



COPPE/UFRJ

IDENTIFICAÇÃO DE REGRAS DE ASSOCIAÇÃO INTERESSANTES EM UMA
BASE DE DADOS SOBRE EXSUDAÇÕES DE ÓLEO NO GOLFO DO MÉXICO

Aretha Felix Thomaz da Silva

Dissertação de Mestrado apresentada ao
Programa de Pós-graduação em Engenharia Civil,
COPPE, da Universidade Federal do Rio de
Janeiro, como parte dos requisitos necessários à
obtenção do título de Mestre em Engenharia Civil.

Orientadores: Luiz Landau
Fernando Pellon de Miranda

Rio de Janeiro
Outubro de 2008

IDENTIFICAÇÃO DE REGRAS DE ASSOCIAÇÃO INTERESSANTES EM UMA BASE
DE DADOS SOBRE EXSUDAÇÕES DE ÓLEO NO GOLFO DO MÉXICO

Aretha Felix Thomaz da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ
COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS
NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM
ENGENHARIA CIVIL.

Aprovada por:

Prof. Luiz Landau, D.Sc.

Prof. Fernando Pellon de Miranda, Ph.D.

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

Prof. Carlos Siqueira Bandeira de Mello, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

OUTUBRO DE 2008

Silva, Aretha Felix Thomaz da

Identificação de regras de associação interessantes em uma base de dados sobre exsudações de óleo no Golfo do México / Aretha Felix Thomaz da Silva. - Rio de Janeiro: UFRJ/COPPE, 2008.

XIV, 91 p.: il.; 29,7 cm.

Orientadores: Luiz Landau e Fernando Pellon de Miranda.

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2008.

Referencias Bibliográficas: p. 86-91.

1. Sensoriamento remoto. 2. Detecção de exsudações de óleo. 3. Regras de associação. 4. Golfo do México. 5. Medidas de interesse objetivas. I. Landau, Luiz *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

O Anjo do Senhor apareceu-lhe numa chama
que saía do meio dum sarça. Moisés olhava:
a sarça ardia, mas não se consumia.

Êxodo (capítulo 3, versículo 2).

AGRADECIMENTOS

Primeiramente, agradeço a Deus por tudo que representa em minha vida. Sou grata também a todos que direta ou indiretamente colaboraram na realização deste estudo.

A meus pais, Edson e Vera, a meu namorado e companheiro Deonel e a minha irmã Anastácia, pelo amor incondicional, carinho e respeito sempre demonstrados e, especialmente, pelo apoio e compreensão nos momentos de ausência.

Ao professor Dr. Fernando Pellon de Miranda (LABSAR/COPPE/UFRJ), pela orientação, constante disponibilidade e condução técnica durante a confecção desta dissertação, muito obrigado pelo seu apoio, incentivo e paciência. Agradeço a sorte que tive em tê-lo como orientador.

Ao professor Dr. Luiz Landau (LAMCE/COPPE/UFRJ), pela orientação e por todo apoio institucional, financeiro e tecnológico.

À minha Coordenadora Antonia, pelo incentivo constante, pelas críticas e sugestões, foi muito mais do que uma colega de trabalho, uma verdadeira amiga.

Ao Laboratório de Sensoriamento Remoto por Radar Aplicado à Indústria do Petróleo (LABSAR/COPPE/UFRJ), por toda infra-estrutura e base de dados disponibilizada. Agradeço a todos os funcionários e amigos do LABSAR, que também contribuíram para a finalização desta pesquisa.

Aos doutores Nelson Francisco Favilla Ebecken (PESC/COPPE/UFRJ) e Carlos Siqueira Bandeira de Mello (CENPES/PETROBRAS), por terem aceitado integrar a banca examinadora.

