCAN, Ester et al. (1996), que é um algoritmo de clusterização baseado em densidade, convertendo-o em um algoritmo baseado em hierarquia, e então extrai *clusters* simples baseado na estabilidade dos mesmos, isso é feito seguindo-se uma série de passos:

a) Transformar o espaço de acordo com densidade/esparsidade;

b) Criar a árvore de abrangência mínima do gráfico ponderado pela distância;

c) Construir uma hierarquia de *clusters* de componentes conectados;

d) Condensar a hierarquia de *cluster* baseado no tamanho mínimo do *cluster*;

e) Extrair os *clusters* estáveis da árvore condensada.

4.5.2.1 Transformar o espaço

Precisamos distinguir o que é dado do que é ruido, para isso precisamos da estimativa de densidade e a maneira mais simples é a distância para o k-ésimo vizinho mais próximo, se tivermos a matriz de distância para nossos dados, pode-se usar diretamente essa matriz. Caso contrário pode se usar uma métrica de distância, que é chamada **distância núcleo** definida pelo parâmetro k para um ponto x e denotado como $\text{núcleo}_k(x)$. Uma vez feito isso precisamos de uma maneira de separar os pontos de baixa densidade, a maneira

mais simples é definir uma nova métrica de distância entre os pontos, que é chamada de **distância de alcançabilidade mútua**, que é definida como:

$$d_{mreach-k}(a,b) = max\{core_k(a), core_k(b), d(a,b)\} \tag{4.6}$$

Onde $d(a,b)$ é a distância métrica original entre a e b. Sob essa métrica, os pontos densos (com baixa distância do núcleo) permanecem à mesma distância um do outro, mas os pontos mais esparsos são afastados para que estejam pelo menos a distância núcleo de qualquer outro ponto. Com a ressalva de que isso depende da escolha de k, valores maiores interpretam mais pontos como sendo ruído.

4.5.2.2 Construir a árvore de abrangência mínima

Uma vez que temos uma nova métrica de alcançabilidade mútua nos dados, queremos começar a encontrar os aglomerados com dados densos. Áreas densas são relativas, e grupos diferentes podem ter densidades diferentes. Conceitualmente, considera-se os dados como um gráfico ponderado com os pontos de dados como vértices e uma aresta entre dois pontos com peso igual à distância de alcance mútuo desses pontos.

Considera-se então um valor limite *(threshold)*, começando alto e diminuindo constantemente. Soltando todas as arestas com peso acima desse limite. Ao soltar as arestas, começaremos a desconectar o gráfico nos componentes conectados. Eventualmente, teremos uma hierarquia de componentes conectados (de completamente conectado a completamente desconectado) em vários níveis de limite.

Na prática, existem n^2 arestas, e o algoritmo deveria ser executado para cada uma destas arestas, então o que se faz é encontrar um conjunto mínimo de arestas, de modo que largar qualquer aresta do conjunto cause uma desconexão dos componentes. Porém é necessário que esse conjunto seja tal que não haja uma borda de peso menor que possa conectar os componentes. O que é resolvido com a árvore de abrangência mínima do gráfico, como podemos ver na Figura 19.

Construímos a árvore uma aresta por vez, sempre adicionando a aresta de menor peso que conecta a árvore atual a um vértice que ainda não está na árvore.

4.5.2.3 Construir a Hierarquia de *Clusters*

Dada a árvore de abrangência mínima, a mesma é convertida numa hierarquia de componentes conectados. Isso é feito classificando as arestas da árvore pela distância (em ordem crescente) e fazendo a iteração, criando um novo *cluster* unido para cada aresta. Porém é necessário identificar os dois *clusters* em que cada borda se unirá, o que pode ser feito por meio de uma estrutura de dados de localização de união.

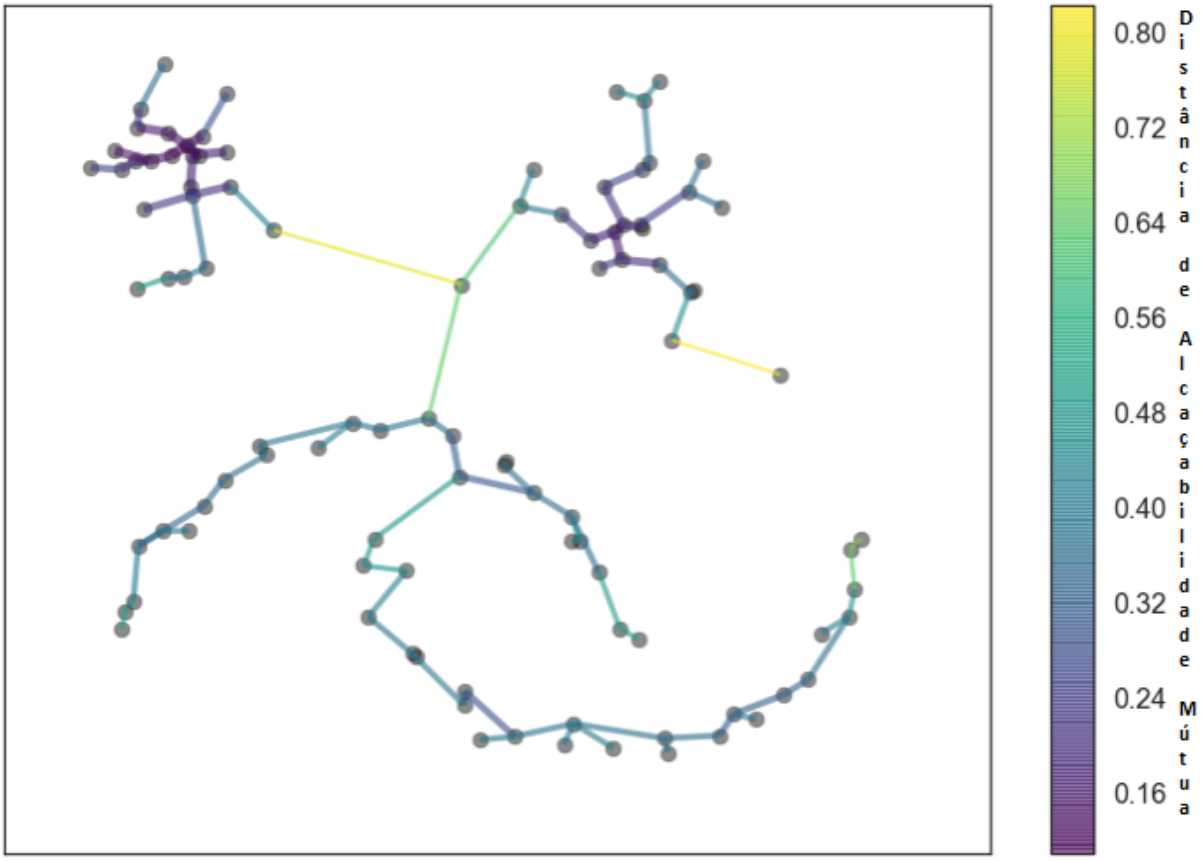

Figura 19 – Essa é a árvore de abrangência mínima para a distância de alcance mútuo, para um valor k de 5, extraído de McInnes, Healy e Astels (2017).

Em *clusters* de ligação única, do inglês *single linkage clustering*, a distância entre dois *clusters* é determinada por um único par de elementos, a saber, os dois elementos (um em cada *cluster*) que estão mais próximos um do outro. O menor desses links que permanece em qualquer etapa causa a fusão dos dois *clusters* cujos elementos estão envolvidos. Em DBSCAN , o ruído é separado dos dados desenhando uma linha horizontal através do diagrama acima e selecionando os *clusters*s pelos quais ela é cortada, declarando quaisquer *clusters* simples no nível de corte como ruído. Esse valor é definido como um parâmetro, ϵ, que muitas vezes não é de fácil escolha.

Porém precisamos lidar com *clusters* de densidade variável e qualquer opção de linha de corte é uma opção de distância de alcance mútuo para cortar e, portanto, um único nível de densidade fixa. Idealmente, queremos ser capazes de cortar a árvore em locais diferentes para selecionar nossos *clusters*.

4.5.2.4 Condensar a árvore de *clusters*

Para extrair os *clusters* temos que condensar a árvore de *clusters* em uma árvore menor, para isso muitos pontos irão juntar-se formando um só *cluster*, o que define isso é o parâmetro $min - cluster\ size$. Qualquer *cluster* que tenha uma quantidade de

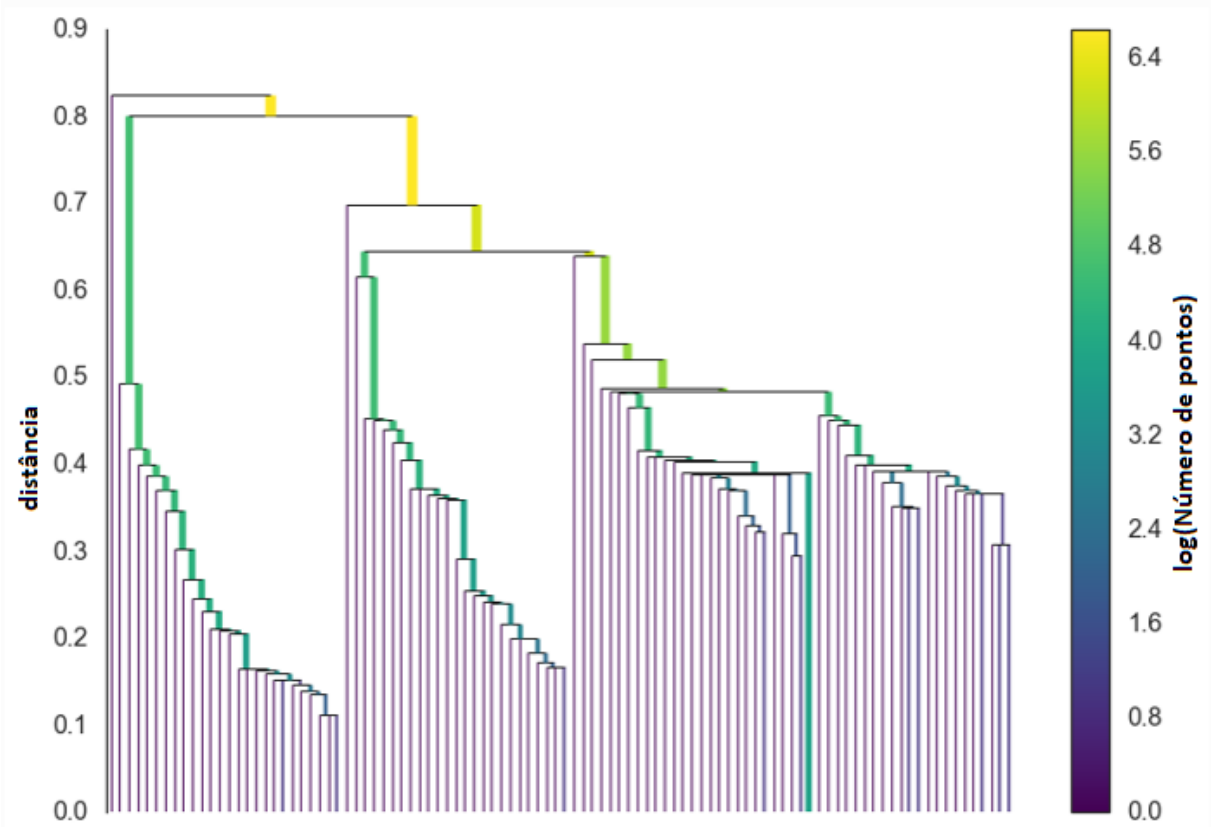

Figura 20 – Dendograma da Árvore de ligação, extraído de McInnes, Healy e Astels (2017).

pontos menor que este parâmetro não será considerado. Isso é feito para cada aresta (galho) da árvore até que todos os *clusters* sejam reduzidos, e terminemos com uma árvore menor que a original e com menos "nós". Isso pode ser visto na Figura 21, que é a representação da Figura 20 com os *clusters* condensados e um parâmetro de $min - cluster\ size$ de 5.

4.5.2.5 Extrair os *clusters*

Então utilizando a árvore condensada os maiores *clusters* são escolhidos, ou seja, os que tem maior persistência, porém após este *cluster* ser escolhido, nenhum "descendente" do mesmo poderá ser escolhido. Para o HDBSCAN é necessário uma medida diferente de distância para considerar a persistência dos *clusters*, com isso é utilizado $\lambda = 1/distancia$.

Para um dado *cluster* são definidos λ_{birth} e λ_{death} como os valores de *lambda* onde o *cluster* se divide e se torna um novo *cluster*. Então, para um dado *cluster*, para cada ponto p neste *cluster* pode-se definir o valor λ_p como o valor de *lambda* no qual aquele ponto "caiu do *cluster*", que é um valor entre λ_{birth} e λ_{death}.

Para cada *cluster* podemos calcular a estabilidade como:

$$\sum p \in cluster(\lambda_p - \lambda_{birth}), \tag{4.7}$$

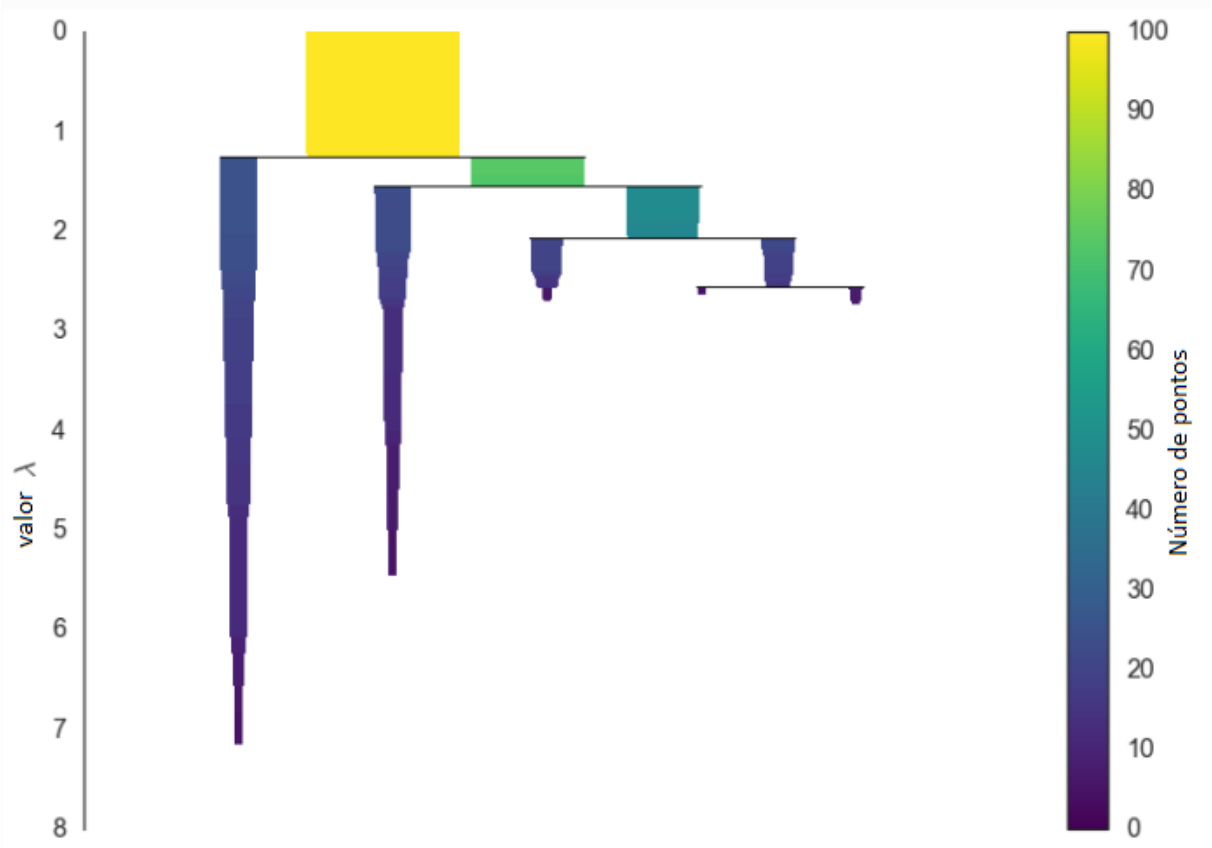

Figura 21 – Árvore condensada, extraído de McInnes, Healy e Astels (2017).

seguindo a ordem de classificação topológica inversa, se a soma das estabilidades dos *clusters* filhos for maior que a estabilidade do *cluster*, definiremos a estabilidade do *cluster* como a soma das estabilidades filhos. Se, por outro lado, a estabilidade do *cluster* for maior que a soma de seus filhos, declaramos que o *cluster* é um *cluster* selecionado e desmarcamos todos os seus descendentes. Quando atingimos o nó raiz, chamamos o conjunto atual de *clusters* selecionados de clusterização simples ou plana. Uma vez que temos os *clusters*, qualquer ponto fora deles é considerado como ruído e é atribuido um rótulo de -1.

4.6 K-Nearest Neighbors

Para podermos falar de *K-Nearest Neighbors*, Altman (1992), ou KNN como é comumente referido, temos que falar também de classificação. Que é um processo relacionado a categorização, o processo no qual idéias e objetos são reconhecidos, diferenciados e entendidos. Em *machine learning*, é um processo de categorizar um determinado conjunto de dados em classes. Ele pode ser executado em dados estruturados ou não estruturados. O processo começa com a previsão da classe de pontos de dados fornecidos. As classes geralmente são chamadas de alvo, rótulo ou categorias.

Existem vários métodos de classificação, e eles se dividem em Paramétricos,

onde a distribuição das características se dá de forma conhecida, como uma distribuição Gaussiana, e Não-Paramétricos onde a distribuição das classes e características se dá de forma aleatória.

Para esse último, um algoritmo largamente utilizado é o KNN que pode ser usado tanto em classificação, quanto para regressão. A entrada consiste nos k exemplos de treinamento mais próximos no espaço de recursos. A saída depende se k-NN é usado para classificação ou regressão.

Na classificação k-NN, a saída é um membro da classe. Um objeto é classificado pelo voto de pluralidade de seus vizinhos, sendo o objeto atribuído a classe mais comum entre os k vizinhos mais próximos (k é um número inteiro positivo, geralmente pequeno). Se $k = 1$, o objeto é simplesmente atribuído à classe do único vizinho mais próximo.

k-NN é um tipo de aprendizado baseado em instância, ou aprendizado lento, em que a função é aproximada apenas localmente e todo o cálculo é adiado até a avaliação da função.

4.6.1 Suposições em KNN

É uma técnica não paramétrica, ou seja, que ela não faz suposições sobre a distribuição de dados subjacente. Em outras palavras, a estrutura do modelo é determinada a partir dos dados. Também é conhecido como algoritmo preguiçoso, ou seja, nenhum aprendizado do modelo é necessário e todo o trabalho acontece no momento em que uma previsão é solicitada. Como tal, o KNN é frequentemente referido como um algoritmo de aprendizado lento.

Aprendizado baseado em instância, ou seja, as instâncias de treinamento brutas são usadas para fazer previsões. Como tal, o KNN é geralmente chamado de aprendizado baseado em instância ou de aprendizado baseado em casos (onde cada instância de treinamento é um caso do domínio do problema).O KNN assume que os dados estão em um espaço de atributos. Mais exatamente, os pontos de dados estão em um espaço métrico. Os dados podem ser escalares ou possivelmente vetores multidimensionais. Como os pontos estão no espaço de atributos, eles têm uma noção de distância, isso não precisa necessariamente ser a distância euclidiana, embora seja a mais usada.

Cada um dos dados de treinamento consiste em um conjunto de vetores e rótulos de classe associado a cada vetor. No caso mais simples, será + ou - (para classes positivas ou negativas). Mas o KNN pode funcionar igualmente bem com um número arbitrário de classes. Também nos é dado um único número "k". Esse número decide quantos vizinhos (onde vizinhos são definidos com base na métrica da distância) influenciam a classificação. Normalmente, esse número é ímpar se o número de classes for 2. Se $k = 1$, o algoritmo é simplesmente chamado de algoritmo vizinho mais próximo.

4.6.2 KNN para Classificação

Como o foco deste trabalho é usar somente o KNN para classificação, na verdade atribuição, não entraremos na parte de regressão do KNN. Uma vez construído o modelo utilizando UMAP+HDBSCAN, ou seja, temos os dados separados em *clusters* e cada *cluster* com seu rótulo, vamos reconstruir o cubo sísmico utilizando KNN, onde pegamos cada amostra e comparamos a distância do mesmo com cada *cluster*, neste caso utilizamos um $k = 5$.

4.6.2.1 Caso 1: k=1 ou Regra do Vizinho mais Próximo

Este é o cenário mais simples. Seja x o ponto a ser rotulado, encontra-se o ponto mais próximo de x, supostamente y. Agora, a regra do vizinho mais próximo pede para atribuir o rótulo de y para x. Que pode ser definido da seguinte forma:

$$C_n^{1nn}(x) = Y(1) \tag{4.8}$$

Suponha que todos os pontos estejam em um plano dimensional D. O número de pontos é razoavelmente grande. Isso significa que a densidade do plano em qualquer ponto é bastante alta. Em outras palavras, dentro de qualquer subespaço, existe um número adequado de pontos. Considere um ponto x no subespaço que também possui muitos vizinhos, seja y o vizinho mais próximo. Se x e y são suficientemente próximos, então podemos assumir que a probabilidade de que x e y pertencem à mesma classe é praticamente a mesma, então, pela teoria da decisão, x e y têm a mesma classe.

À medida que o tamanho do conjunto de dados de treinamento se aproxima do infinito, o classificador vizinho mais próximo garante uma taxa de erro não inferior a duas vezes a taxa de erro de Bayes (a taxa de erro mínima possível, dada a distribuição dos dados).

O limite da classificação é dado por:

$$P^* \leq P \leq P^*(2 - \frac{c}{c-1}P^*), \tag{4.9}$$

onde P^* é a taxa de erro de Bayes, c é o número de classes e P é a taxa de erro do vizinho mais próximo.

4.6.2.2 Caso 2: k = K ou Regra dos K-Vizinhos mais Próximos

É uma extensão direta de 1NN. Basicamente, o que fazemos é tentar encontrar o k vizinho mais próximo e fazer uma votação por maioria. Normalmente k é ímpar quando o número de classes é 2, digamos $k = 5$ e há 3 instâncias de $C1$ e 2 instâncias de $C2$. Nesse

caso, KNN diz que o novo ponto deve ser rotulado como $C1$, pois constitui a maioria. Seguimos um argumento semelhante quando há várias classes.

Uma das extensões simples é não dar 1 voto a todos os vizinhos. Uma coisa muito comum a fazer é KNN ponderado, onde cada ponto tem um peso que normalmente é calculado usando sua distância. Por exemplo, sob ponderação de distância inversa, cada ponto tem um peso igual ao inverso de sua distância até o ponto a ser classificado. Isso significa que os pontos vizinhos têm uma votação mais alta do que os pontos mais distantes.

4.6.2.3 O Classificador de Vizinhos Mais Próximo Ponderado

O classificador de k-vizinhos mais próximos pode ser visto como atribuindo aos k vizinhos mais próximos um peso $\frac{1}{k}$ e a todos os outros peso 0. Isso pode ser generalizado para classificadores de vizinhos mais próximos ponderados. Ou seja, onde ao i-ésimo vizinho mais próximo é atribuído um peso w_{ni}, com $\sum_{i=1}^{n} w_{ni} = 1$.

Considere que C_n^{wnn} denota o classificador mais próximo ponderado com pesos $\{w_{ni}\}_{i=1}^{n}$. Sujeito às condições de regularidade nas distribuições de classe, o risco excessivo tem a seguinte expansão assintótica:

$$R_R(C_n^{wnn}) - R_R(C^{Bayes}) = (B_1 s_n^2 + B_2 t_n^2)\{1 + o(1)\}, \tag{4.10}$$

para as constantes B_1 e B_2 onde:

$$s_n^2 = \sum_{i=1}^{n} w_{ni}^2 \tag{4.11}$$

e

$$t_n = n^{\frac{-2}{d}} \sum_{i=1}^{n} w_{ni}\{i^{1+2/d} - (i-1)^{1+2/d}\}. \tag{4.12}$$

O esquema ótimo de ponderação $\{w_{ni}^*\}_{i=1}^{n}$, que balanceia os termos (4.11) e (4.12), é dado a seguir:

$k^* = \left\lfloor Bn^{\frac{4}{d+4}} \right\rfloor, w_{ni}^* = \frac{1}{k^*}[1 + \frac{d}{2} - \frac{d}{2k^{*2/d}}\{i^{1+2/d} - (i-1)^{1+2/d}\}]$, para $i = 1, 2, ..., k^*$

e

$w_{ni}^* = 0$ para $i = k^* + 1, ..., n$.

4.6.3 Métricas de Distância

Para determinar quais das informações do K no conjunto de dados de treinamento são mais semelhantes a uma nova entrada, uma medida de distância é usada. Para variáveis de entrada de valor real, a medida de distância mais popular é a distância euclidiana. É uma boa medida de distância para usar se as variáveis de entrada principais são semelhantes em tipo (por exemplo, todas as larguras e alturas medidas).

Distância Euclidiana: é calculada como a raiz quadrada da soma das diferenças quadráticas entre um novo ponto x e um ponto existente x_i em todos os atributos de entrada j.

É definida por: $D_e = \sqrt{\sum_{i=1}^{n}(x_j - x_{ij})^2}$, onde n é o numero de dimensões.

Outras medidas populares de distância incluem:

Distância Manhattan: Calcula a distância entre vetores reais usando a soma de sua diferença absoluta. Também chamado de *"City Block Distance"*. É uma boa medida para usar se as variáveis de entrada não forem semelhantes em tipo (como idade, sexo, altura, etc).

É definida por: $D_m = \sum_{i=1}^{n} |x_j - x_{ij}|$

Distância Minkowski: Generalização da distância Euclidiana e Manhattan .

É definida por: $D = \sqrt[p]{\sum_{i=1}^{n} |x_j - x_{ij}|^p}$

Distância de Hamming: Calcula a distância entre vetores binários, mede a semelhança entre duas palavras de mesmo comprimento. A distância de Hamming entre duas palavras do mesmo comprimento é o número de posições nas quais os caracteres correspondentes são diferentes.

Existem muitas outras medidas de distância que podem ser usadas, como a distância de Tanimoto, Jaccard , Mahalanobis e cosseno. A melhor métrica de distância é escolhida com base nas propriedades dos dados.

5 Metodologia, Aplicação e Resultados

Conforme introduzido nos capítulos anteriores, o método que será descrito adiante permite a identificação das sismofácies correspondentes aos sais de baixa (taquidrita, carnalita e silvita) e alta velocidade (anidrita e gipsita) além da halita, com isso melhores modelos de velocidade podem ser gerados e o método pode até mesmo ser usado na área de geomecânica ao ajudar o projetista dos poços a identificar previamente na sísmica as prováveis camadas de sal de baixa velocidade que tem por característica a alta solubilidade, e por consequência sua dissolução no fluido, o que pode levar a perda de circulação e até mesmo a perda do poço causando prejuízos milionários.

O fluxograma da Figura 22 apresenta o encadeamento das etapas que definem os resultados deste projeto.

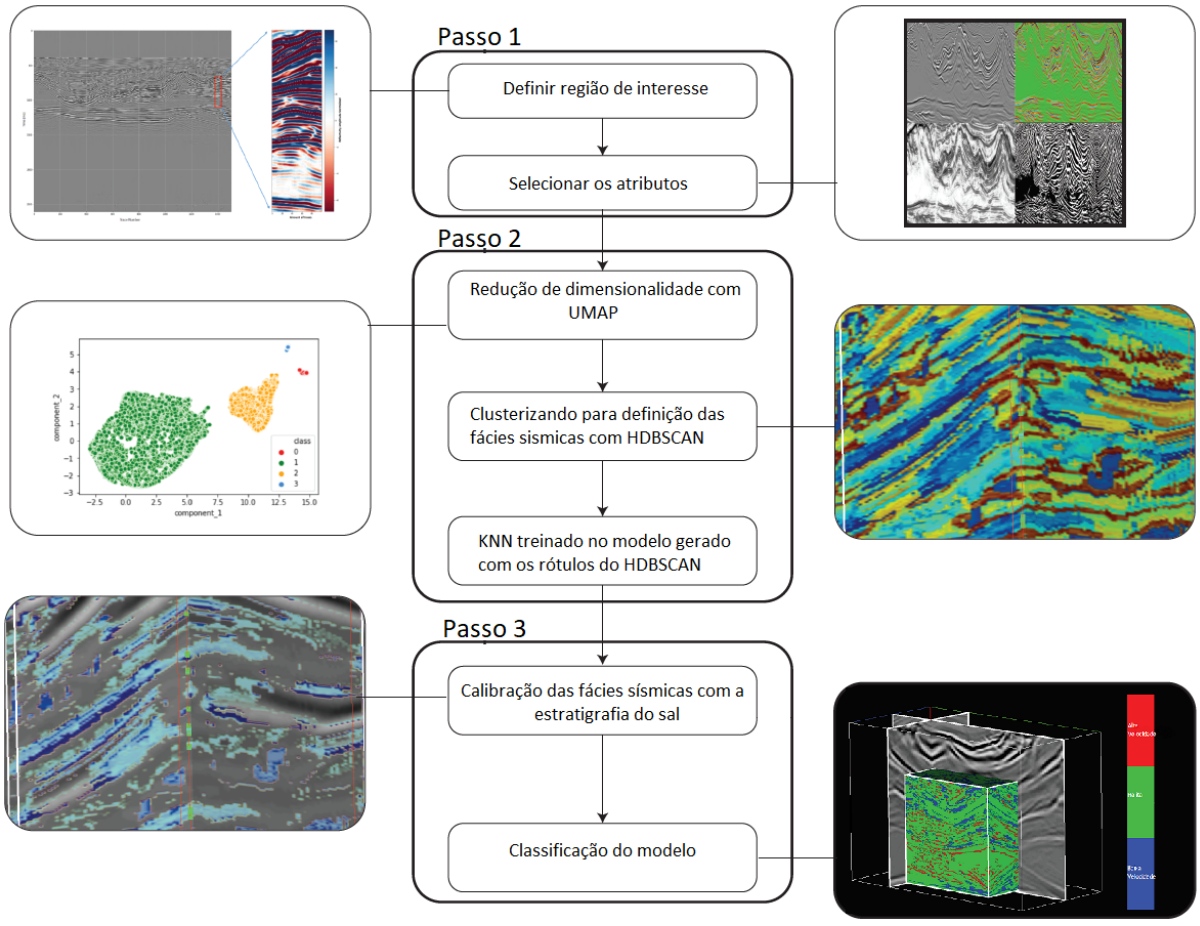

Figura 22 – Fluxograma do método.

5.1 Passo 1: Dados de entrada

O primeiro passo numa interpretação multi-atributos utilizando *machine learning* é a seleção dos dados de entrada como mencionado no capítulo de atributos sísmicos. Com base nos poços disponíveis selecionamos aqueles que tinham perfis elétricos ou perfis de amostras de calha (perfil composto) dentro da camada de sal. Após isso selecionamos uma região do cubo sísmico 3D que cobrisse esses poços e extraímos um cubo sísmico menor, deste cubo menor geramos os atributos de volume, ver Figura 23

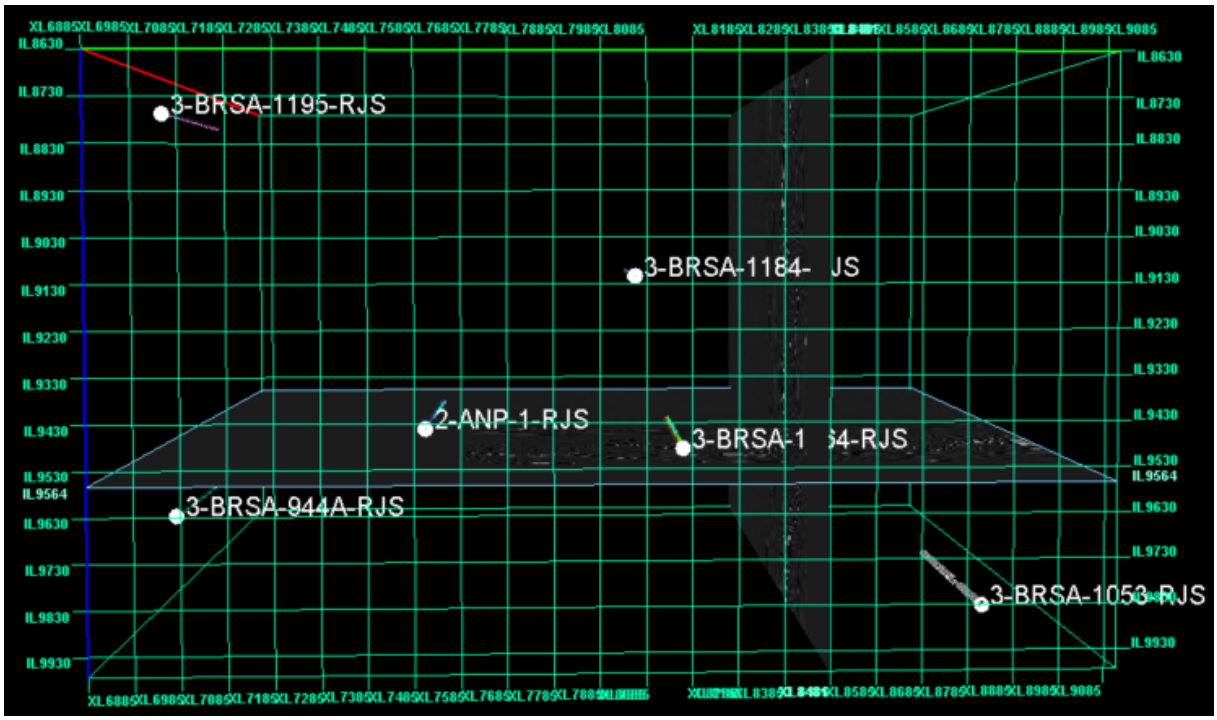

Figura 23 – Com base nos poços disponíveis para o projeto tivemos que buscar os que tinham perfil composto na nossa zona de interesse, neste caso a camada de sal.

Na sequência, mapeamos os horizontes do topo e base do sal, como na Figura 15 e extraímos as *inlines/crosslines* em cima de cada poço para cada atributo, toda essa parte é realizada em um software de interpretação.

Como falado no Capítulo 4, após a seleção dos atributos de entrada, selecionamos a zona de interesse em cada poço, ou seja, somente a camada de sal, e a respectiva seção sobre a linha sísmica, do topo a base da camada de sal e um raio de 50 traços ao redor do poço, como na Figura 16.

Uma vez selecionado a região, fazemos uma janela deslizante sobre a região de interesse, ver Figura 17. Essa é uma prática bastante comum em processamento de imagens, embora estejamos usando o traço sísmico, a matriz de dados em uma *inline/crossline* pode ser tratada da mesma maneira que uma imagem, dependendo da quantidade de dados a ser utilizada pode ser até mesmo necessário paralelização em GPU, Yang et al. (2016).

Porém uma das desvantagens é o aumento da dimensão da matriz de dados após passar pela janela deslizante.

Com isso acabamos com 4 matrizes contendo todo o dado que usaremos como entrada para o próximo passo. Combinamos então essas matrizes como um tensor, que chamaremos de I, e nos referiremos ao valor de um simples atributo como $I|h, w, c|$ onde h é a linha da matriz, w é a coluna e $c \in \{1, 2, 3, 4\}$ representa o atributo.

5.2 Passo 2: Redução de Dimensionalidade e Clusterização

Antes de falarmos de redução de dimensionalidade, temos que falar de alguns passos de pré-condicionamento. Normalmente em um fluxo de trabalho de *machine learning* temos que normalizar ou escalonar os dados de entrada de forma que todos os dados estejam na mesma escala e um atributo não acabe tendo mais ênfase que outro durante os passos do fluxo de trabalho, é uma etapa de pré-condicionamento.

5.2.1 Escalonamento e Normalização

Tem como objetivo alterar os valores das colunas numéricas no conjunto de dados para uma escala comum, sem distorcer as diferenças nos intervalos de valores. Para *machine learning*, nem todos os conjuntos de dados requerem normalização, esse método é necessário somente quando os parâmetros tiverem intervalos muito diferentes. O resultado de um escalonamento pode ser visto na Figura 24.

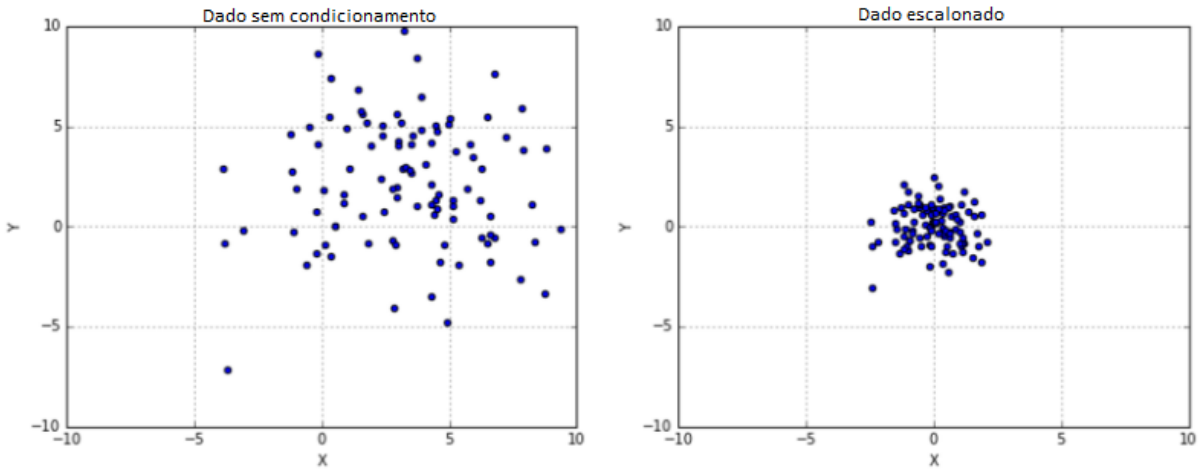

Figura 24 – Exemplo de dado escalonado.

Muitos algoritmos de redução de dimensionalidade como PCA ou SOM necessitam desta normalização. Uma das vantagens do UMAP com relação a estes outros algoritmos é a não necessidade deste pré-condicionamento dos dados, o próprio algoritmo já funciona como um normalizador/regularizador dos dados, o resultado de

UMAP+HDBSCAN com diferentes pré-condicionamentos podemos ver na Figura 25. Aqui o número de clusters não é passado, e o software se encarrega de determinar a quantidade de acordo com o dado, resultando em mais de mil *clusters* neste caso. Cada sismofácie acaba sendo representada por um *cluster*, resultando numa melhor representação das estratificações.

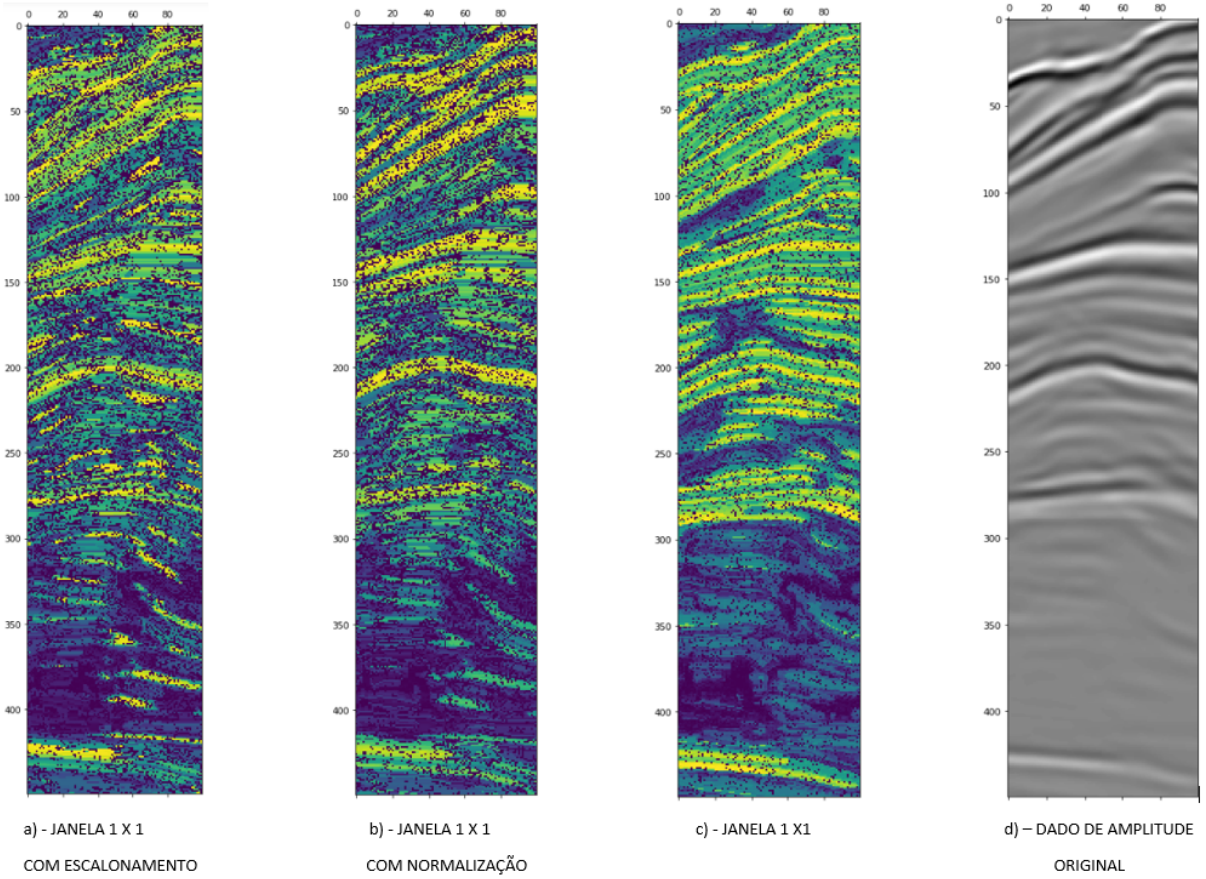

Figura 25 – Comparação da saída de UMAP+HDBSCAN, em (a) usando escalonamento,(b) normalização, (c) nenhum condicionamento de entrada e (d) dado de amplitude original.

5.2.2 Redução de Dimensionalidade

Uma vez concatenados os dados de entrada ficamos com uma dimensão alta e precisamos reduzi-la de forma a poder trabalhar em dimensão baixa, nesse trabalho vamos reduzir para duas dimensões, mas podemos acrescentar mais dimensões caso necessário. Existem várias técnicas de redução de dimensionalidade, listamos algumas na Figura 39.

Neste trabalho testamos quatro algoritmos:

a) Análise de Componentes Principais, do inglês *Principal Component Analysis* (PCA), F.R.S. (1901);

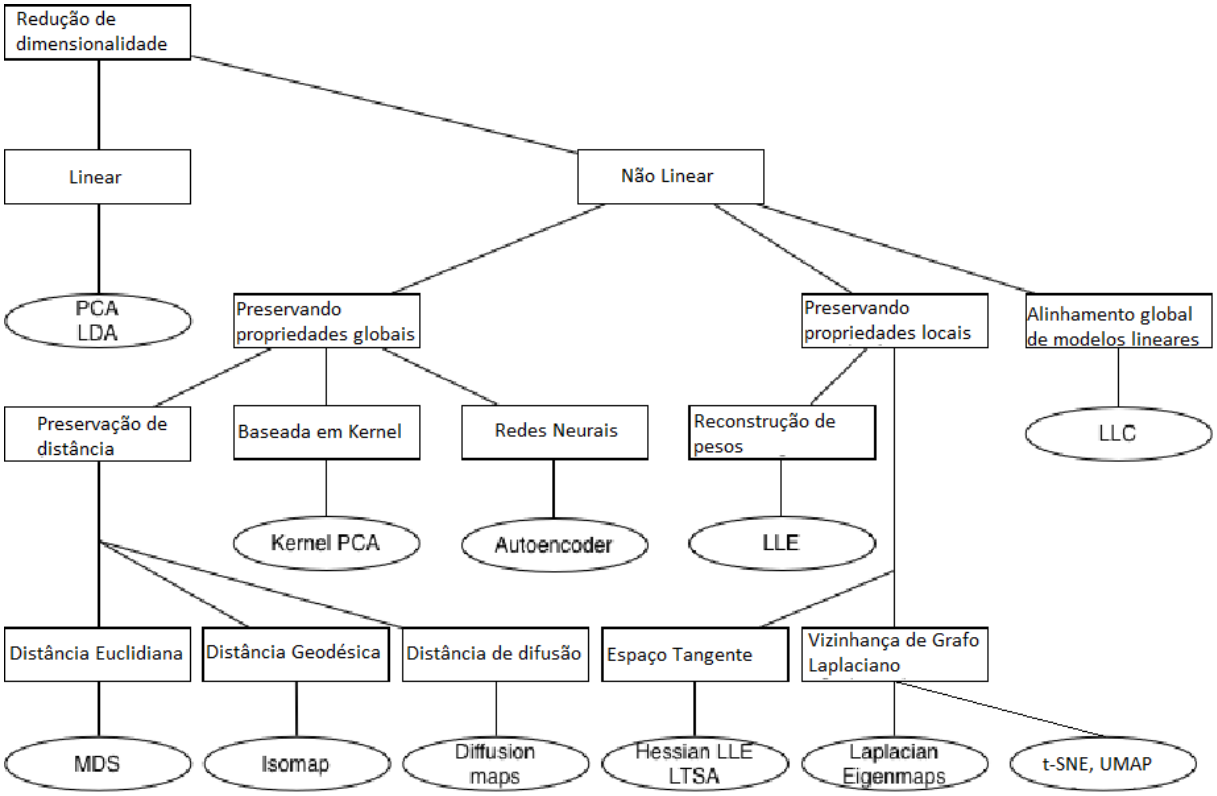

Figura 26 – Técnicas de redução de dimensionalidade. Editado de Maaten, Postma e Herik (2009).

b) Incorporação de Vizinho Estocástico Distribuído t, do inglês *t-Distributed Stochastic Neighbor embedding* (t-SNE), Maaten e Hinton (2008);

c) Mapa Auto Organizável, do inglês *Self Organizing Maps* (SOM), (KOHONEN, 2001);

d) Aproximaçao por Variedade Uniforme e Projeção, do inglês, *Uniform Manifold Approximation and Projection* (UMAP), McInnes, Healy e Melville (2018).

Nas próximas seções explicaremos resumidamente algumas características destes métodos.

5.2.2.1 PCA + *Agglomerative Clustering*

A maioria dos trabalhos de classificação de sismofácies usam a técnica de PCA para redução de dimensionalidade seguido de uma clusterização, normalmente utilizando*k-means*, como no trabalho de Ferreira et al. (2019). Porém devido a natureza não linear do dado sísmico, usar PCA simples não é a melhor escolha neste caso, como demonstrado no trabalho de Maaten, Postma e Herik (2009), onde o mesmo compara PCA com algumas técnicas de redução de dimensão não lineares.

Num primeiro momento foi testado a utilização do PCA no nosso conjunto de dados, na Figura 27 vemos os dados de entrada após seleção e concatenação formando uma nuvem de dados. Após escalonamento fazemos um PCA nesses dados e o resultado pode ser visto na Figura 28.

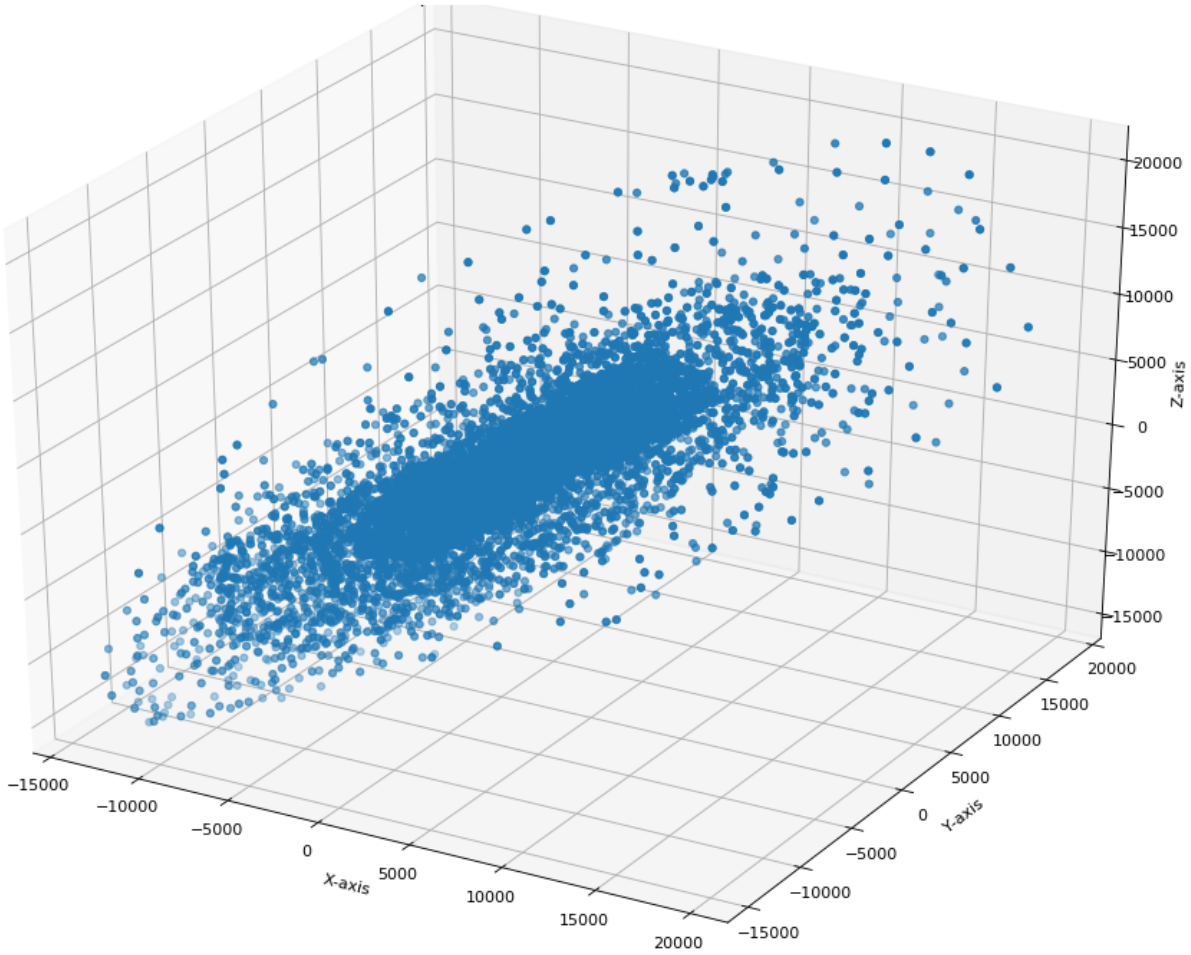

Figura 27 – Nuvem de dados de entrada. Aqui demonstrado as três primeiras componentes.

Após a redução de dimensionalidade, clusterizamos na saída do PCA considerando somente 5 *clusters* (equivalente a quantidade de litologias que estamos buscando), aplicamos um *agglomerative clustering*, que é um tipo de clusterizador hierárquico disponível nas bibliotecas do Scikit-learn, Pedregosa et al. (2011) e o resultado pode ser visto na Figura 29.

5.2.2.2 t-SNE + *Agglomerative Clustering*

Devido aos mal resultados resolvemos tentar algoritmos mais elaborados e que conseguem lidar melhor com a não linearidade dos dados sísmicos, tentamos então uma técnica relativamente nova chamada t-SNE, Maaten e Hinton (2008). Porém a mesma demonstrou ser de difícil manuseio e muito custosa computacionalmente, levando as vezes

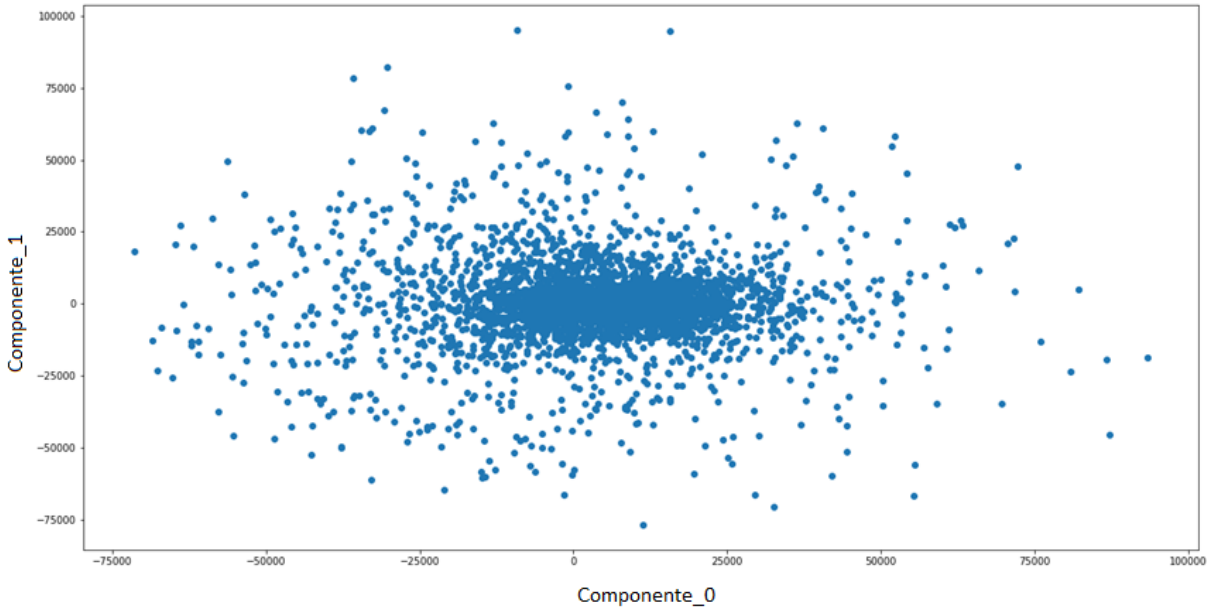

Figura 28 – Dados de entrada após redução de dimensionalidade com PCA.

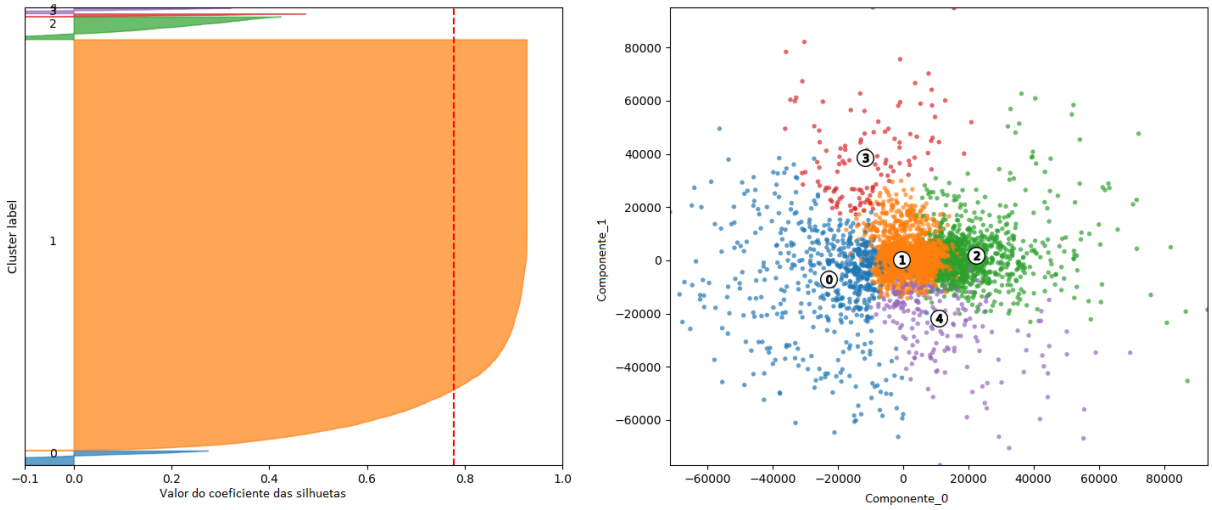

Figura 29 – Clusterização na saída do PCA, usando um clusterizador aglomerativo.

dias de processamento no mesmo conjunto de dados de entrada, além de ter parâmetros que não são tão intuitivos, como a perplexidade. Que é um parâmetro que determina a quantidade de vizinhos próximos a cada ponto, levando a várias iterações do algoritmo até encontrar o valor ótimo do parâmetro. Além do fato de o t-SNE não preservar bem a estrutura global dos dados, o mesmo não é capas de transformar um novo ponto, não servindo para classificação do dado completo. O resultado da redução de dimensionalidade com t-SNE pode ser visto na Figura 30.

Após a redução de dimensionalidade com t-SNE, clusterizamos na saída do mesmo considerando somente 5 *clusters* (equivalente a quantidade de litologias que estamos buscando), e aplicamos novamente um *agglomerative clustering*, e o resultado pode ser

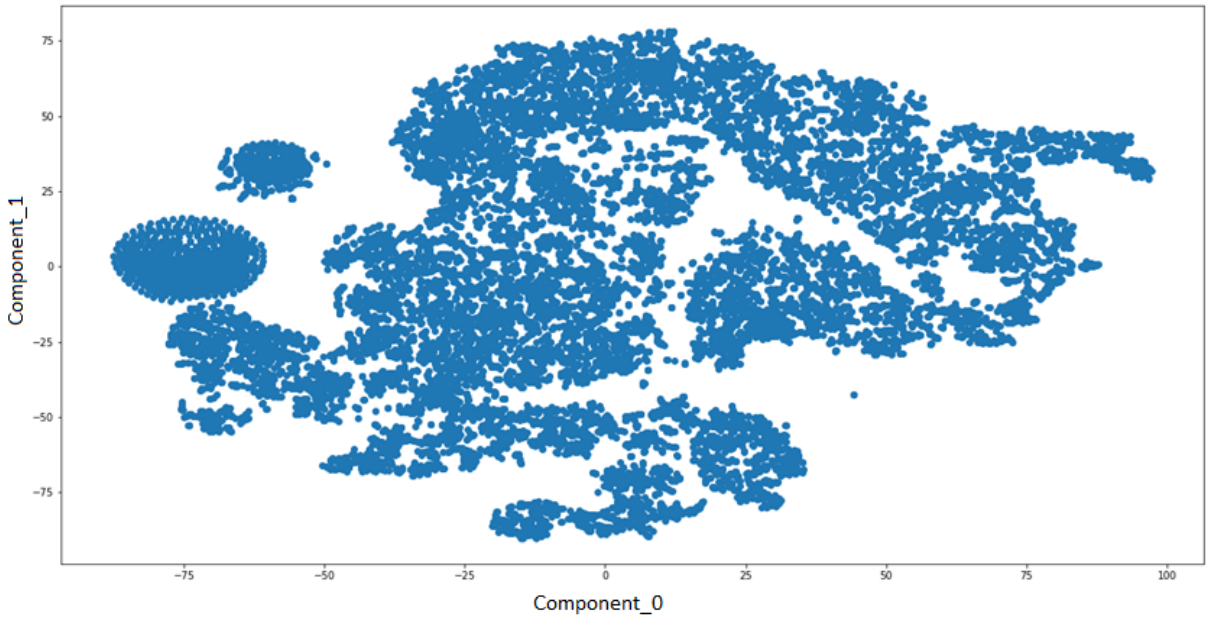

Figura 30 – Dados de entrada após redução de dimensionalidade com t-SNE.

visto na Figura 31.

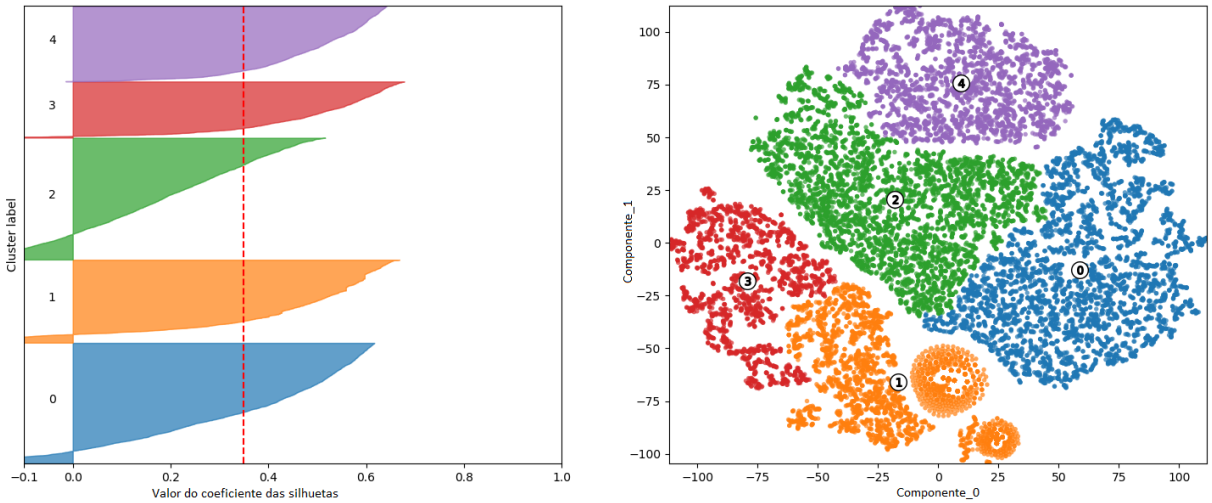

Figura 31 – Clusterização na saída do t-SNE, usando um clusterizador aglomerativo.

5.2.2.3 UMAP + *Agglomerative Clustering*

Como a tentativa de uso do t-SNE para redução de dimensionalidade não foi conclusiva, resolvemos utlizar novos métodos que pudessem lidar melhor com a não linearidade dos dados sísmicos, preservasse melhor a estrutura tanto global quanto local e fosse computacionalmente mais rápido e fácil de utilizar. Com isso encontramos o UMAP, que como já descrito no Capítulo 4 seção 4.4, é um método de redução de dimensionalidade não linear que é adequado para embutir em duas ou três dimensões para visualização como um gráfico de dispersão. É uma técnica relativamente nova, mas muito eficaz para

visualizar *clusters* ou grupos de pontos de dados e suas proximidades relativas, e que faz um bom trabalho em aprender a estrutura local dentro dos dados, mas também tenta preservar as relações entre seus grupos. É rápido, escalonável e pode ser aplicado diretamente a matrizes esparsas, eliminando a necessidade de executar PCA como uma etapa de pré-processamento. Ele suporta uma ampla variedade de medidas de distância, permitindo uma mais fácil exploração dos dados, além de permitir transformar novos dados de forma relativamente rápida usando a árvore do algoritmo, e isso nos permite classifcar o cubo sísmico.

Um estudo comparativo de como UMAP como redutor de dimensionalidade antes da clusterização melhora consideravelmente a performance desses algoritmos pode ser visto no trabalho de Allaoui, Kherfi e Cheriet (2020). O resultado da redução de dimensionalidade dos dados de entrada utilizando UMAP pode ser visto na Figura 32, neste caso utilizamos a distância Euclidiana como métrica de distância.

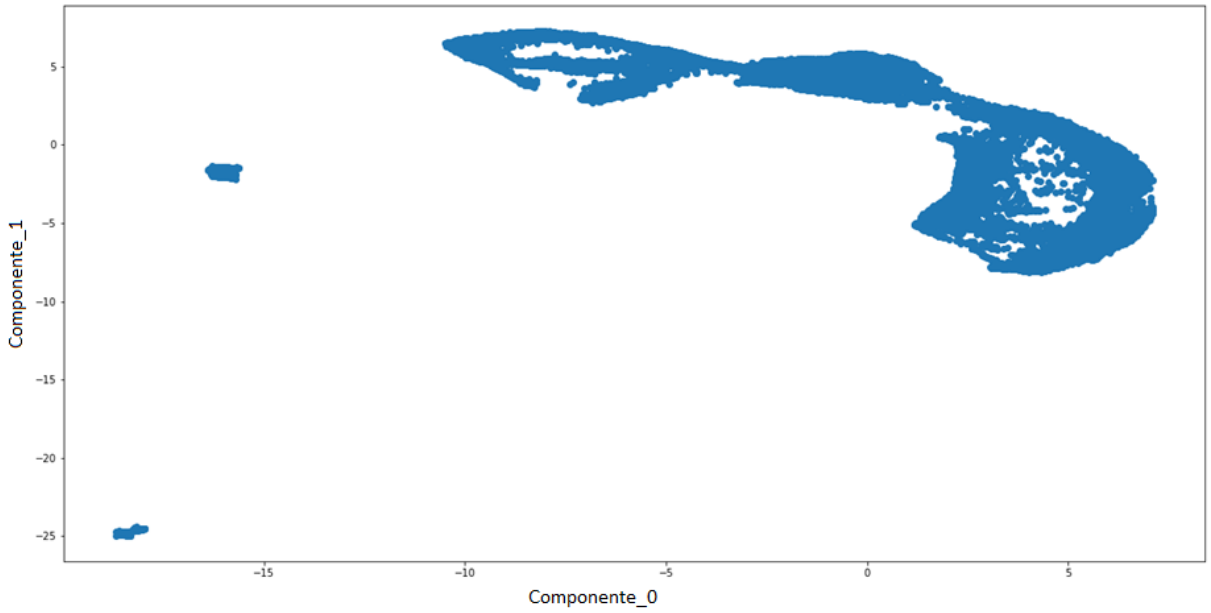

Figura 32 – Dados de entrada após redução de dimensionalidade com UMAP.

Após a redução de dimensionalidade vemos que os grupos são melhor separados, demonstrando uma melhor eficiência do UMAP em lidar com nossos dados, o que é confirmado quando realizamos a clusterização, como pode ser visto na Figura 33.

5.2.2.4 SOM + k-means

Um dos *benchmark* ao qual estamos comparando nosso método é o do SOM, trata-se de um método de redução de dimensionalidade e clusterização que trata bem dados não lineares além de ser computacionalmente rápido. Porém o SOM por si só não consegue identificar sismofácies sozinho, sendo necessário uma etapa posterior de identificação, onde

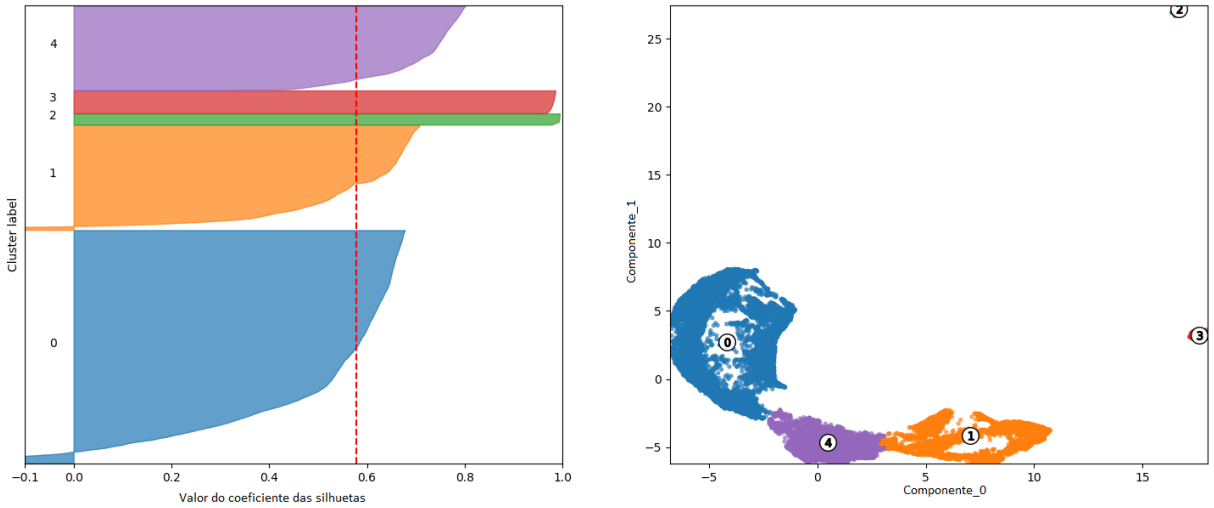

Figura 33 – Clusterização na saída do UMAP, usando um clusterizador aglomerativo.

muitos usam o k-means, um exemplo é o trabalho de identificação de fácies carbonáticas feito por Bronizeski (2018).

Para fins de comparação testamos o SOM nos nossos dados de entrada, o número de neurons foi definido de acordo com a Equação 5.1

$$M = 5 \times \sqrt{N}, \tag{5.1}$$

onde M é o numero de neurons e N é o número de observações ou dados de entrada. Com isso ficamos com uma matriz de neurons de 32 x 32, o mapa produzido pode ser visto na Figura 34.

Após isso utilizamos a mesma técnica de aplicar *k-means* na saída do SOM, porém precisamos primeiro definir qual o número de *clusters* correto, para isso utilizamos o chamado "Método do Cotovelo", Thorndike (1953), onde o ponto de inflexão da curva indica a quantidade que *clusters* que melhor representa seu dado. O resultado pode ser visto na Figura 35.

Com a definição da quantidade de *clusters* a utilizar, neste caso 5, que é a mesma quantidade de litologias que estamos buscando, fizemos então o *k-means*, o resultado pode ser visto na Figura 36.

Com isso podemos ver que os *clusters* resultantes lembram muito a saída produzida pelo PCA. Existem algumas desvantagens neste método, um deles é a não preservação da estrutura dos dados, outra é a necessidade de normalização e uso de PCA como etapa de pré-condicionamento dos dados, além de que o uso do *k-means* como clusterizador não é a melhor escolha. De acordo com Pedregosa et al. (2011), apesar de ser de fácil compreensão e sempre convergir, esta convergência nem sempre é para o mínimo global. Com isso pode haver uma classificação equivocada dos grupos, principalmente onde

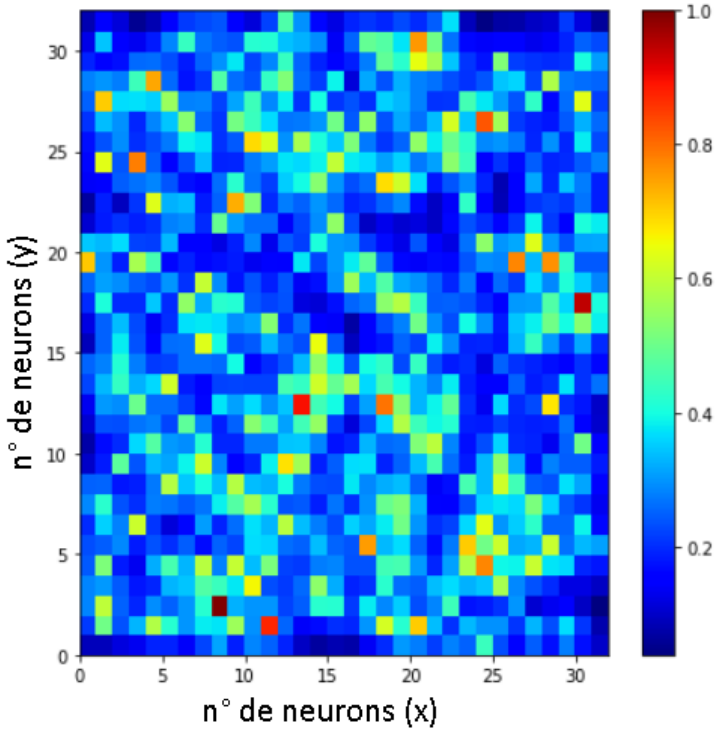

Figura 34 – Mapa bi-dimensional produzido pelo SOM nos dados utilizados.

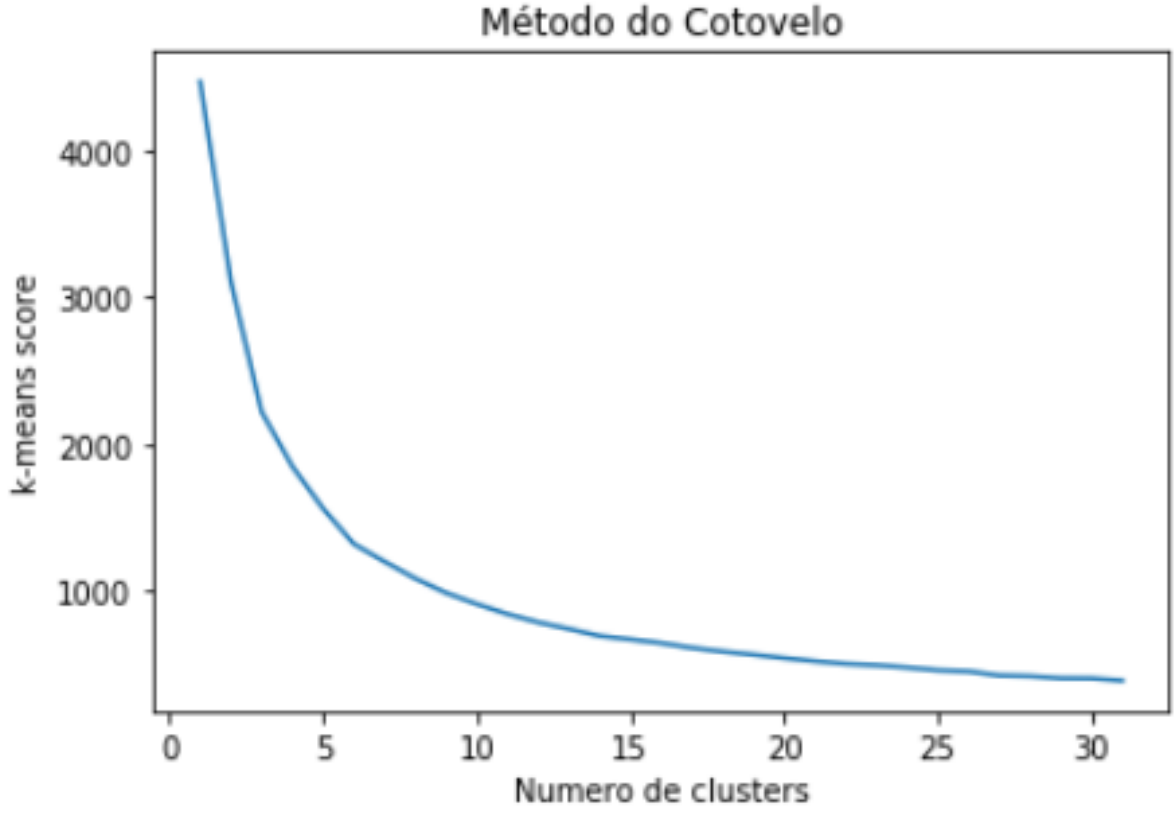

Figura 35 – Método do cotovelo.

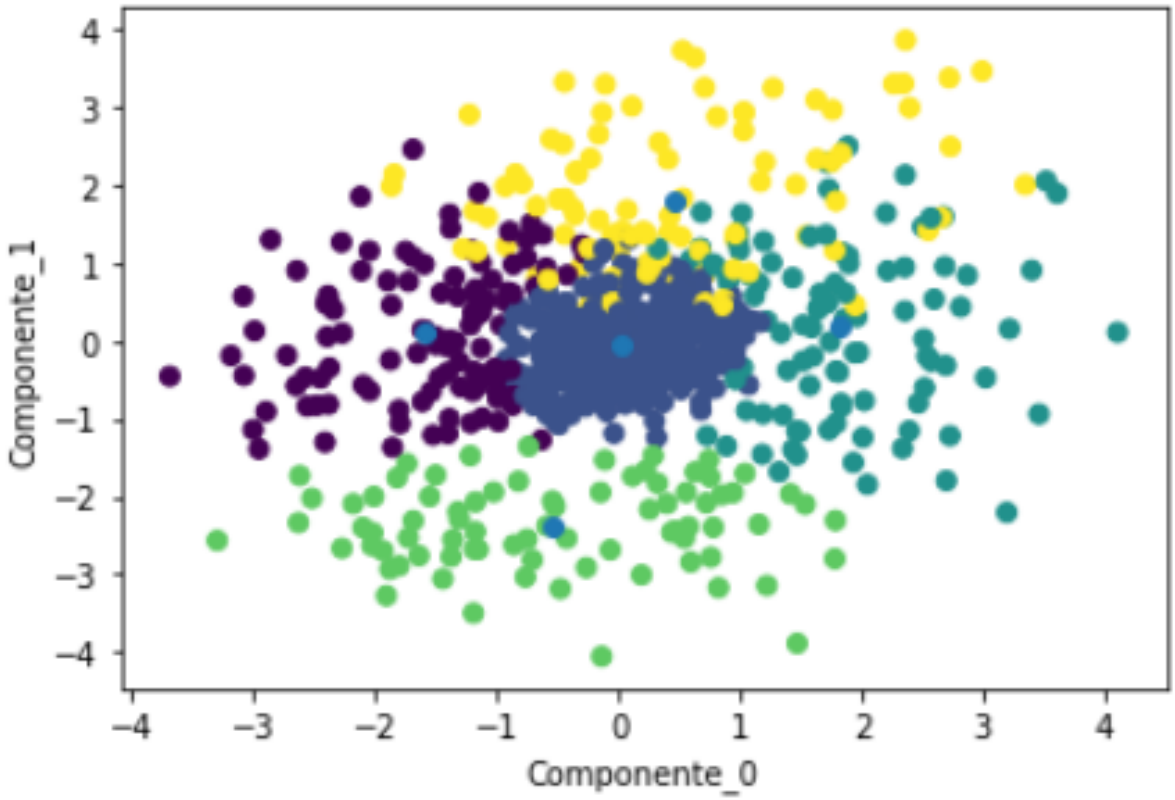

Figura 36 – Clusterização na saída do SOM usando *k-means*.

os conjuntos de dados tem formato alongado, ou formas irregulares, como é o nosso caso.

Na Figura 37, fazemos uma comparação do método de clusterização proposto (HDBSCAN) para separação dos *clusters* no dado de saída do UMAP quando comparado com o método *k-means* no mesmo dado, aqui passamos a quantidade de *clusters* esperada.

Podemos ver que mesmo quando a estrutura é preservada utilizando UMAP, se o método de clusterização não for apropriado, boa parte da informação é interpretada erroneamente.

5.2.3 UMAP + HDBSCAN

Uma vez reduzido a dimensão dos dados com UMAP, usamos HDBSCAN para definir as fácies sísmicas, fazemos isso desta vez sem passar a quantidade de *clusters* esperada, mas deixando o dado ”falar por si mesmo”. O UMAP foi criado para ser usado principalmente como um método de redução dimensional, mas sua formulação também tem um efeito de regularização. Neste trabalho estamos mais interessados no efeito regulatório. O UMAP pode ser usado para transformar grupos de pontos que são muito densos em grupos com densidade de pontos mais uniforme, a distância mínima entre os pontos no *embedding* pode ser controlada pelo hiper-parâmetro de distância mínima. Isso tem um efeito positivo porque facilita o trabalho de algoritmos de clusterização que trabalham

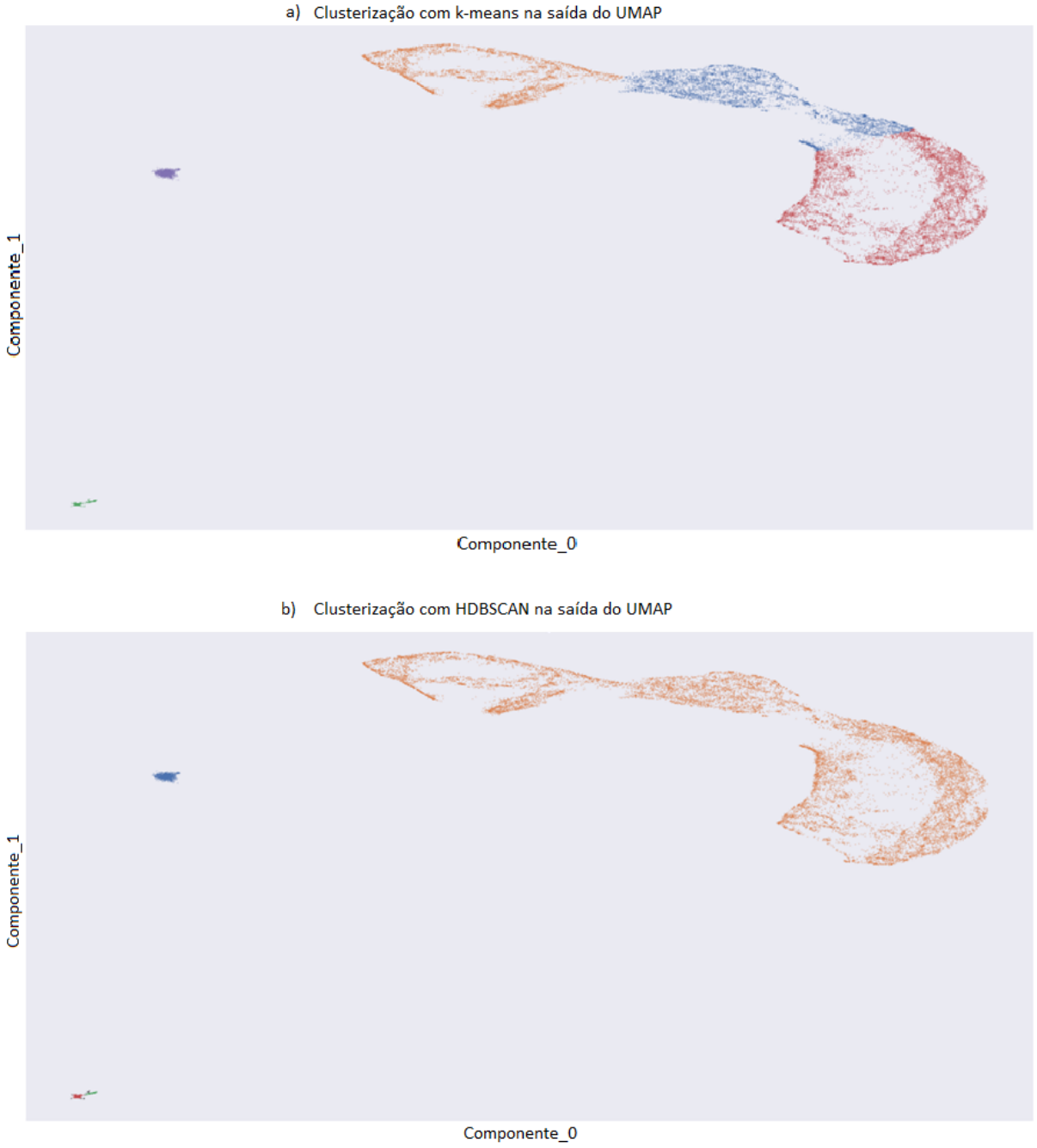

Figura 37 – Comparação entre: (a) usar *k-means* como clusterizador e (b) usar HDBSCAN como clusterizador.

expandindo a seleção de amostra, como DBSCAN e HDBSCAN. Na Figura 38 fizemos uma comparação entre clusterizar os dados com UMAP + K-means (a), usar somente o HDBSCAN sem UMAP (b) e usar UMAP + HDBSCAN (c). Comparando o resultado da Figura 38 (b) com a Figura 38 (c) podemos ver o efeito dessa regularização, quando empregamos o UMAP as fácies sísmicas são muito mais definidas. Os pontos pretos na imagem são os pontos os quais o algoritmo não conseguiu definir como pertencente a nenhum *cluster* e portanto classifica como ruído, que é uma característica do HDBSCAN.

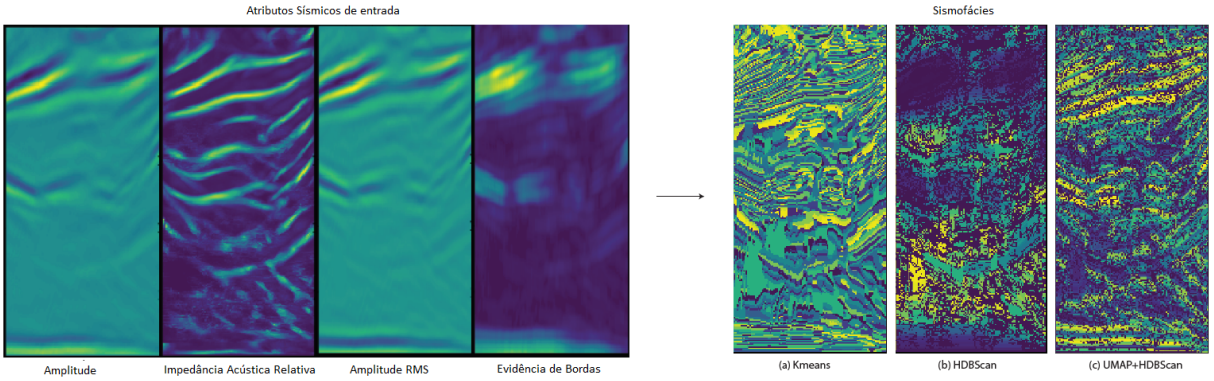

Figura 38 – Atributos sísmicos de entrada e sismofácies encontradas por diferentes métodos. (a) UMAP+K-means, (b) somente HDBSCAN, (c) UMAP+HDBSCAN.

Integrar UMAP com um algoritmo de clusterização hierárquico fornece a melhor evidência da existência de estruturas de dados local e global. Aqui usamos um algoritmo que faz essa clusterização sobre valores variados de ϵ (epsilon) e integra o resultado para encontrar um agrupamento que dê a melhor estabilidade sobre este ϵ, que é o parâmetro que especifica quão próximos os pontos devem estar de cada outro para ser considerado parte de um *cluster*. Isso permite ao HDBSCAN encontrar *clusters* de densidades variáveis (ao contrário do DBSCAN), e ser mais robusto para a seleção de parâmetros.

Aqui utilizamos como parâmetro uma quantidade mínima de amostras igual a 5, tamanho mínimo de *cluster* igual a 50 e uma métrica de distância de Mahalanobis. Essa métrica foi escolhida devido a sua formulação em que mede a distância entre um ponto e uma distribuição, como um *cluster* por exemplo, diferentemente da distância Euclidiana que mede a distância entre dois pontos. No caso em que a distribuição dos grupos não é esférica mas altamente irregular e de tamanhos diferentes, como é o nosso caso, é a métrica mais indicada pois remove informação redundante de variáveis que são altamente correlacionadas, isso pode ser visto no trabalho de Xiang, Nie e Zhang (2008). Além de ser invariável em escala, e dessa forma mais robusta a possível saída fora de escala da transformação UMAP. Isso faz com que esta métrica seja extremamente útil, tendo aplicações na detecção de anomalias multivariadas, classificação em conjuntos de dados altamente desbalanceados e classificação de uma só classe.

Neste trabalho o UMAP foi usado para criar um *embed space* com duas componentes, ver Figura 39, a partir da entrada dos quatro atributos. Essas duas novas componentes serão a entrada para o algoritmo de clusterização. O UMAP tem dois principais hiper-parâmetros: o **número de vizinhos** que faz o balanço entre estrutura global versus local nos dados e a **distância mínima** que controla quão juntos os pontos são permitidos a ficarem. Neste caso usamos o número de vizinhos como 35 e distância mínima 0.01, selecionamos esses valores utilizando uma busca de *grid* e inspeção visual da

distribuição de fácies. As fácies sísmicas obtidas com esta metodologia são apresentadas na Figura 38 (c).

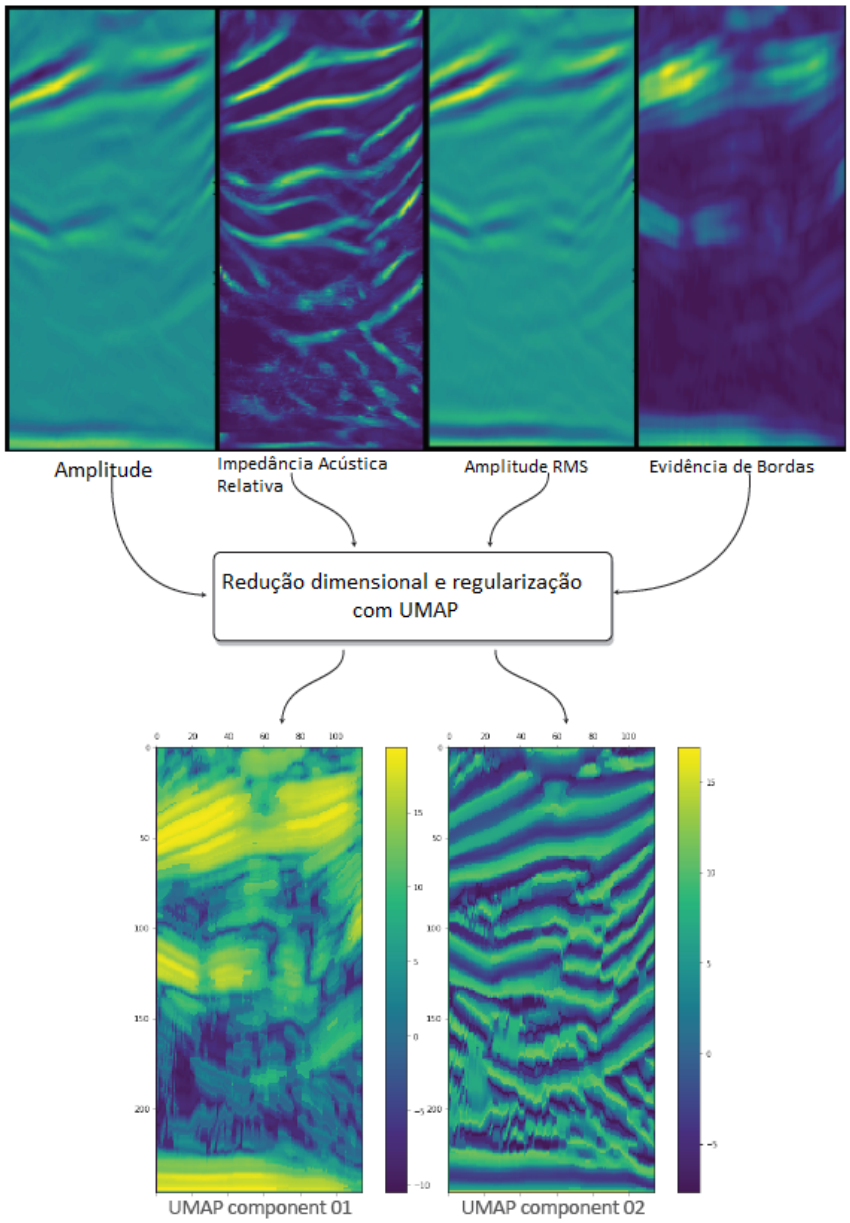

Figura 39 – Atributos sísmicos de entrada acima e saída em duas componentes após transformação UMAP abaixo.

5.2.3.1 HDBSCAN Modificado

Um problema com HDBSCAN é que na sua formulação ele tenta separar os *cluster* do dado de fundo (*background*), ou seja, os dados que não foram classificados como pertencente a nenhum *cluster* são considerados como ruído. Esse efeito pode ser visto na Figura 40 (a), onde há muitos pontos pretos. Para resolver esse problema modificamos o algoritmo HDBSCAN original, e agora para cada ponto classificado como ruído busca-

mos pelo ponto classificado mais próximo e transferimos seu rótulo. O resultado desta modificação pode ser visto na Figura 40 (b).

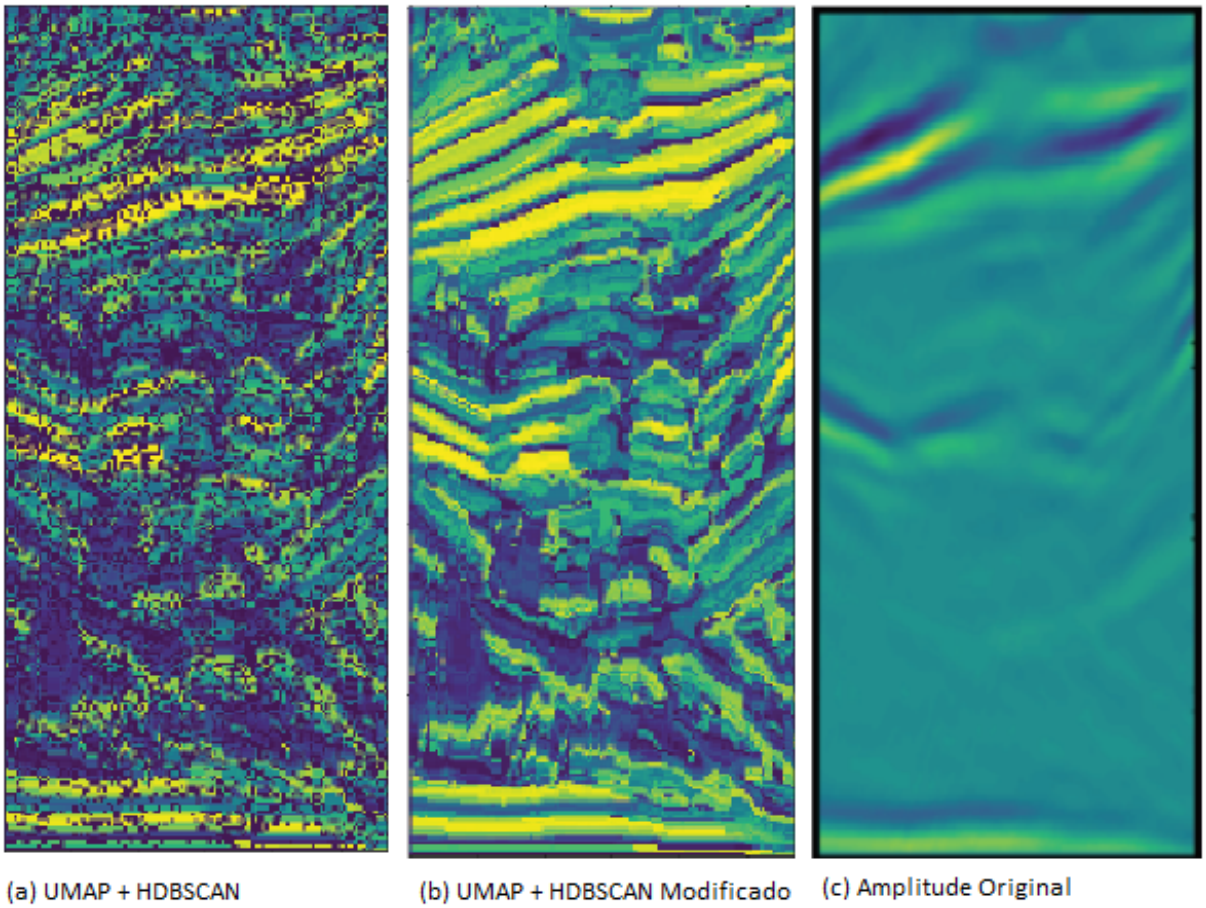

Figura 40 – Comparação entre usar o HDBSCAN com e sem modificação. (a) UMAP+HDBSCAN original; (b) UMAP+HDBSCAN modificado; (c) Amplitude Original.

5.2.4 Propagar os Rótulos

Após a aplicação do método de clusterização proposto (UMAP + HDBSCAN Modificado), terminamos com um mapa de sismofácies da região de interesse em cima do poço, porém queremos expandir essa classificação de sismofácies para o cubo sísmico ao redor. Para fazer isso usamos o modelo diretamente, transformando novas amostras com UMAP e daí podemos proceder de duas maneiras:

a) Treinar um KNN com os rótulos encontrados pelo HDBSCAN e usar esse KNN para predizer os rótulos das novas amostras, fazendo desta maneira, o método fica mais rápido, porém devido a usar uma etapa a mais de treinamento acaba tornando-se mais complexo.

b) Usar a árvore condensada criada pelo método do HDBSCAN para predizer os rótulos das novas amostras, o que já é feito através de um KNN modificado interno ao próprio HDBSCAN.

Independente do método escolhido terminamos com um cubo sísmico com as fácies rotuladas como na Figura 41 (c).

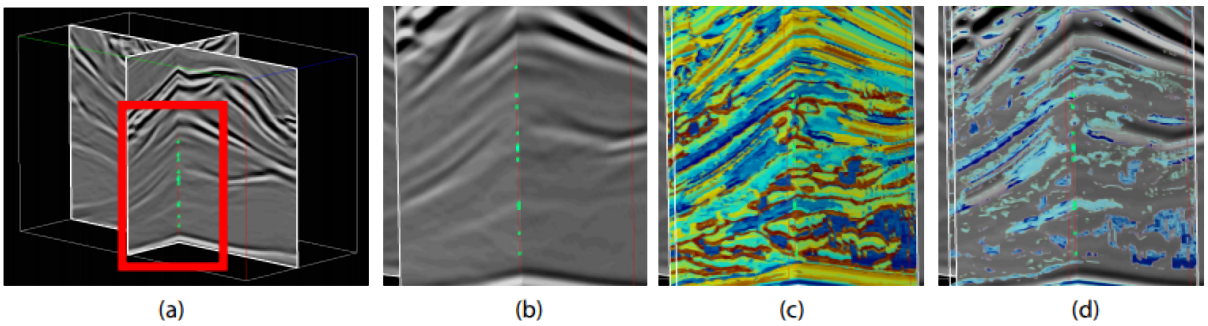

Figura 41 – (a) Cubo sísmico mostrando a *inline e crossline* sobre o poço; (b) Cubo sísmico em detalhe e poço, pontos verdes indicam LVS no poço; (c) fácies sísmicas propagadas; (d) fácies de LVS selecionadas.

5.3 Passo 3: Calibração das Fácies e Geração do Modelo

Neste passo mapeamos as fácies sísmicas do Passo 2 com a estratigrafia do sal. Porém este é um problema mal posto, de informações incompletas devido ao fato da sísmica não ser perfeita e conter várias fontes de incerteza e da baixa resolução sísmica.

Gao (2007) usou o método de calibrar os clusters utilizando o controle do poço, dando origem ao que é chamado de supervisão a posteriori, onde ele usa poços junto com seu conhecimento sobre o ambiente deposicional em que estava trabalhando para calibrar os clusters encontrados. Baseado nesses trabalhos usamos o mesmo processo, onde agora os mesmos poços que estamos usando como referência servem para calibrar as fácies sísmicas, para isso um passo de processamento tem que ser realizado antes, que é a amarração sísmica-poço para colocar as informações no mesmo domínio e calibradas em profundidade.

Vejamos por exemplo o caso da Figura 42 (a), o dado de poço mostra que há sais de baixa velocidade (pontos verdes), porém a sísmica não mostra uma reflexão sobre o poço, mas para a direita e esquerda as reflexões estão presentes. Com isso existem muitas possíveis possibilidades de mapeamento da sísmica para as camadas de sal, neste sentido, para criar esse mapeamento temos que usar a interpretação, e para isso usamos duas fontes de informação: perfis de poços e conhecimento sobre a distribuição do sal nesta Bacia.

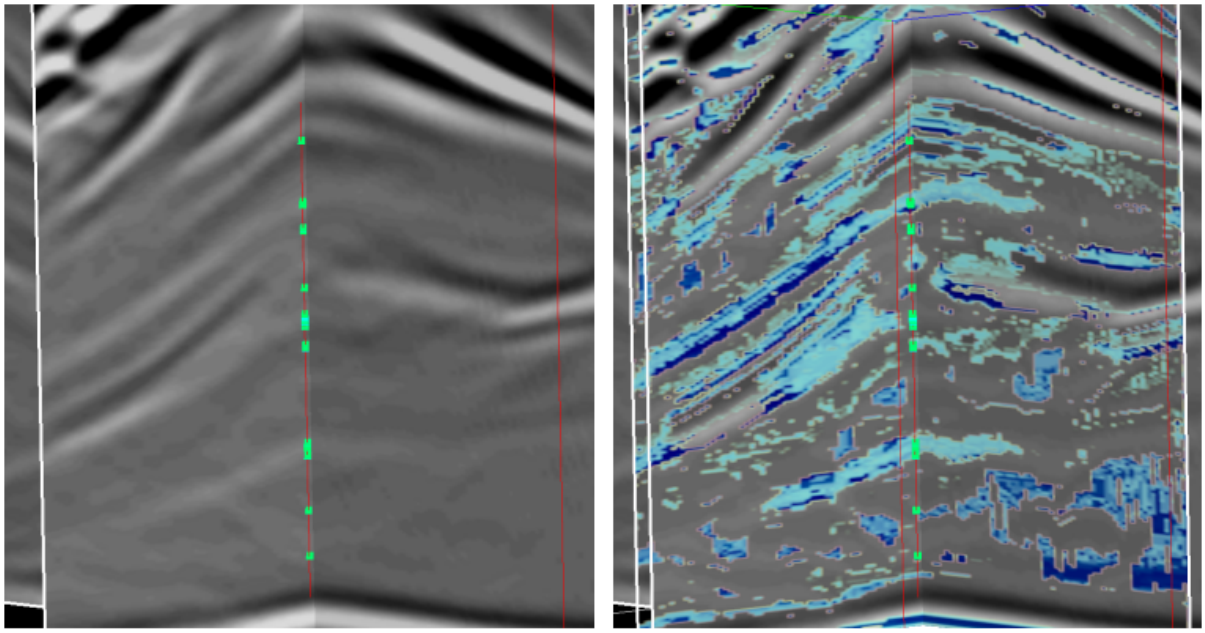

Figura 42 – (a) Cubo sísmico em detalhe e poço, pontos verdes indicam LVS no poço; (b) fácies de LVS selecionadas.

Neste trabalho queremos mapear as fácies sísmicas aos três diferentes grupos estratigráficos de sal: LVS, Halita (*background*) e HVS. Como apresentado na Tabela 3.

GRUPO	COMPOSIÇÃO	MINERAL	DENSIDADE (g/cm3)	VELOCIDADE INTERVALAR (m/s)
LVS	Taquidrita	CaMg2Cl6.12(H20)	1,57	3.300
	Carnalita	KMgCl3.6(H20)	1,66	3.910
	Silvita	KCl	1,86	3.910
BACKGROUND	Halita	NaCl	2,10	4.550
HVS	Gypsita	CaSO4.2(H20)	2,35	5.810
	Anidrita	CuSO4	2,98	6.100

Tabela 3 – Grupos minerais e suas respectivas propriedades, valores médios adaptados por Maul, Santos e Silva (2018) considerando 182 poços na Bacia de Santos.

Para criar este mapeamento salvamos este cubo sísmico classificado com as fácies sísmicas no formato SEGY e o carregamos em um software de interpretação, neste caso utilizamos Petrel 2016, onde podemos visualizar os dados de poço sobre o dado sísmico, como na Figura 41. Os pontos verdes sobre o poço indicam as regiões com sais de baixa velocidade (LVS), então selecionamos as fácies sísmicas associadas com essas regiões. O resultado desta seleção pode ser visto na Figura 41 (d) e com mais detalhes na Figura 42 (b).

A seleção das fácies é muito facilitada pelo fato de empregarmos HDBSCAN, geralmente algoritmos de clusterização geram rótulos que não tem uma ordem específica, como pode ser visto na Figura 38 (a). O HDBSCAN diferentemente gera rótulos para as fácies sísmicas de forma ordenada, os rótulos são números inteiros, o que significa que

em geral, fácies com características similares tem rótulos mais próximos, essa estrutura pode ser vista na Figura 40 (a) e (b). O mesmo processo é usado para selecionar as fácies associadas com os sais de alta velocidade (HVS) e as fácies remanescentes são associadas a Halita, o resultado final é um mapa 3D com a classificação do sal, isso pode ser visto na Figura 43.

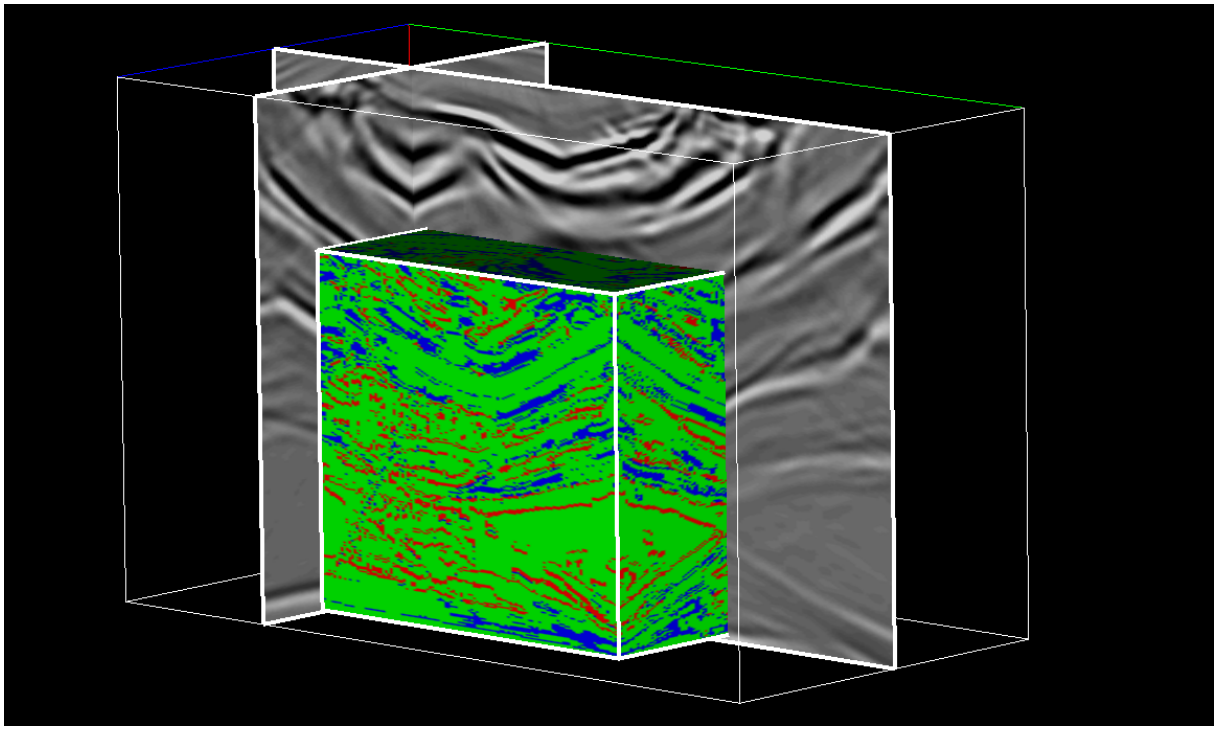

Figura 43 – Volume sísmico 3D classificado e carregado em um software de interpretação. Em azul LVS, em verde a Halita e em vermelho HVS, sísmica original ao fundo.

Uma vez que interpretamos as sismofácies usando os poços como referência e sabemos quais rótulos pertencem a qual classe podemos então atribuirr valores de acordo com a litologia. Nesse caso usamos os valores médios do sal de acordo com a Tabela 3, onde são apresentados o agrupamento mineral como proposto por Maul, Santos e Silva (2018), indicando a fórmula quimica de cada mineral bem como sua propriedade acústica. Outros valores também podem ser atribuídos tais como valores de impedância por exemplo. Uma vez que temos o modelo de velocidade gerado finalmente exportamos o cubo sísmico ou simplesmente a *inline*/*crossline* em formato ”bin” ou SEGY, que pode então ser utilizado para fins de migração ou geomecânica por exemplo.

Na Figura 44, temos um exemplo de uma *inline* classificada e a mesma inline calibrada após interpretação.

Na Figura 45 (a) temos um exemplo de outra *inline* sobre outro poço, abaixo em (b) a mesma *inline* já com os valores de velocidade considerando os três grupos de sal, neste caso fizemos a classificação para toda a *inline* desconsiderando as seções pré e

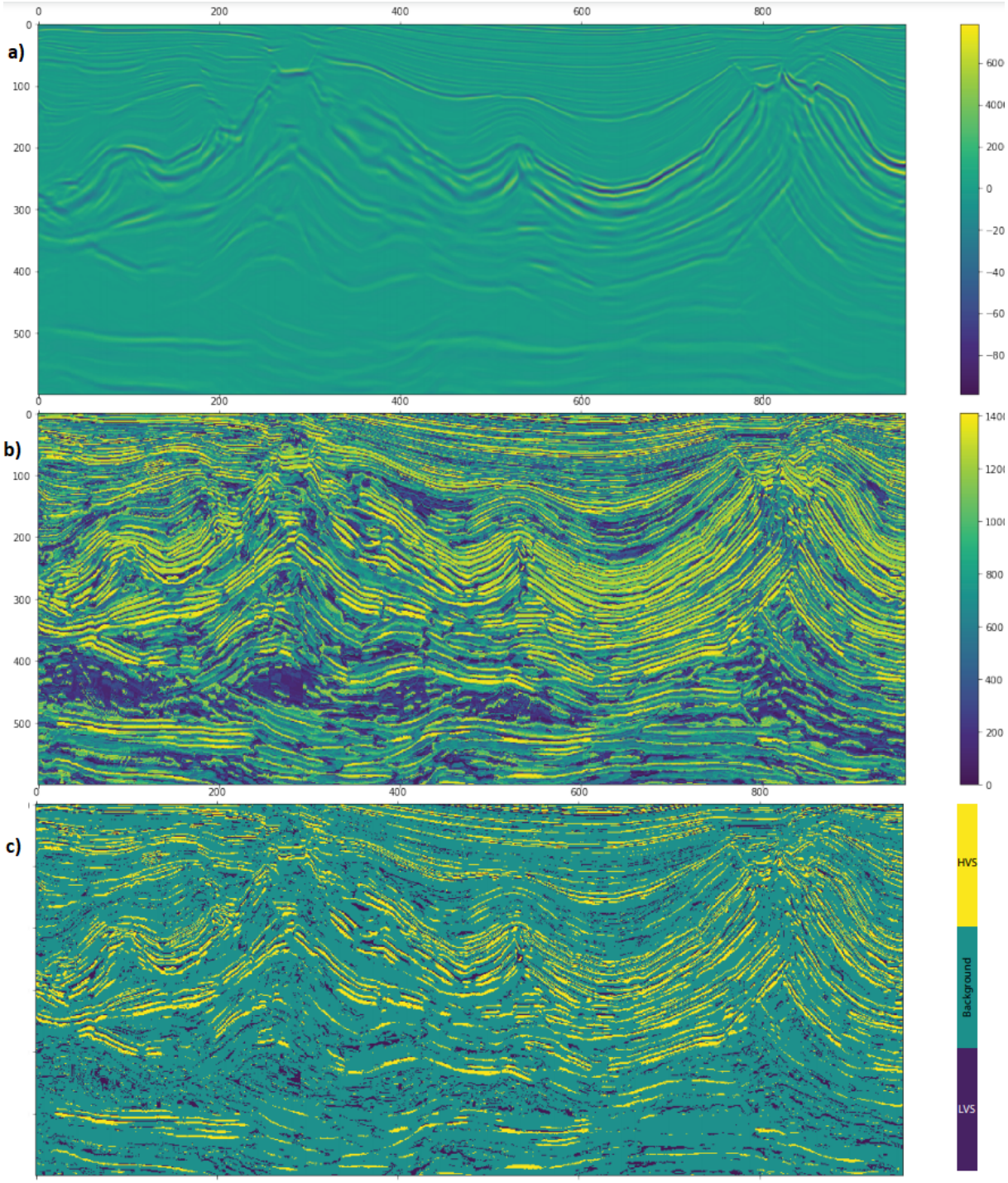

Figura 44 – (a) Amplitude sísmica original; (b) Fácies sísmicas detectadas pelo algoritmo proposto UMAP + HDBSCAN Modificado; (c) modelo com a litologia do sal calibrada.

pós sal. A mesma metodologia pode ser usada em outros ambientes somente mudando o modelo, onde cada litologia é uma classe e os atributos sísmicos de entrada devem ser os apropriados para cada ambiente, o recorte dos dados também deve ser o apropriado para a seção a ser estudada.

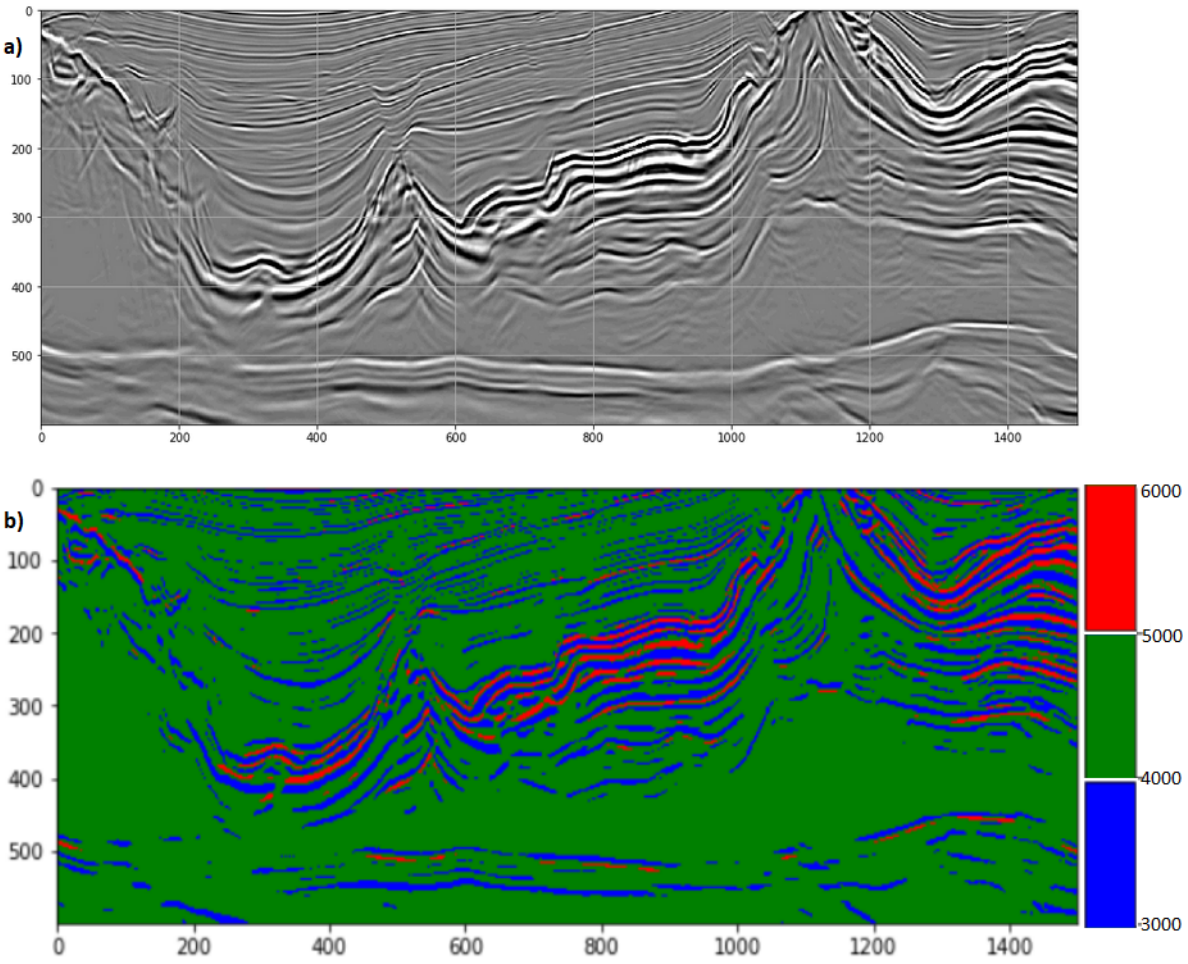

Figura 45 – Acima em (a) amplitude sísmica original; abaixo em (b) modelo com as velocidades do sal.

6 Conclusões

Do que pudemos aprender neste trabalho é que é extremamente difícil mapear os sais de baixa velocidade na sísmica devido a várias razões, uma delas é que a estratificação pode ser de somente alguns metros ficando muitas vezes abaixo da resolução sísmica, e portanto não gerando uma reflexão visível. Mais ainda é diferenciar a Carnalita da Taquidrita. Muitos métodos vem sendo utilizados nesse sentido, como inversão, migração, criação de modelos, classificação Bayesiana e classificação de sismofácies. Neste último decidimos tentar por métodos não convencionais, como já explicado anteriormente a natureza não linear do traço sísmico favorece o uso de redes neurais, e portanto decidimos pelo uso de *machine learning* para identificação dessas estratificações.

Uma vez selecionados os dados de entrada (atributos) e as referências(poços) que usaríamos como balizamento (rótulos), passamos então a testar métodos de classificação de sismofácies. Muitos autores ainda usam métodos que não são os mais apropriados como PCA, *K-means*, entre outros. Vimos que alguns desses métodos como o PCA podem ser utilizados na seleção dos atributos, mas nao funcionam bem como redutor de dimensionalidade devido justamente à natureza não linear dos dados, além de não preservar a estrutura global. Alguns autores usam métodos mais elaborados como SOM, porém complementados com classificadores ou clusterizadores que não são apropriados para o tipo de dado.

Tentamos então usar novos métodos como o t-SNE que se provou de difícil manejo e computacionalmente custoso além de não dar bons resultados, sendo mais apropriado para visualização do que para redução de dimensionalidade, além de também não preservar bem a estrutura dos dados. Foi então que decidimos utilizar métodos bem recentes que ainda não tinham sido utilizados em dados sísmicos como o UMAP, que é mais comumente utilizado na área de Engenharia Genética.

O mesmo se provou ser bastante eficaz para o nosso tipo de problema, funcionando de forma bastante rápida mesmo com computadores simples, preservando bastante da estrutura dos dados além de ser de fácil manejo, requerendo poucos e bem intuitivos parâmetros. O UMAP demonstrou também ser um bom regularizador dos dados, evitando assim a necessidade de passos de pré-processamento dos dados como escalonamento ou mesmo o uso de PCA, como requerido pelo SOM.

Após essa etapa buscamos um clusterizador que melhor se adequasse ao nosso tipo de dado. Como explicado na seção 4.5.1.1, o mais apropriado para nosso problema foi o HDBSCAN, que é um tipo de clusterizador hierárquico baseado em densidade. Porém muito do nosso dado era classificado como ruído, foi aí então que decidimos modificar o

algoritmo e passar a atribuir ao ruído o rótulo mais próximo. Essa abordagem se mostrou bastante frutífera, como pode ser visto na comparação feita na Figura 40.

Além disso utilizamos outras características do algoritmo que é a geração dos rótulos de maneira ordenada, colocando os grupos (*clusters*) mais próximos por similaridade. Isso facilitou bastante para fazermos a seleção do que seria cada tipo de sal de acordo com a referência (poço) ao carregarmos o dado classificado num software de interpretação. Após essa interpretação, pudemos então atribuirr os valores de velocidade de acordo com os grupos de sal e assim gerar o modelo de velocidade, que pode então ser exportado em formato ”bin” ou SEGY.

Esse trabalho serviu também para encontrarmos maneiras melhores e mais eficazes de classificação de sismofácies além do que tradicionalmente vem sendo feito. Aqui introduzimos o uso do UMAP como redutor de dimensionalidade e regularizador, além do uso de um HDBSCAN modificado que substitui a necessidade de um passo a mais que seria o uso de *K-means* ou KNN como normalmente utilizado por vários autores.

A combinação de UMAP + Modified HDBSCAN mais os passos de interpretação e calibração demonstrou ser a melhor escolha em termos de *shallow learning* para classificação de sismofácies e geração de modelo, seja de velocidade ou impedância. O método se revelou bastante versátil e pode ser usado também em outros ambientes deposicionais, tais como os silicíclasticos ou carbonáticos, tudo dependendo claro da correta seleção dos atributos de entrada.

6.1 Estudos Futuros

Muito ainda pode ser feito para melhorar o resultado, tais como etapas de pré-condicionamento dos dados de entrada, tipo filtragem do dado sísmico, normalização de fator Q para recuperar os efeitos de atenuação, entre outros. Outros atributos podem ser testados bem como outros parâmetros. O modelo gerado ainda necessita ser submetido a um processo de inversão para verificação da melhoria da imagem, agora que estão sendo considerados as estratificações do sal, porém este não era o escopo deste trabalho.

Referências

AGRAWAL, R. et al. Automatic subspace clustering of high dimensional data for data mining applications. In: *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*. IBM Almaden Research Center, San Jose, Calif.: Association for Computing Machinery, 1998. p. 49–60. Disponível em: <https://doi.org/10.1145/276305.276314>. Citado na página 48.

ALLAOUI, M.; KHERFI, M. L.; CHERIET, A. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In: _____. [S.l.: s.n.], 2020. p. 317–325. ISBN 978-3-030-51934-6. Citado na página 66.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. In: *The American Statistician*. [S.l.: s.n.], 1992. v. 46, p. 175–185. Citado na página 53.

ANKERST, M. et al. Some methods for classification and analysis of multivariate observations. In: *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data*. Association for Computing Machinery, 1999. p. 49–60. Disponível em: <https://doi.org/10.1145/304182.304187>. Citado na página 48.

AQRAWI, A. A.; BOE, T. H. *Improved fault segmentation using a dip guided and modified 3D Sobel filter*. [S.l.], 2011. Disponível em: <http://library.seg.org/>. Citado na página 37.

BARNES, A. E. *Handbook of Poststack Seismic Attributes*. [S.l.]: Society of Exploration Geophysicists, 2016. Citado 3 vezes nas páginas 10, 31 e 32.

BRONIZESKI, E. D. *Classificação de sismofáceis carbonáticas a partir da técnica Self-Organizing Maps (SOM)*. Dissertação de Mestrado em Engenharia Mineral — Escola Politécnica, 2018. Citado na página 66.

BROWN, A. R. Seismic attributes and their classification. *The Leading Edge*, v. 15, n. 10, p. 1090–1090, 1996. Disponível em: <https://doi.org/10.1190/1.1437208>. Citado na página 32.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160–172. ISBN 978-3-642-37456-2. Citado na página 46.

CARNEIRO, C. de C. et al. Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the brazilian amazon. *GEOPHYSICS*, v. 77, n. 4, p. K17–K24, 2012. Disponível em: <https://doi.org/10.1190/geo2011-0302.1>. Citado na página 39.

CHANG, H. et al. Sistemas petroliferos e modelos de acumulação de hidrocarbonetos na bacia de santos. *Revista Brasileira de Geociências*, v. 38, p. 29–46, 2008. Citado na página 22.

CHOPRA, S.; MARFURT, K. *Seismic Attributes for Prospect Identitification and reservoir Characterization*. [S.l.]: Society of Exploration Geophysicists, 2007. Citado na página 31.

CHOPRA, S.; MARFURT, K. J. Seismic attributes — A historical perspective. *GEOPHYSICS*, 2005. ISSN 0016-8033. Citado na página 18.

CONNOLLY, P. Elastic impedance. *The Leading Edge*, v. 18, n. 4, p. 438–452, 1999. Disponível em: <https://doi.org/10.1190/1.1438307>. Citado na página 36.

DUMAY, J.; FOURNIER, F. Multivariate statistical analyses applied to seismic facies recognition. *GEOPHYSICS*, v. 53, n. 9, p. 1151–1159, 1988. Disponível em: <https://doi.org/10.1190/1.1442554>. Citado na página 39.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: SIMOUDIS, E.; HAN, J.; FAYYAD, U. M. (Ed.). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. [S.l.]: AAAI Press, 1996. p. 226–231. Citado 3 vezes nas páginas 46, 48 e 49.

FERREIRA, D. J. A. et al. Unsupervised seismic facies classification applied to a presalt carbonate reservoir, santos basin, offshore brazil. *AAPG Bulletin*, v. 103, n. 4, p. 997–1012, 2019. ISSN 01491423. Citado na página 62.

FIDUK, J. C.; ROWAN, M. G. Analysis of folding and deformation within layered evaporites in blocks bm-s-8 & -9, santos basin, brazil. *Geological Society, London, Special Publications*, Geological Society of London, v. 363, n. 1, p. 471–487, 2012. ISSN 0305-8719. Disponível em: <https://sp.lyellcollection.org/content/363/1/471>. Citado na página 28.

FREITAS, J. *Ciclos deposicionais evaporíticos da Bacia de Santos: Uma análise cicloestratigráfica a partir de dados de 2 poços e de traço de sísmica*. Tese de Mestrado — Instituto de Geociências, 2006. Citado na página 27.

F.R.S., K. P. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor Francis, v. 2, n. 11, p. 559–572, 1901. Disponível em: <https://doi.org/10.1080/14786440109462720>. Citado na página 61.

GAMBOA, L. et al. Evaporitos estratificados no atlântico sul. In: _____. *Sal: Geologia e Tectônica. Exemplos nas Bacias Brasileiras*. [S.l.]: Beca Edições, 2008. p. 91–163. Citado 4 vezes nas páginas 10, 24, 27 e 28.

GAMBOA, L. A. P.; RABINOWITZ, P. D. The evolution of the rio grande rise in the southwest atlantic ocean. *Marine Geology*, v. 58, p. 35–58, 1984. Citado na página 24.

GAO, D. Application of three-dimensional seismic texture analysis with special reference to deep-marine facies discrimination and interpretation: An example from offshore angola, west africa. *AAPG Bulletin*, v. 91, p. 1665–1683, 2007. ISSN 1860-0980. Citado na página 74.

GARCIA, S. *Restauração estrutural da halotectônica na porção central da Bacia de Santos e implicações para os sistemas petrolíferos*. Tese (Tese de Doutorado) — Escola de Minas - Universidade Federal de Ouro Preto, Ouro Preto - Minas Gerais - Brasil, 2012. Citado 2 vezes nas páginas 10 e 24.

GUERRA, M. C. M.; UNDERHILL, J. R. Role of halokinesis in controlling structural styles and sediment dispersal in the santos basin, offshore brazil. *Geological Society, London, Special Publications*, Geological Society of London, v. 363, n. 1, p. 175–206, 2012.

ISSN 0305-8719. Disponível em: <https://sp.lyellcollection.org/content/363/1/175>. Citado na página 27.

HAMPSON, D. P.; SCHUELKE, J. S.; QUIREIN, J. A. Use of multiattribute transforms to predict log properties from seismic data. *GEOPHYSICS*, v. 66, n. 1, p. 220–236, 2001. Disponível em: <https://doi.org/10.1190/1.1444899>. Citado na página 35.

HOSSAIN, S. Application of seismic attribute analysis in fluvial seismic geomorphology. *Journal of Petroleum Exploration and Production Technology*, v. 10, n. 3, p. 1009–1019, 2020. Disponível em: <https://doi.org/10.1007/s13202-019-00809-z>. Citado na página 36.

JACKSON, C. A.-L. et al. Internal structure, kinematics, and growth of a salt wall: Insights from 3-D seismic data. *Geology*, v. 42, n. 4, p. 307–310, 04 2014. ISSN 0091-7613. Disponível em: <https://doi.org/10.1130/G34865.1>. Citado 3 vezes nas páginas 10, 28 e 29.

KALKOMEY, C. T. Potential risks when using seismic attributes as predictors of reservoir properties. *The Leading Edge*, v. 16, n. 3, p. 247–251, 1997. Disponível em: <https://doi.org/10.1190/1.1437610>. Citado na página 34.

KARNER, G. D.; GAMBOA, L. A. P. Timing and origin of the south atlantic pre-salt sag basins and their capping evaporites. *Geological Society, London, Special Publications*, Geological Society of London, v. 285, n. 1, p. 15–35, 2007. ISSN 0305-8719. Disponível em: <https://sp.lyellcollection.org/content/285/1/15>. Citado na página 24.

KOHONEN, T. *Self Organizing Maps*. 3. ed. [S.l.: s.n.], 2001. 105–114 p. Citado na página 61.

LAI, J. et al. Prediction of diagenetic facies using well logs: Evidences from upper triassic yanchang formation chang 8 sandstones in jiyuan region, ordos basin, china. *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles*, v. 71, n. 3, p. 34, 2016. Disponível em: <https://doi.org/10.2516/ogst/2014060>. Citado na página 39.

MAATEN, L. V. D.; POSTMA, E.; HERIK, J. V. D. Dimensionality reduction: A comparative review. In: . [S.l.: s.n.], 2009. Citado 2 vezes nas páginas 11 e 62.

MAATEN, L. van der; HINTON, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, v. 9, p. 2579–2605, 2008. Disponível em: <http://www.jmlr.org/papers/v9/vandermaaten08a.html>. Citado 2 vezes nas páginas 61 e 63.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967. p. 281–297. Disponível em: <https://projecteuclid.org/euclid.bsmsp/1200512992>. Citado na página 48.

MAUL, A.; SANTOS, M. C.; SILVA, C. Evaporitic section characterization and its impacts over the pre-salt reservoirs, examples in santos basin, offshore. In: . [S.l.: s.n.], 2018. Citado 6 vezes nas páginas 13, 19, 28, 49, 75 e 76.

MCINNES, L.; HEALY, J.; ASTELS, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, The Open Journal, v. 2, n. 11, mar 2017. Disponível em: <https://doi.org/10.21105/joss.00205>. Citado 4 vezes nas páginas 11, 51, 52 e 53.

MCINNES, L.; HEALY, J.; MELVILLE, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. Disponível em: <http://arxiv.org/abs/1802.03426>. Citado 4 vezes nas páginas 11, 42, 47 e 62.

MONTARON, B.; TAPPONNIER, P. A quantitative model for salt deposition in actively spreading basins. In: . [S.l.]: AAPG International Conference and Exhibition, 2009. Citado 4 vezes nas páginas 10, 25, 26 e 27.

MOQBEL, A. A.; WANG, Y. Carbonate reservoir characterization with lithofacies clustering and porosity prediction. *Journal of Geophysics and Engineering*, v. 8, n. 4, p. 592–598, 2011. ISSN 17422132. Citado na página 39.

MOREIRA, J. et al. Bacia de santos. In: *Boletim de Geociências da Petrobrás*. Rio de Janeiro: [s.n.], 2007. v. 15, n. 2. Citado 4 vezes nas páginas 10, 21, 22 e 23.

NG, R. T.; HAN, J. Clarans: A method for clustering objects for spatial data mining. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, v. 14, n. 5, p. 1003–1016, 2002. Disponível em: <http://www.cs.ecu.edu/dingq/CSCI6905/readings/CLARANS.pdf>. Citado na página 48.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 63 e 68.

RICCOMINI, C.; SANT'ANNA, L.; TASSINARI, C. Pré-sal: Geologia e exploração. *Revista USP*, v. 95, p. 33–42, 2012. Citado 2 vezes nas páginas 10 e 22.

RUSSELL, B. et al. Multiattribute seismic analysis. *The Leading Edge*, v. 16, n. 10, p. 1439–1443, 10 1997. ISSN 1070-485X. Citado na página 39.

SARHAN, M. A. The efficiency of seismic attributes to differentiate between massive and non-massive carbonate successions for hydrocarbon exploration activity. *NRIAG Journal of Astronomy and Geophysics*, v. 6, n. 2, p. 311 – 325, 2017. ISSN 2090-9977. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2090997717300652>. Citado na página 36.

TANER, M. T.; KOEHLER, F.; SHERIFF, R. E. Complex seismic trace analysis. *GEOPHYSICS*, v. 44, n. 6, p. 1041–1063, 1979. Disponível em: <https://doi.org/10.1190/1.1440994>. Citado na página 31.

TEIXEIRA, L. *Interpretação sísmica e estratigráfica dos evaporitos da Bacia de Santos*. 2020. Apresentação de exame de qualificação de Doutorado. Citado 2 vezes nas páginas 10 e 33.

THORNDIKE, R. L. Who belongs in the family? *Psychometrika*, v. 18, p. 267–276, 1953. ISSN 1860-0980. Disponível em: <https://doi.org/10.1007/BF02289263>. Citado na página 68.

WANG, W.; YANG, J.; MUNTZ, R. Sting: A statistical information grid approach to spatial data mining. In: *VLDB*. [S.l.: s.n.], 1997. Citado na página 48.

WU, Y. et al. Machine learning for locating organic matter and pores in scanning electron microscopy images of organic-rich shales. *Fuel*, v. 253, p. 662 – 676, 2019. ISSN 0016-2361. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0016236119307628>. Citado na página 39.

XIANG, S.; NIE, F.; ZHANG, C. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, v. 41, n. 12, p. 3600 – 3612, 2008. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320308002057>. Citado na página 71.

YANG, P. et al. Gswo: A programming model for gpu-enabled parallelization of sliding window operations in image processing. *Signal Processing: Image Communication*, v. 47, p. 332 – 345, 2016. ISSN 0923-5965. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0923596516300601>. Citado na página 59.

YILMAZ, O. Seismic data analysis. In: _____. Society of Exploration Geophysicists, 2001. ISBN 978-1-56080-094-1. Disponível em: <https://doi.org/10.1190/1.9781560801580>. Citado na página 18.

Anexos

ANEXO A – Artigo Submetido a Revista Brasileira de Geofísica.

New method for identifying low velocity salt stratifications using the machine learning approach over seismic attributes

Flavio Costa de Mesquita[a] (GISIS-UFF), Marco Antonio Cetale Santos[a] (GISIS-UFF), Alexandre Rodrigo Maul[a,b] (Petrobras/GISIS-UFF) and Alex Laier Bordignon[c] (IME-UFF)

[a]*Seismic Imaging and Inversion Group (GISIS), Av. Gen. Milton Tavares de Souza, s/n - Gragoata - Campus da Praia Vermelha, Niteroi - RJ, Brazil*
[b]*Petrobras - Reservoir Geophysics, Av. Henrique Valadares, 23 - 6º andar, Centro, Rio de Janeiro – RJ, Brazil*
[c]*Mathematics and Statistics Institute (IME), Rua Professor Marcos Waldemar de Freitas Reis, s/n, Blocos G e H, Campus do Gragoata - Niteroi - RJ, Brazil*

ARTICLE INFO

Keywords:
machine learning
salt drilling
risk reduction
seismic attributes

ABSTRACT

The formation of evaporites occur in stages following many specified environmental conditions. The stages order are the result of evaporation rates which generate a layering pattern, also denominated stratifications. To facilitate this referred layer identification using seismic data some have purposed to group the main evaporitic mineral into three major facies: Halite, High Velocity Salts - HVS (anhydrite and gypsum) and Low Velocity Salts - LVS (carnallite, sylvite and tachyhydrite). Inside the last group tachyhydrite is one of the major causes for fluid loss during the salt drilling operation in the Santos Basin, Brazil Offshore. Due to its high solubility leading sometimes to events of well abandon causing a great financial problem to the project. In this work we used the traces of seismic attribute sections, which served as input for a selected Machine Learning algorithm to identify the main occurence of the existing minerals. Thereby, allowing to identify from these 2D sections the stratifications within the salt layer and consequently their characterization of the corresponding salt type, giving to the well drilling engineers a prior knowledge of the lithology ahead and the ability to adjust the drilling parameters accordingly.

1. Introduction

Problems associated with the drilling of oil and gas wells are largely due to the disturbances of earth stresses around the borehole. These are caused by the creation of the hole itself and by drilling mud/formation interaction. Thus, a hole is kept open (or stable) by maintaining a balance between earth stresses and pore pressure on one side and well bore mud pressure and chemical composition on the other side. Whenever this balance is disturbed, well bore problems occur. Drilling operations in salt zones have gained importance in Brazil due to the discovery of large oil and gas reserves in the Pre-Salt province. Thus, the pursuit of excellence in such operations is requiring considerable development of new operational practices and technologies. High operating costs associated to deep water drilling is placing additional emphasis on drilling performance in order to reduce the operational time, without losing the quality of the wells.

In the Brazilian Pre-Salt scenario, the most common salts encountered are anhydrite (CaSO4), halite (NaCl), carnallite (KCl.MgCl2 .6H2O) and tachyhydrite (CaCl2.2MgCl2 .12H2O), the deposition sequence normally follows the same order. While the salt solubility in water, for these salts, is the opposite: Tachyhydrite >Carnallite > Halite > Anhydrite.

One of the most common causes of drilling problems on the Santos Basin, Brazilian Offshore, is loss of fluid circulation due to mineral absorption. It is the significant and continuing loss of whole mud to a formation, it is probably the most common, and overall, the most costly drilling well problem. Depending on its severity, lost circulation can lead to:

- Increased costs for drilling mud and associated materials.
- Formation damage and decreased productivity.
- Wellbore fluid level drops, resulting in increased potential for stuck pipe, borehole instability and kicks.
- Lost formation evaluation data, since the information normally obtained from drilled cuttings and mud returns may be unobtainable

This paper aims the creation of a model, based on seismic data, for the salt section in the Santos Basin, taking in consideration the different types of salt and their stratifications. A good model able to identify some specific salt compositions on the seismic section may be of great use to the drilling engineers, allowing the change of drilling parameters before entering the Low Velocity Salt layers that have high solubility, therefore preventing lost circulation problems.

One of the methods to avoid lost circulation is the correct use of mud types for each section of well, as mentioned in Lomba *et al*., 2013[13]: *The drilling of evaporites with high mobility with synthetic based muds may result in wellbore collapse, high torques, difficult reaming, stuck pipe, deviations, casing collapse and, eventually, loss of the well. In some cases, stress concentration, after drilling the salt, may cause stuck bit, mainly during the connections, demanding the injection of fresh water pills for its liberation. In some*

⋆This document is the result of the research project funded by Petrobras.

ORCID(s): 0000-0002-4540-520X (F.C.d. Mesquita); 0000-0002-3556-3140 (A.R. Maul)

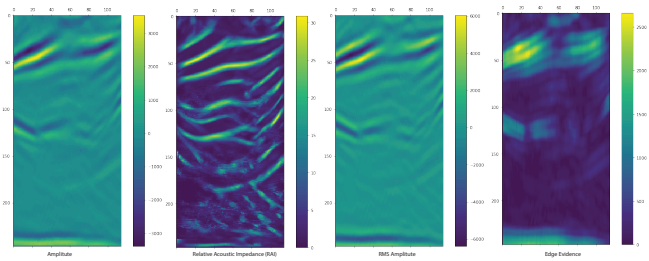

Figure 1: Attributes used at this seismofacies classification. From left to right: Amplitude, RAI, RMS Amplitude and Edge Evidence.

wells, the frequent use of those pills results in such an irregular section as if it had been drilled with a water based fluid, thus rendering the mentioned problems more critical.

A less expensive way of avoiding a constant changing of drilling fluid, which is time and resource demanding, is to know a priori the lithology ahead. In this work, we propose modeling salt layers with a combination of sismofacies identification based on machine learning algorithms on seismic attributes. The model can be used during drilling, and can be the key to avoiding loss of circulation in basins where salt stratifications with LVS are a problem.

The remainder of this paper is structured as follows. Section 2 presents related works and define the seismic attributes used. Section 3 describes the proposed workflow. Section 4 discusses about the obtained results. Finally in Section 5 we present conclusions and possible future works.

2. Materials and Methods

The field-reservoir under study is located in the central portion of the Santos Basin, about 180 km off the coast of the municipality of Rio de Janeiro in a water depth of approximately 1,900 m. The reservoirs of this field are situated between 5,000 and 6,000 m below the sea level and under a layer of salt, the Ariri Formation, which can range from a few hundred meters to over 2,000 m, (Mohriak *et al*., 2012) [21].

It is know that this formation is not homogeneous and that it is composed of different types of stratified minerals, the evaporites. Depending on the complexity of the field and the saline structuring, there are variations in these “strata“ and in their thickness. Oliveira *et al*. (2015)[23], indicated an inverse correlation between “salt thickness“ and “salt velocity variation“. There are different types of evaporite minerals within the evaporitic section in the Santos and Campos basins, and research carried out in log analysis shows that not all these types of minerals will be seismically detectable by the amplitude (Gobatto *et al*., 2016)[8].

To date, Machine Learning has been used in the seismic scale to predict geological structures, stratigraphy, and rock and fluid properties, usually through seismic interpretation and inversion. In order to achieve it, fully convolutional deep networks have been used in the area of fault interpretation (Long *et al*.,2015)[14], 3D convolutional neural networks (Waldeland *et al*., 2018)[26] and deep encoder-decoder networks for stratigraphic interpretation (Badrinarayanan *et al*., 2015)[3]. These techniques classify a 3D Post-Stack data set based on 3D sub-cubes or 2D sections, and require a relatively low number of labels. Interpretation in seismic images has long used texture attributes to better identify and highlight areas of interest. These can be seen as feature maps in seismic texture. In the salt case, it can be noted that the texture in the salt masks are quite chaotic, where the surrounding seismic is more ”striped”. However this is not always true, as we can see in the Santos Basin, where the salt layer is highly stratified.

Due to the low amount of wells containing samples of the 4 types of salt we want to identify, we propose using a multi-attribute technique, although the analysis of several attributes presents a problem that is of difficult solution. It makes the task almost impossible trough traditional methods, requiring the application of algorithms capable of identifying the intrinsic particularities between the attributes highly necessary. Artificial neural networks (ANNs) are tools capable of capturing these relationships between attributes including its non-linear relationships (Hagan *et al*., 2002)[9]. Comparisons between the application of supervised and unsupervised ANN in geoscience show that the choice is dependent on the analysis in each case, there is no rule to be applied to all cases such as in Bhattacharya; Carr; Pal, 2016[4] or Sayago *et al*., 2012[25].

In the specific case of this work, we propose a technique that is composed of two classification steps: one unsupervised and one supervised, as in Moqbel and Wang (2011) [22]. Our first attempt to solve the mentioned problem was presented by Mesquita *et al*., 2019[20], however we realized the results were high computationally demanding and the parameters not very intuitive, so we decided to follow another way as we are presenting here that is comparable to apply Self Organizing Maps - SOM (Kohonen, 2001)[12], with the advantage we have better control of each step. The SOM is a non-supervised network which main characteristic is to perform dimensionality reduction and clustering resulting in a bi-dimensional map representing the input data grouped by similarity.

Differently to follow the guides proposed by Wang2011, we use the raw seismic trace instead of working with pixels, which makes the algorithm faster besides producing better results. For this we use a novel dimensionality reduction technique followed by an unsupervised state of the art clusterization algorithm to model a set of sismofacies. The sismofacies are then calibrated to match the types of salt most common in the Santos Basin: Tachyhydrite, Carnallite, Halite and Anhydrite. Once we have the labels (classes) calibrated, we can perform classification of a Post-Stack Depth Migration (PSDM) data.

The input data consists of pieces of volume attributes extracted from a Post-Stack Depth Migration seismic type. From these attribute volumes, inlines were extracted on the position of each well that shows ocurrences of these salts, in special the LVS.

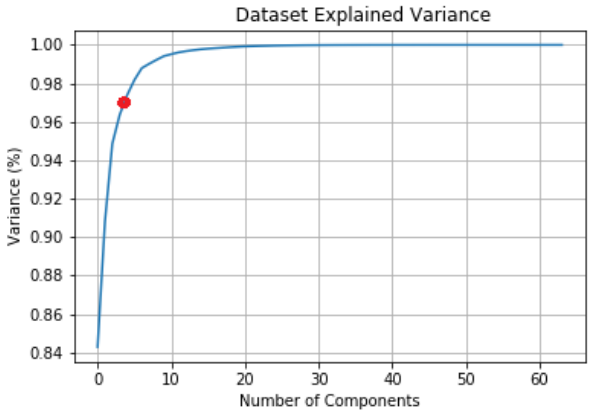

Figure 2: Dataset Explained Variance, with 4 attributes we have 97% of the explained variance.

2.1. Seismic Attributes

Seismic attributes are the components of the seismic data which are obtained by measurement, computation, and other methods from the seismic data. Seismic Attributes were introduced as a part of the seismic interpretation in early 1970, (Chopra and Marfurt, 2005)[6]. Since then many new attributes were derived and computed. Seismic attribute analysis can extract information from seismic data that is, otherwise, hidden in the data and have been used to identify prospects, ascertain depositional environments. There are now dozens of distinct seismic attributes calculated from seismic data and applied to the interpretation of geologic structure, stratigraphy, and rock/pore fluid properties[6]. Extracting all the potential information hidden in the seismic data using a single attribute almost never occurs. Therefore a combination of attributes or multiattribute analysis is carried out to gauge more information overall than what is possible with only one attribute.

2.1.1. Selection of the Attributes

A crucial problem in any multiattribute analysis is the selection and the number of seismic attributes to be used. Kalkomey (1997), [11] showed that the probability of observing a spurious correlation increases as the number of control points decreases and also as the number of seismic attributes being used increases. Based on knowledge and experience we selected the attributes that best highlighted the layers we are focusing on this work, the LVS. Then we did a variance analysis in order to know the reasonable amount of attributes to use, see Figure 2.

According to the graph of Figure 2, with only 4 attributes we have 97% of the dataset variance, so, in order to avoid redundancy and a possible overfitting of the model, knowing that we only need 4 attributes, we eliminated the ones that had high correlation plus the ones that not highlighted the salt reflections when matched to the well log as proposed by Hampson et al. (2001)[10], the correlation result can be seen on Table 1

According to the correlation results presented on Table 1, we end up choosing the following 4 attributes:

	Amplitude	RMS Amplitude	Relative Acoustic Impedance	1st Derivative	Instantaneous Frequency	TECVA	Edge Evidence	Amplitude Contrast
Amplitude	1,000	0,674	0,585	0,149	0,149	0,122	0,023	0,175
RMS Amplitude		1,000	0,584	0,118	0,197	0,080	0,026	0,173
Relative Acoustic Impedance			1,000	0,208	0,988	0,812	0,016	0,865
1st Derivative				1,000	0,109	0,037	0,018	0.018
Instantaneous Frequency					1,000	0,991	0,056	0.004
TECVA						1,000	0,482	0,482
Edge Evidence							1,000	0,034
Amplitude Contrast								1,000

Table 1
Correlation of the Proposed Attributes

- Amplitude - The seismic standard reflectivity amplitude, is an attribute related to the physical properties of the subsurface as a function of the reflections at the different acoustic interfaces;

- Root Mean Square (RMS) Amplitude - It is a basic amplitude attribute, a statistical measure of the magnitude of variation in amplitude throughout a dataset. Generally, higher acoustic impedance variations (associated with variations within stacked lithology) will result in higher RMS values. It is computed in a sliding tapered window of N samples as the square root of the sum of all the trace values x squared where w and n are the window values as presented in Equation 1.

$$x_{rms} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} w_n x_n^2} \tag{1}$$

- Relative Acoustic Impedance (RAI) - It is a stratigraphic method, and is the product of density and seismic velocity, which varies among different rock layers and commonly symbolized as Z. This attribute shows apparent acoustic contrast, indicates sequence boundaries, unconformity surfaces and discontinuities. It can also indicate porosity or fluid content in the reservoir. It defines the density contrast encountered at the interface between two distinct lithologies and is calculated by integrating the trace, then passing the result through a High-pass filter to reduce potentially introduced low-frequency noise, (Connolly, 1999).[7] In our case a cutoff frequency of 10 Hz was used, because it was the dominant frequency of the wavelet on the salt layer. It can be expressed as equation 2 below:

$$< Z_n >=< Z_0 > \exp(2\sum_{j=0}^{n} R_j \Delta t) \tag{2}$$

  Where $< Z >$ indicates the average impedance about a layer j, R is the reflectivity, j is the time sample varying from 0 to n, and t is time.

- Edge Evidence (Structural Method) - It is a statistical edge enhancement method used to delineate fault and

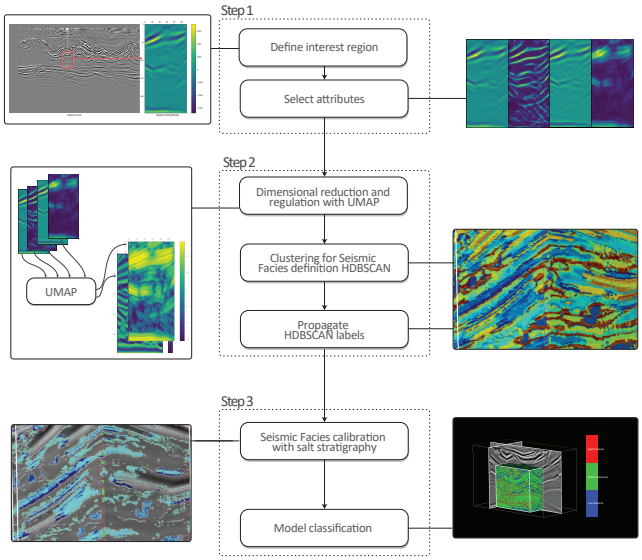

Figure 3: Workflow for seismofacies classification.

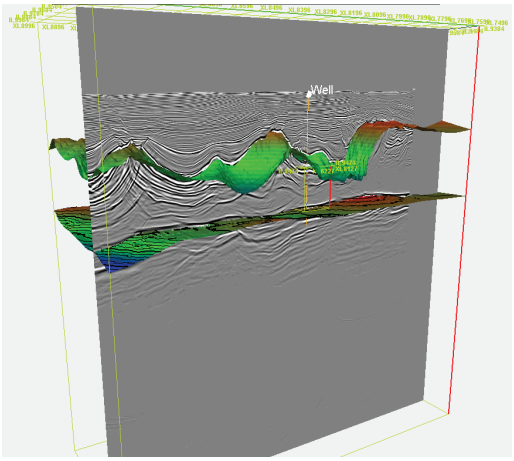

Figure 4: Seismic line from the volume seismic with the salt top and base mapped. The red, green and blue colors are associated with the surface curvature. In the middle, the yellow line is the well trajectory.

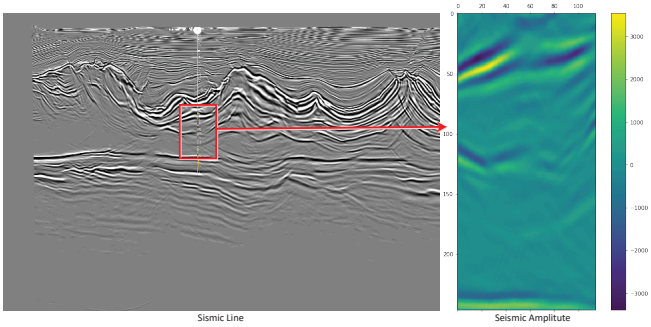

Figure 5: Data slice from top to bottom of salt layer on a 50 traces radius from the well location on the inline.

salt body borders within seismic data.The algorithm is related to the Radon transform and Hough transforms but uses an integral to detect edges within an image, and is limited to a user-defined window. The edge evidence attribute works by searching locally in all directions for line segments where the values on the line differ significantly from the surrounding values, (Aqrawi and Boe, 2011). [2]

The four attributes computed over a seismic line can be seen in Figure 1 and in the left side of Figure 7 we show the attributes computed on the region of interest.

The same method can be used with an arbitrary number of attributes and can be used to assist in mapping other lithologies.

3. Workflow

In this section we will explain individually the processing steps that were used to arrive at the model of the types of salts, to facilitate reading we have grouped the processing into three steps depicted in Figure 3.

Step 1: The first step in a multi-attribute machine learning interpretation workflow is to select the data that will be used as input. As we are interested in modelling the salt layers, we start by modeling the top and the base of the salt, see Figure 4. In this region we have one well data with information about the salt velocity, and we select the seismic line crossing this well, see Figure 5 (left). In this seismic line, we selected an rectangular window with 120 traces, and 250 samples for each trace, with the well in the middle as depicted in Figure 5. For this interest region we selected four seismic attributes, as in Figure 7. We end this step with four matrices containing all the data we will use as input for the next step. We combine these matrices as a tensor, lets call it I, and we refer to a single attribute value as $I[h, w, c]$ were h is the row of the matrix, w is the column and $c \in \{1, 2, 3, 4\}$ represents the attribute. Also we will denote by $I[h, w]$ a four dimensional vector composed by the attribute samples:

$$I[h, w] = [I[h, w, 1], I[h, w, 2], I[h, w, 3], I[h, w, 4]]$$

Step 2 With the input data I from the previous step, we define a set of samples, where each sample $I[h, w]$ is a vector with four dimensions. These samples are then transformed using the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [19] and the transformed samples are then clustered using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [18].

The UMAP technique (McInness, 2018)[19], it is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. Dimension reduction seeks to produce a low dimensional representation of high dimensional data that preserves relevant structure. Dimension reduction algorithms tend to fall into two categories; those that seek to preserve the distance structure within the data and those that favor the preservation of local distances over global distance, UMAP falls into the latter category.

The theoretical foundations for UMAP are largely based

in manifold theory and topological data analysis. At a high level, UMAP uses local manifold approximations and patches together their local fuzzy simplicial set representations to construct a topological representation of the high dimensional data. Given some low dimensional representation of the data, a similar process can be used to build an equivalent topological representation. UMAP then optimizes the layout of the data representation in the low dimensional space, to minimize the cross-entropy between the two topological representations. The theoretical description of the algorithm works in terms of fuzzy simplicial sets. Computationally this is only tractable for the one skeleton which can ultimately be described as a weighted graph. This means that, from a practical computational perspective, UMAP can ultimately be described in terms of, construction of, and operations on weighted graphs. In particular this situates UMAP in the class of k- neighbour based graph learning algorithms such as Laplacian Eigenmaps, Isomap and t-SNE, vanDerMaaten (2008)[15].

The UMAP was created to be used mainly as a dimensional reduction method, but its formulation also has a regularization effect. In this work we are more interested in the regulatory effect. The UMAP can be used to transform groups of points that are very dense into groups with more uniform point density, the minimal distance between points in the embedding can be controlled by the minimal distance hyperparameter. This has a positive effect because it facilitates the work of clustering algorithms that work towards expanding sample selection, such as DBSCAN and HDBSCAN. Comparing the result of in Figure 7 (b) to Figure 7 (c) we can see the effect of this regularization, when we employ the UMAP the seismic facies are much more defined.

Integrating UMAP with a hierarchical clustering algorithm provides the best evidence of existing local and global structures of data. Here we used Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInness, 2017) [18]. It performs DBSCAN, (Campello, 2013)[5] over varying epsilon values and integrates the result to find a clustering that gives the best stability over epsilon, which specifies how close points should be to each other to be considered a part of a cluster. This allows HDBSCAN to find clusters of varying densities (unlike DBSCAN), and be more robust to parameter selection. In Figure 7 we compare the result of the proposed clustering method (UMAP + HDBSCAN) against the k-means algorithm, MacQueen (1967) [16] and with the original HDBSCAN without the UMAP step.

In this work, we have chosen four seismic attributes and the UMAP was used to create an embed space with two components. The two new components will be the input of the clustering algorithm. The UMAP have two principal hyperparameters: number of neighbors that balances local versus global structure in the data and minimal distance that controls how tightly UMAP is allowed to pack points together. In this work we set number of neighbors to 35 and minimal distance to 0.01. We selected these values based on a grid search and visual inspection of the facies distribution. In

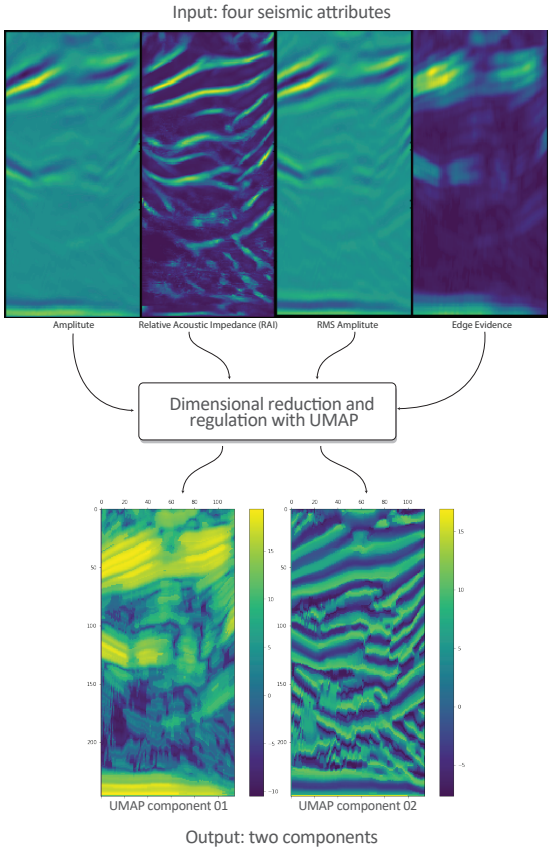

Figure 6: Dimensional reduction and regularization with UMAP

Figure 6 we depict the inputs and outputs generated by the UMAP algorithm with the selected hyperparameters.

The HDBSCAN was used with the default parameters, except for the metric. We chose the Mahalanobis metric, this metric is useful because it is scale invariant, in this way more robust to the possible unscaled output of the UMAP transform. The seismic facies obtained with this methodology are presented in Figure 7 (c).

One problem with HDBSCAN is that in its formulation it tries to separate the clusters from the background data, that is it detects the points that are between the clusters and classifies them as noise. This effect can be seen in Figure 7 (c), where a lot of points are classified as belonging to the noise class (the darker blue points). To fix this problem we modified the HDBSCAN algorithm, for each point classified as noise we look for the nearest classified point and transfer its label. The result of this modification is presented in Figure 7 (d).

After the application of the proposed clustering method (UMAP + Modified HDBSCAN), we ended with a seismofacies map of the interest region, we want to expand this seismofacies classification to all the seismic data. To make this expansion we directly use the model, transforming new samples with the UMAP and use the condensed tree created by HDBSCAN method to predict the labels [18]. The classification for new data points took approximately half millisecond

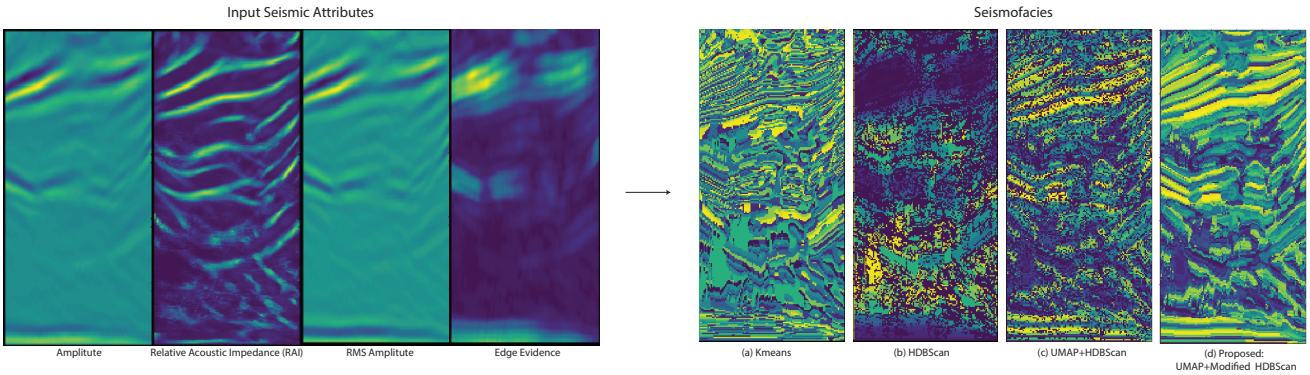

Figure 7: Input seismic attributes and Seismofacies found by different methods

per sample in a Intel Core I5-9600KF CPU, this is a little on the slow side, but the process can be directly done using parallel computing. See Figure 8 (c) where the seismic facies information was propagated to the seismic cube.

Step 3: In this step we want to map the seismic facies from Step 2 with the salt stratigraphy. This is a problem with incomplete information, this is due to the fact that the seismic its not perfect and contains several sources of uncertainty. Look for example the case presented in Figure 9 (a), the well data show that there are LVS salts (green points) in the middle of the picture, but the seismic do not display a reflection over the well trace, but to the left and right of the well, the reflections are present. There are several possible correct mappings from the seismic to the salt layers, in this sense, to create the mapping we use interpretation, to facilitate the process of creating this map we propose two main sources of information: knowledge about salt distribution and well logs.

As we can see in Table 2, a study performed at Santos Basin considering 182 wells shows the average percentage for the different types of salts found, where the Halite counts for 81%, the HVS counts for 12% and the LVS counts for 7%. The well logs also can be used, but there are several difficulties, like in the case were the wells are not perfectly adjusted with the seismic data, or in regions where the seismic have problems, due to physical properties of the salt or other problems.

In this work we want to map the seismic facies to three different stratigraphic groups: LVS, Background (Halite) and HVS as presented in Table 3. To create the mapping we loaded the seismic facies obtained in Step 2 into an interpretation software, were we can visualize the well data and the seismic facies data, as shown in Figure 8. The green points over the well trace indicate the regions with the LVS, and we select the seismic facies associated with those regions. The resulting process of this selection is presented in Figure 8 (d) and in more detail in Figure 9 (b).

The selection of facies is greatly facilitated by the fact that we employ the HDBSCAN algorithm. Usual cluster algorithms generate labels that do not have an specific order as in Figure 7 (a), differently the HDBSCAN generate ordered seismic facies labels, the labels are integer numbers, this means that,in general, facies with similar attribute characteristics have closer labels, this structure can be seen in Figure 7 (c) and (d). The same process is used to select the facies associated with HVS and the remaining facies are associated to the background Halite, the final result is a 3D map with the salt classification, see Figure 12.

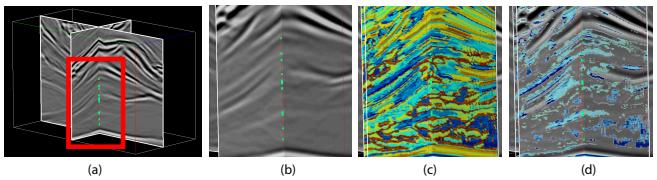

Figure 8: (a) 3d Seismic data and Well data, green points indicate LVS in the well trace; (b) 3d Seismic data and Well data, green points indicate LVS in the well trace, detail; (c) The seismic facies; (d) LVS selected facies.

Once we classify the data we can then assign values accordingly using our prior knowledge of the lithology using the wells as reference. We used the theoretical values to the salt layers according to Table 3. It presents the mineral grouping as proposed by Maul et al.(2018)[17], indicating the chemical formula of each mineral as well as their acoustic properties.

The whole process is very accessible to anyone with minimum programming skills, this workflow was done using Python 3.7 and ScikitLearn package[24] plus Petrel interpretation software, but any other could be used.

4. Results

As can be seen in Figure 10 and in more detail in Figure 11, the Amplitude inline was classified and values were assigned to each class found, where the color green represents the background lithology, dark blue color represents a low velocity lithology and yellow color the high velocity lithology. In this case we did the classification for the whole inline disregarding the post and pre salt zones, the same methodology can be used only changing the model where each different lithology is a class and the input seismic attributes

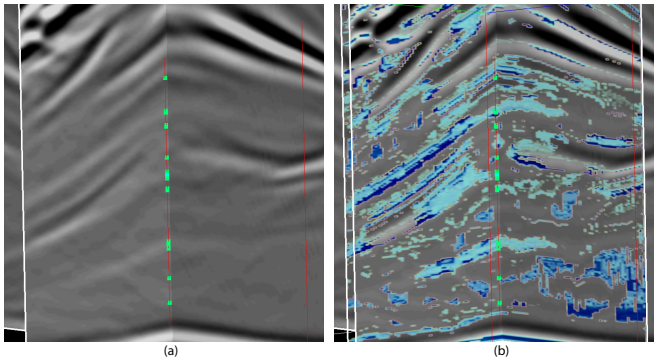

Figure 9: (a) 3d Seismic data and Well data, green points indicate LVS in the well trace, (b) LVS selected facies.

Field	# Wells	% LVS	LVS AIV	% Halite	HALITE AIV	% HVS	HVS AIV	WIV
1	20	8	4018,56	83	4480,88	8	5210,27	4462,56
2	29	9	4218,47	82	4563,69	9	4975,84	4567,53
3	17	12	4054,42	77	4498,25	12	4989,92	4505,66
4	3	13	3971,00	71	4507,09	16	4927,59	4505,04
5	5	3	4167,00	84	4538,00	13	5123,33	4576,00
6	7	3	4264,19	80	4509,87	17	5061,36	4596,05
7	72	8	4122,33	81	4526,47	11	5105,84	4560,03
8	25	4	4182,53	88	4533,59	8	5003,35	4547,16
9	4	6	4055,63	81	4486,58	13	5077,49	4535,67
TNW	182							
AVG		7	4117,13	81	4516,05	12	5052,78	4539,52

Table 2
LVS: Low Velocity Salts; HVS: High Velocity Salts; AIV: Average Interval Velocity; WIV: Weighted Interval Velocity; TNW: Total Number of Wells; Interval Velocity (m/s); AVG: Average. Modified from Maul et al. (2018)[17].

GROUP	MINERAL	COMPOSITION	DENSITY (g/cm3)	INTERVAL VELOCITY (m/s)
LVS	Tachyhydrite	CaMg2Cl6.12(H20)	1,57	3.300
	Carnallite	KMgCl3.6(H20)	1,66	3.910
	Sylvite	KCl	1,86	3.910
BACKGROUND	Halite	NqCl	2,10	4.550
HVS	Gypsum	CaSO4.2(H20)	2,35	5.810
	Anhydrite	CuSO4	2,98	6.100

Table 3
Mineral groups and respective properties, average values adapted by Maul et al. (2018) covering 182 wells in the Santos Basin.

for these depositional environments (siliciclastic and carbonatic) are different from the ones used in this work. Also the data slice chosen must match the environment to be studied. The same can be done for 3D volumes, as can be seen in Figure 12. Any value could be attributed to each class found, we could assign impedance values as well, the versatility of the method can be further explored.

5. Conclusions

From what we could learn on this work, a Machine Learning workflow combined with the right seismic attributes, can be applied to identify salt stratification and generate a more accurate model to seismic processing in regions where salt stratification is a problem. The same can be used for geomechanics purpose as we proposed here, the seismic resolution is still a problem, although could be somehow mitigated using the right attributes and the well control. The 3 different groups of salt can be differentiated, LVS, Halite and

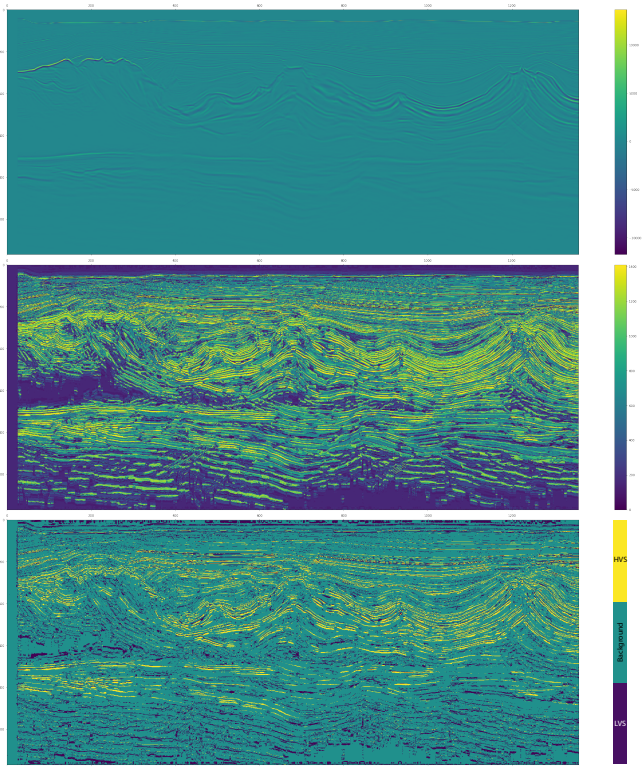

Figure 10: Method applied to an entire seismic line: top, original seismic amplitudes; middle, seismic facies detected by the proposed UMAP + Modified HDBSCAN algorithm; bottom, calibrated salt lithology.

HVS. Although sometimes was not possible to differentiate the Carnallite from the Tachyhydrite, the results can still be used as an aid to the drilling projectist to plan the drilling parameters before the well is actually drilled, avoiding the fluid losses when entering the LVS layers that leads to many time and financial issues.

Many pre-process steps can be added as noise filtering and Q factor normalization to recover the attenuation effects, but even without these steps the method show its strength as a much better alternative to the classic PCA+K-means normally used, being comparable or better than Self Organizing Maps which by itself doesn´t solve the problem without a pre-conditioning PCA plus further steps.

Also we introduce here the use of UMAP as a regularization step to the data, diminishing the need of normalization, filtering, PCA and many other steps normally used in traditional methods, plus a modified HDBSCAN, which is a hierarchical clusterization method which works with varying densities distribution producing much beter results than K-means or KNN, Altman (1992)[1]. It all starts from a criterious data selection, starting from the wells, to the data slice on the seismic inlines. Then applying dimensional reduction and clusterization followed by a calibration step. Finally we end up with a classified seismic volume. The classified data is then assigned values according to the classes and a velocity model. Summarizing, we started from an unsupervised classification technique, generated the model and used the

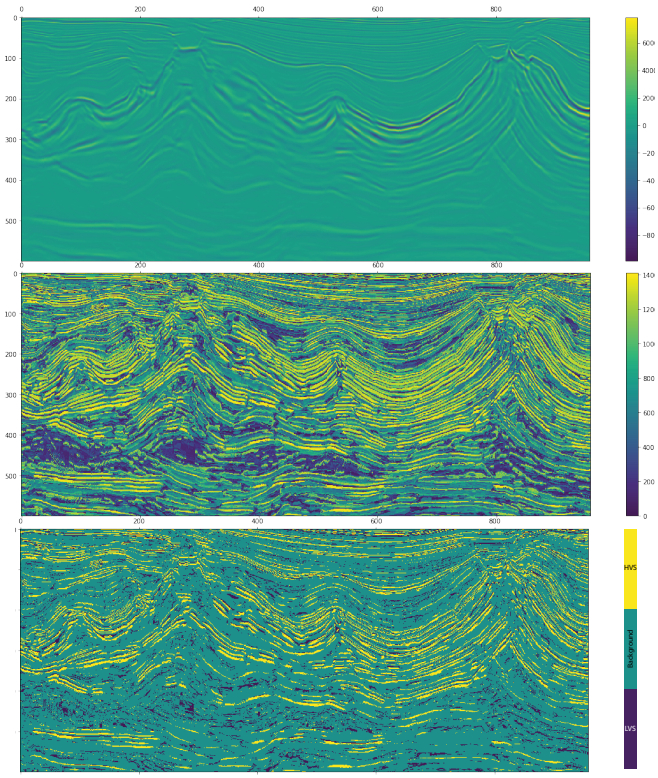

Figure 11: Detail from Figure 10: top, original seismic amplitudes; middle, seismic facies detected by the proposed UMAP + Modified HDBSCAN algorithm; bottom, calibrated salt lithology.

wells and the salt proportions as base to map the seismofacies to the salt lithology. We ended up with a model showing the distribution of the three salt stratification's groups.

As a future work we plan to use autoencoders for detection of rare events, since the tachyhydrite is not a common salt, we judge this technique to be promissing.

6. Acknowledgements

We would like to acknowledge the Brazilian National Petroleum Agency (ANP) for providing the dataset for this work, the Euclides da Cunha Foundation (FEC) for the scholarship, Schlumberger for the academic license of Petrel software and GISIS for the academic suppport and infrastructure.

References

[1] Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 3, 175–185. doi:10.1080/00031305.1992.10475879.

[2] Aqrawi, A., Boe, T., 2011. Improved fault segmentation using dip guided and modified Sobel filter. SEG Technical Program Expanded Abstracts , 999–1003doi:10.1190/1.3628241.

[3] Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, arXiv preprint. URL: http://arxiv.org/abs/1511.00561.

[4] Bhattacharya, S., Carr, T.R., Pal, M., 2016. Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the bakken and mahantango-marcellus shale, usa. Journal of Natural Gas Science and Engineering 33, 1119–1133.

[5] Campello, R.J., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7819 LNAI, 160–172. doi:10.1007/978-3-642-37456-2_14.

[6] Chopra, S., Marfurt, K.J., 2005. Seismic Attributes - A historical perspective. GEOPHYSICS, VOL. 70, NO. 5 , 3–28doi:10.1190/1.2098670.

[7] Connolly, P., 1999. Elastic Impedance. The Leading Edge 18, 438–452. doi:10.1190/1.1438307.

[8] Gobatto, F., Maul, A., Falcão, L., Teixeira, L., Boechat, J.B., González, M., González, G., 2016. Refining velocity model within the salt section in Santos Basin: an innovative workflow to include the existing stratification and its considerations. Technical Report. URL: http://library.seg.org/.

[9] Hagan, M., Demuth, H., Beale, M., 2002. Neural Network Design. Number v. 10 in Neural network design, Campus Pub. Service, University of Colorado Bookstore. URL: https://books.google.com.br/books?id=bUNJAAAACAAJ.

[10] Hampson, D.P., Schuelke, J.S., Quirein, J.A., 2001. Use of multiattribute transforms to predict log properties from seismic data. GEOPHYSICS 66, 220–236. URL: https://doi.org/10.1190/1.1444899, doi:10.1190/1.1444899, arXiv:https://doi.org/10.1190/1.1444899.

[11] Kalkomey, C.T., 1997. Potential risks when using seismic attributes as predictors of reservoir properties. The Leading Edge 16, 247–251. URL: https://doi.org/10.1190/1.1437610, doi:10.1190/1.1437610, arXiv:https://doi.org/10.1190/1.1437610.

[12] Kohonen, T., 2001. Self Organizing Maps. 3 ed. doi:10.1007/978-3-642-56927-2.

[13] Lomba, R.F.T., Pessanha, R.R., Cardoso Jr, W.F., Lomba, B., Folsta, M., Goncalves, J.T., Teixeira, G.T., 2013. Lessons Learned in Drilling Pre-Salt Wells With Water Based Muds, in: OTC Brasil, Offshore Technology Conference. URL: http://www.onepetro.org/doi/10.4043/24355-MS, doi:10.4043/24355-MS.

[14] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440. URL: http://arxiv.org/abs/1411.4038.

[15] van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[16] MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, Calif.. pp. 281–297. URL: https://projecteuclid.org/euclid.bsmsp/1200512992.

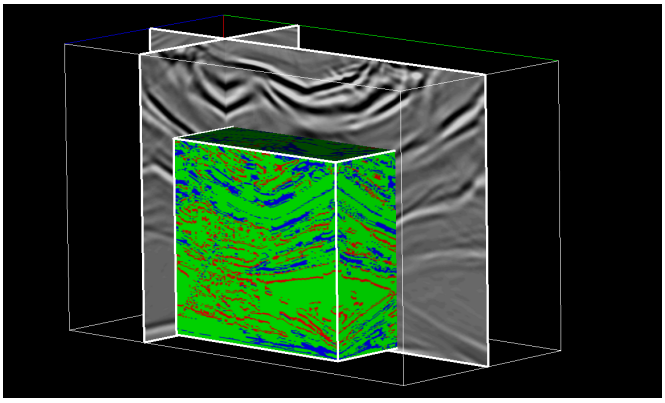

Figure 12: 3D seismic volume classified and loaded to interpretation software.In blue LVS, green represents the Background Halite and red indicates the HVS.

[17] Maul, A.R., Santos, M.A.C., Silva, C.G., 2018. Few Considerations, Warnings and Benefits for the E&P Industry When Incorporating Stratifications Inside Salt Sections. Revista Brasileira de Geofísica doi:10.22564/rbgf.v36i4.1981.

[18] McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software 2. URL: https://doi.org/10.21105/joss.00205, doi:10.21105/joss.00205.

[19] McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction URL: http://arxiv.org/abs/1802.03426, arXiv:1802.03426.

[20] Mesquita, F., Cetale Santos, M., Maul, A., Bordignon, A., 2019. Identifying salt stratifications performing the machine learning approach over seismic attributes, 16th International Congress of the Brazilian Geophysical Society & EXPOGEF. Rio de Janeiro - RJ, Brazil.

[21] Mohriak, W.U., Szatmari, P., Anjos, S., 2012. Salt: geology and tectonics of selected brazilian basins in their global context. Geological Society, London, Special Publications 363, 131–158. URL: https://sp.lyellcollection.org/content/363/1/131, doi:10.1144/SP363.7, arXiv:https://sp.lyellcollection.org/content/363/1/131.full.pdf.

[22] Moqbel, A., Wang, Y., 2011. Convolutional neural networks for automated seismic interpretation. Journal of Geophysics and Engineering. 8, 592.

[23] Oliveira, L.C., Fernandes, L.F., Maul, A.R., Rosseto, J.A., Gonzalez, M., Gonzalez, G., 2015. Geological Velocity Approach in Order to Obtain a Detailed Velocity Model for the Evaporitic Section at Santos Basin , 1374–1377doi:10.1190/sbgf2015-273.

[24] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

[25] Sayago, J., et al, 2012. Characterization of a deeply buried paleokarst terrain in the loppa high using core data and multiattribute seismic fades classification. AAPG Bulletin 96, 1843–1866.

[26] Waldeland, A., Jensen, A., Gelius, L., Solberg, A., 2018. Convolutional neural networks for automated seismic interpretation. The Leading Edge. 7, 529–537.

ANEXO B – Artigo apresentado no 16º Congresso Internacional da Sociedade Brasileira de Geofísica

Identifying salt stratifications performing the machine learning approach over seismic attributes

Flavio Costa de Mesquita(GISIS/UFF), Marco Antonio Cetale Santos(GISIS/UFF), Alexandre Rodrigo Maul(Petrobras/UFF) and Alex Laier Bordignon(UFF)

Abstract

To build a seismic image we need to process the information from rock interfaces reflections. These reflections occurs as function of the impedance properties differences among the rocks, which is calculated as a combination of density and compressional velocity (inverse of slowness) measurements. Halite, usually the most abundant mineral on the section so-called salt, has an average density about 2,14 g/cm3 and compressional velocity in the order of 4,500 m/s. In terms of seismic studies, until the extreme recent time, the evaporite section was considered to be approximately constant, and reflecting the properties of halite. However, with the evolution of seismic migration algorithms and the computational capacity, it was perceived the need to make the salt section less homogeneous, since the evaporites formation (the evaporation process) occurs in stages, according to specific evaporation rates , generating the observed layering, still also denominated as stratifications, "enigmatic" reflectors/structures. In order to overcome this problem, we need better-elaborated velocity models that take into account these stratifications. In this work we used images of seismic sections and two types of seismic attributes, which will serve as input for a selected Machine Learning algorithm. Thereby allowing identifying from these 2D sections the stratifications within the salt layer allowing the characterization of the corresponding salt type.

Introduction

The field-reservoir under study is located in the central portion of the Santos Basin, about 180 km off the coast of the municipality of Rio de Janeiro in a water depth of approximately 1,900 m depth. The reservoirs of this field are situated between 5,000 and 6,000 m below the sea level and under a layer of salt, the Ariri Formation, which can range from a few hundred meters to over 2,000 m. We do know that this formation is not a homogeneous one, and that it is composed of different types of stratified minerals, the (evaporites). Usually, in the exploratory wells, the log registering of this layer of evaporites is carried out because these are "unknown" areas. In development wells, in general, the logs are no longer acquired within this layer mainly because the project economy. Depending on the complexity of the field and the saline structuring, there are variations in these "strata" and in their thickness, Oliveira et al. (2015), indicated an inverse correlation between "salt thickness" and "salt velocity variation". There are many types of evaporite minerals within the evaporitic section in the Santos and Campos basins, the most common being halite, anhydrite, gypsum, carnallite, tachyhydrite, sylvite. Studies carried out in log analysis show that not all these types of minerals will be seismically detectable by amplitude (Gobatto et al., 2016). Thus, to facilitate the strata identification, the evaporitic minerals in the salt section were grouped into three major facies: halite, high velocity salts (anhydrite and gypsite) and low velocity salts (carnallite, sylvite and tachyhydrite) as per emphasized in Maul et al.(2018). Table 1 presents the mineral grouping as proposed by Maul et al. (2018), indicating the chemical formula of each mineral as well as their acoustic properties.

GROUP	MINERAL	COMPOSITION	DENSITY (g/cm3)	COMPRESSIONAL VELOCITY (m/s)
LVS	Tachyhydrite Carnallite Sylvite	CaMg2Cl6.12(H2O) KMgCl3.6(H2O) KCl	1.57 1.66 1.86	3,300 3,910 3,910
BACKGROUND	Halite	NaCl	2.10	4,550
HVS	Gypsum Anhydrite	CaSO4.2(H2O) CaSO4	2.35 2.98	5,810 6,100

Table 1: Mineral groups and respective properties, average values compiled by Maul et al. (2018) covering more than 200 well in Santos Basin. Low Velocity Salts (LVS), Halite (Background) and High Velocity Salts (HVS).

Maul et al. (2018) based on the methodology proposed by Amaral et al. (2015), compiled the information from more than 200 wells in Santos Basin, showing the halite predominance, over 80% of occurrence. This percentage explains why it is considered as the background class and the main reason the seismic processing starts from the halite velocity property for the entire salt section at the tasks the velocity is needed. The remaining 20% of mineral occurrence can be split as:

- Anhydrite, gypsum, with high values of density and compressional velocity (compared to Halite), facies with low solubility. The basal Anhydrite represents the main seal of reservoirs in the Santos Basin.
- Tachyhydrite, carnallite, and sylvite, whose densities and compressional velocities are smaller than the

groups described above. They are also more soluble, a factor that makes drilling wells more difficult in this basin.

- Other studies also considered halite, the background mineral in this section, representing 75-90% of occurrences (Yamamoto et al., 2016, Gobatto et al., 2016).The proportion of salt varies for each well and its magnitude is related to the geological conditions where the well was drilled. In the "salt walls" Halite normally represents more than 90% and in the mini basins the proportion of salt with high and low velocities varies between 10 and 20% each, consequently decreasing the halite content.

These differences in density and compressional velocity create the clear reflections within the "saline section". These seismic responses have received several denominations, such as: Enigmatic Reflector (Mohriak et al., 2004); Enigmatic Structure (Jackson et al., 2015); Stratification (Maul et al., 2015)). Ji et al. (2011) already indicates that even the insertion of randomly heterogeneities into the salt (the so-called "dirty salt") already assists in the production of better seismic images by their selected migration algorithms.

Materials and Methods

To date, Machine Learning has been used in the seismic scale to predict geological structures, stratigraphy, and rock and fluid properties, usually through seismic interpretation and inversion. In order to do so, fully convolutional deep networks have been used in the area of fault interpretation (Long et al., 2015), 3D convolutional neural networks (Waldeland et al., 2018) and deep encoder-decoder networks for stratigraphic interpretation (Badrinarayanan et al., 2015). These techniques classify a 3D Post-Stack data set based on 3D sub-cubes or 2D sections, and require a relatively low number of labels. Interpretation in seismic images has long used texture attributes to better identify and highlight areas of interest. These can be seen as feature maps in seismic texture. For salt, it can be noted that the texture in the salt masks are quite chaotic, where the surrounding seismic is more "striped". However this is not always true, as we can see in the Santos Basin, where the salt layer is highly stratified.

In this work we use an unsupervised classification algorithm, where the classes we wish to find are the 3 evaporite groups, according to the nomenclature mentioned in Maul et al., (2018): LVS, Halite and HVS.

The seismic data used here is a piece of a Pre-Stack Depth Migration type (PSDM). From this data was extracted an ***in-line*** in SEG-Y and the amplitude response was used as one of the inputs of the algorithm.

Another seismic attribute used was the Relative Acoustic Impedance (RAI), which is a good seismic attribute for the quantitative analysis of beddings (especially in thin strata such as in our case), due to its low amount of low frequency content. Samples with lower seismic amplitude have RAI values related to LVS; samples with higher seismic amplitude have RAI values related to HVS; the remainder can be characterized as halite. To calculate the attribute we use the amplitude response in the extracted ***in-line*** and used as the second input for the algorithm.

The third seismic attribute entry was the TECVA (Amplitudes Volume Technique) type. This attribute aims at the generation of amplitude maps and vertical and horizontal seismic sections that reflect, as far as possible, subsurface geology. Where the knowledge of geology is very dependent on seismic information, it is necessary the details of imaging at the boundaries between the seismic sequences or their inner layers (Bulhões and Amorim, 2005). The Elementary SeismoLayer is the rock layer of smaller thickness that the seismic data can solve and it is defined as the key element of weighting for the calculation and obtaining of the seismic data with the TECVA.

In our case the salt stratifications can be of the magnitude of only a few meters, way below the seismic resolution. Therefore, we need seismic attributes that focus on the high frequencies such as RAI and TECVA combined, in order to better define the stratifications within the salt. A comparison of the three can be seen in figure 1.

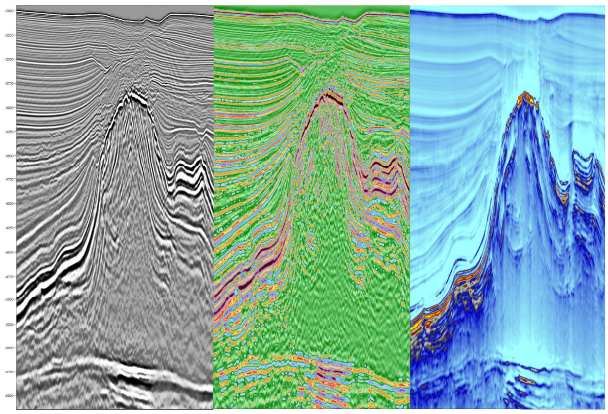

Figure 1: Images used as input: on the left, seismic amplitude section; on the center, relative acoustic impedance; on the right, the TECVA.

Clusterizing

In our case, because it is an image with many features, we used an unsupervised classification that separates the pixels by classes in clusters, for this a sliding window was used on the image as in figure 2, as per defended by Yang et al. (2016).

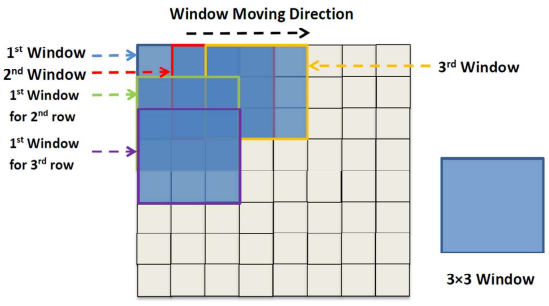

Figure 2: Sliding window. Source: Yang et al. (2016).

For the clusters classification , we use the algorithm t-SNE (van der Maaten and Hinton, 2008). It is a nonlinear dimensionality reduction technique suitable for use with

high-dimensional data for visualization in a small space, as two or three dimensions. Specifically, it models each high-dimensional object by a two-dimensional or three-dimensional point such that similar objects are modeled by nearby points and distant points model different objects with high probability. The resulting clustering can be seen in figure 3.

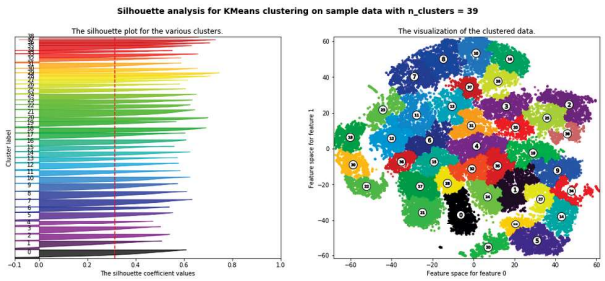

Figure 3: Pixels classified into clusters, each representing a class found in the input images

t-SNE is based on probability distributions with random walk in neighborhood graphs to find the structure within the data. Given a set of *N* high-dimensional objects $\mathbf{x}_1, ..., \mathbf{x}_N$, t-SNE first computes probabilities p_{ij} that are proportional to the similarity of objects $\mathbf{x}_i$ and $\mathbf{x}_j$, as follows:

$$p_{j|i} = \frac{\exp(-\mathbf{x}_i - \mathbf{x}_j^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\mathbf{x}_i - \mathbf{x}_k^2/2\sigma_i^2)}, \quad (1)$$

As van der Maaten and Hinton (2008) explained: "The similarity of datapoint $x_j x_j$ to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i."

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad (2)$$

Moreover, the probabilities with *i = j* are set to zero : $p_{ij} = 0$

The bandwidth of the Gaussian kernels σ_i, is set in such a way that the perplexity of the conditional distribution equals a predefined perplexity using the bisection method. As a result, the bandwidth is adapted to the density of the data: smaller values of σ_i are used in denser parts of the data space.

Since the Gaussian kernel uses the Euclidean distance $x_i - x_j$, it is affected by the curse of dimensionality, and in high dimensional data when distances lose the ability to discriminate, the p_{ij} become too similar (asymptotically, they would converge to a constant). Schubert and Gertz (2017) proposed to adjust the distances with a power transform, based on the intrinsic dimension of each point, to alleviate this.

t-SNE aims to learn a *d*-dimensional map $y_1, \ldots, y_N$ (with $y_i \in R^d$) that reflects the similarities p_{ij} as well as possible. To this end, it measures similarities q_{ij} between two points in the map y_i and y_j, using a very similar approach. Specifically, q_{ij} is defined as:

$$q_{ij} = \frac{(1 + \mathbf{y}_i - \mathbf{y}_j^2)^{-1}}{\sum_{k \neq l} (1 + \mathbf{y}_k - \mathbf{y}_l^2)^{-1}} \quad (3)$$

Herein a heavy-tailed Student's t-distribution (with one-degree of freedom, which is the same as a Cauchy distribution) is used to measure similarities between low-dimensional points in order to allow dissimilar objects to be modeled far apart in the map. Note that also in this case we set $q_{ii} = 0$.

The locations of the points y_i in the map are determined by minimizing the (non-symmetric) Kullback–Leibler divergence of the distribution *Q* from the distribution *P*, that is:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (4)$$

The minimization of the Kullback–Leibler divergence with respect to the points y_i is performed using descending gradient. The result of this optimization is a map that reflects the similarities between the high-dimensional inputs.

Classifying

After finding all of the classes for the input images in separate clusters, we saved the resulting cluster labels. Then we train a classifier using these labels as a target variable and use it for classifying a new image. At this classification, we use two methods, K-Means centers and cKDTree in order to compare.

- *K-Means classification is a method of vector quantization,it aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The center points are vectors of the same length as each data point vector and are the "X's" in figure 3. Each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it.*

- *cKDTree provides an index into a set of k-dimensional points which can be used to rapidly look up the nearest neighbors of any point.The used algorithm is described in Maneewongvatana and Mount (1999). The general idea is that the KDTree is a binary tree in which every leaf node is a k-dimensional point. Every non-leaf node can be thought of as implicitly generating a splitting hyperplane that divides the space into two parts, known as half-spaces. Points to the left of this hyperplane are represented by the left subtree of that node and points to the right of the hyperplane are represented by the right subtree. The hyperplane direction is chosen in the following way: every node in the tree is associated with one of the k dimensions, with the hyperplane perpendicular to that dimension's axis. So, if for a particular split the "x" axis is chosen, all points in the subtree with a smaller "x" value than the node will appear in the left subtree and all points with larger "x" value will be in the right subtree. In such a case, the hyperplane would be set by the x-value of the point, and its normal would be the unit x-axis, (Bentley, 1975).*

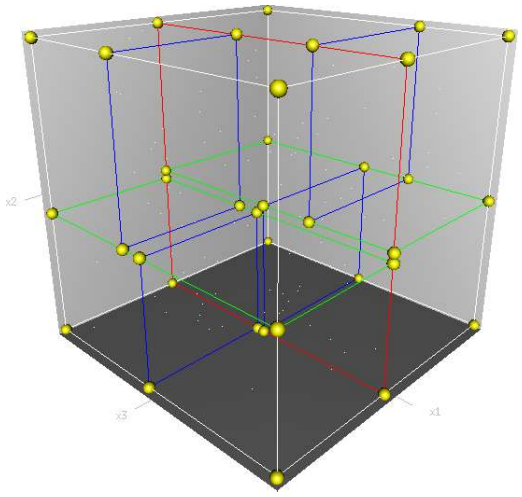

Figure 4: A 3-dimensional cKDTree. The first split (the red vertical plane) cuts the root cell (white) into two subcells, each of which is then split (by the green horizontal planes) into two subcells. Finally, four cells are split (by the four blue vertical planes) into two subcells. Since there is no more splitting, the final eight are called leaf cells. Source: http://www.stat.purdue.edu/ btyner/packages.html

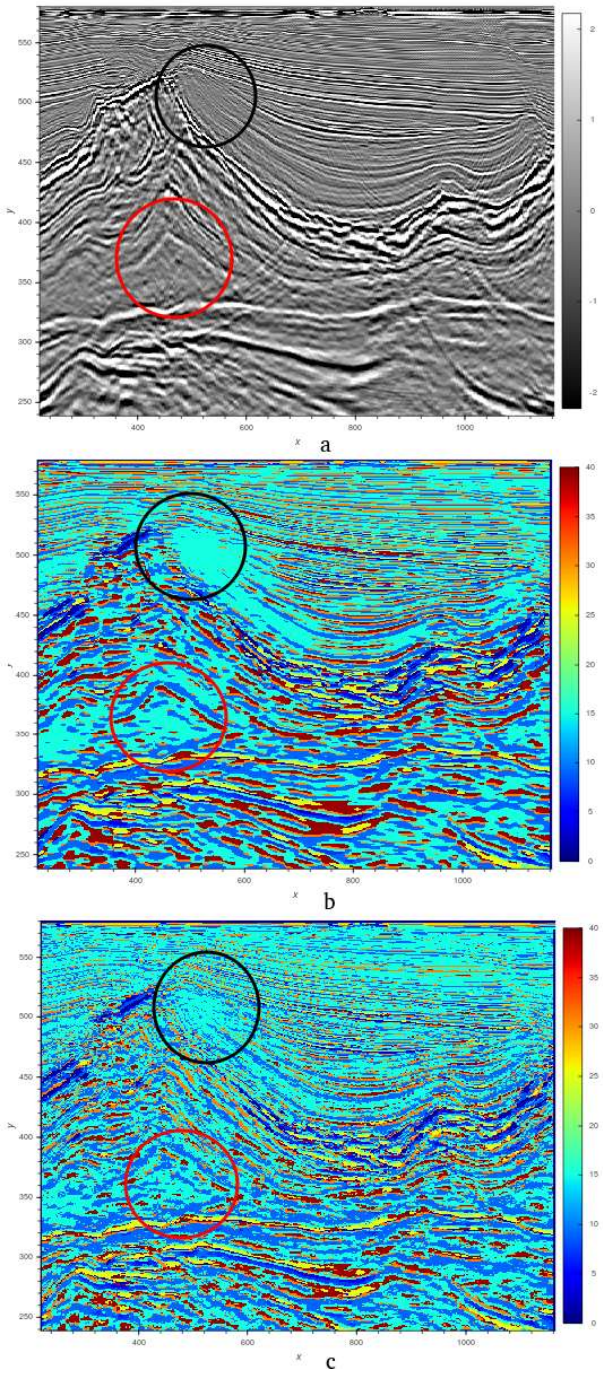

Figure 5: In (a) we see a seismic section image for classification using the cluster labels found; (b) we see the resulting classification using K-Means algorithm; (c) we see the resulting classification using cKDTree algorithm.

Results

From what we could observe until this stage of this research, the t-SNE algorithm is the proper one for this kind of work, which involves the information clusterization, better identifying the image features in a timely manner and a small error of only 2.66% after 1,000 iterations. From that, we used the cluster labels generated to classify another image with two different algorithms, K-Means and cKDTree.

When we compare the resultant images, we clearly can see that cKDTRee was better to classify the image not only on the salt region but also in its surrounds. As we can see on figure 5 (a) a thinly laminated sequence outside the salt, inside the black circle, posed a challenge to imaging. It is due to the low seismic resolution, which is limited to 25 m, causing the image to be blurred and with small definition on this region. Beyond this, some parts of the salt tend to show a homogeneous behavior, as shown inside the red circle.

After using the K-Means classification, figure 5 (b) it had a small improvement. However, it still showed that blurred behavior of the original image.

When we used the cKDTree classifier, figure 5 (c), we could finally see some features we could not see before, on the original image. Some regions of the salt started to present a clear stratification showing different classes, colors, where it was previously more homogeneous . The blurred region outside the salt started to show the laminations we could not see before.

Discussion and Conclusion

From what we could learn on this work, the Machine Learning technique combined with the right seismic attributes can be applied to improve image and generate a more accurate model to seismic processing in regions where salt stratification is known as a problem. Not only on salt but also on regions of thin laminations where the layers are smaller than the seismic resolution, the correct attributes, in this case RAI and TECVA , can aid to realize features of the subsurface we could not see before.

As a future work, we plan to extend this research to a 3D

dimension data and test other relevant seismic attributes to see if it can improve the image not only on the salt region but also on thinly laminated structures, providing better and more accurate velocity models.

Acknowledgments

We would like to acknowledge the ANP for providing the dataset for this work, the Euclides da Cunha Foundation (FEC) for the scholarship, Schlumberger for the academic license of Petrel software and GISIS for the academic suppport and infrastructure.

References

BADRINARAYANAN, V., KENDALL, A. AND CIPOLLA, R., [2015]. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint, arXiv:1511.00561.

BENTLEY, J. L., [1975]. "Multidimensional binary search trees used for associative searching". Communications of the ACM. 18 (9): 509. doi:10.1145/361002.361007.

BULHÕES, E. AND AMORIM, W., [2005] – Princípio da Camada Elementar e sua aplicação à Tecnica Volume de Amplitudes (TECVA). Ninth International Congress of the Brazilian Geophysical Society

GOBATTO, F., MAUL, A., FALCÃO, L., TEIXEIRA, L., BOECHAT, J.B., GONZÁLEZ, M. AND GONZÁLEZ, G., [2016]. Refining Velocity Model within the Salt Section in Santos Basin: an Innovative Workflow to include the Existing Stratification.

LONG, J., SHELHAMER, E. AND DARRELL, T., [2015]. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3431-3440. arXiv:1411.4038.

MANEEWONGVATANA S. AND MOUNT D. M., [1999]. It's okay to be skinny, if your friends are fat. 4th Annual CGC Workshop on Computational Geometry.

MAUL, A., CETALE, M., AND GUIZAN, C., [2018] - Few Considerations, warnings and benefits for the E&P industry when incorporating stratifications inside salt sections. Revista Brasileira de Geofísica, Vol. 36(4), 2018 DOI: http://dx.doi.org/10.22564/rbgf.v36i4.1981.

OLIVEIRA, L., FALCÃO, L., MAUL, A., ROSSETO, J., GONZÁLEZ, M. AND GONZÁLEZ, G., [2015]. Geological Velocity Approach in Order to Obtain a Detailed Velocity Model for the Evaporitic Section, Santos Basin.

SCHUBERT, E. AND GERTZ, M., [2017]. Intrinsic t-Stochastic Neighbor Embedding for Visualization and Outlier Detection. SISAP 2017 – 10th International Conference on Similarity Search and Applications. pp. 188–203.

VAN DER MAATEN, L.J.P. AND HINTON, G.E., [2008]. Visualizing Data Using t-SNE . Journal of Machine Learning Research. 9: 2579–2605.

WALDELAND, A.U., JENSEN, A.C., GELIUS, L.J. AND SOLBERG, A.H.S., [2018]. Convolutional neural networks for automated seismic interpretation. The Leading Edge, 37(7), 529-537.

YAMAMOTO, T., MAUL, A., MARTINI, A., BORN, E. AND GONZALEZ, M., [2017]. Evaporitic Section Characterization Using Inversion and Bayesian Classification

YANG PO, GORDON C., FENG D., VALERIU C., DAVID W., BAOQUAN L., JOS BTM, ROERDINK AND ZHIKUN D., [2016] - GSWO: A programming model for GPU-enabled parallelization of sliding window operations in image processing. Signal Processing: Image Communication Volume 47, September 2016, Pages 332-345