Agradeço à Universidade Federal do Rio de Janeiro (UFRJ), pela oportunidade, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão de bolsa de estudos durante parte da pesquisa.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

IDENTIFICAÇÃO DE REGRAS DE ASSOCIAÇÃO INTERESSANTES EM UMA
BASE DE DADOS SOBRE EXSUDAÇÕES DE ÓLEO NO GOLFO DO MÉXICO

Aretha Felix Thomaz da Silva

Outubro/2008

Orientadores: Luiz Landau

Fernando Pellon de Miranda

Programa: Engenharia Civil

Esta dissertação tem por objetivo a identificação de regras de associação interessantes em uma base de dados de exsudações de óleo, obtida a partir da interpretação de imagens RADARSAT-1 do Golfo do México, Baía de Campeche. Associação é uma tarefa de mineração de dados que tem sido amplamente utilizada para representação de conhecimento implícito. Porém o grande número de regras que podem ser geradas dificulta o reconhecimento de conhecimento relevante ao usuário. Com propósito de minimizar esse problema, a metodologia proposta é composta por duas etapas principais. Na primeira, foram geradas as regras com o uso do aplicativo CBA; na segunda, foi realizado o pós-processamento das regras com uso de medidas de interesse objetivas e subjetivas de avaliação do conhecimento. Essas medidas foram utilizadas para aquilatar a qualidade das regras e selecionar algumas potencialmente interessantes, conforme a opinião de um especialista na detecção de exsudações de óleo por satélite. Assim, foram identificadas, dentre as regras geradas, aquelas que são realmente interessantes e inovadoras de acordo com o conhecimento obtido durante a avaliação. Como resultado, foi possível estabelecer uma seqüência de procedimentos com potencial para viabilizar o uso operacional da metodologia proposta, assim como definir um conjunto de padrões ambientais e de localização úteis na detecção de exsudações de óleo na área de estudo.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

IDENTIFICATION OF INTERESTING ASSOCIATION RULES IN A
DATABASE ABOUT OIL SEEPS IN THE GULF OF MEXICO

Aretha Felix Thomaz da Silva

October/2008

Advisors: Luiz Landau
Fernando Pellon de Miranda

Department: Civil Engineering

The objective of this dissertation is to identify interesting associations rules in a database about oil seeps generated as a result of RADARSAT-1 image interpretation in the Gulf of Mexico, Campeche Bay. Association is a data mining procedure that has been often used to represent intrinsic knowledge. However, the excessive number of rules that can be obtained this way makes difficult the recognition of knowledge that is relevant to the user. With the objective of minimizing such a problem, the methodology herein proposed is constituted of two main steps. In the first one, association rules are generated using the algorithm CBA; in the second one, post-processing of rules is carried out with the aid of objective and subjective interest measures for knowledge evaluation. These measures are used to form an idea of the quality of rules and to select those potentially interesting, according to the opinion of an expert in satellite oil seep detection. Therefore, the association rules that are indeed interesting and innovative are identified according to knowledge acquired during the evaluation process. As a result, it is possible to establish a sequence of procedures with potential to make feasible the operational use of the proposed methodology, as well as to define environmental and shape patterns relevant to oil seep detection in the study area.

SUMÁRIO

AGRADECIMENTOS	v
RESUMO	vi
ABSTRACT	vii
LISTA DE FIGURAS.....	x
LISTA DE TABELAS.....	xiii
CAPÍTULO 1 – INTRODUÇÃO.....	1
1.1 – Motivação para realização da pesquisa	4
1.2 – Objetivos.....	5
1.3 – Localização da área de estudo.....	6
1.4 – Visão geral da metodologia proposta	8
CAPÍTULO 2 – DETECÇÃO DE EXSUDAÇÕES DE ÓLEO NA SUPERFÍCIE DO MAR UTILIZANDO SENSORIAMENTO REMOTO POR RADAR.....	10
2.1 – Conceitos básicos.....	10
2.2 – Rugosidade superficial dos alvos	14
2.3 – O efeito de redução do espalhamento Bragg pela presença de óleo.....	16
2.4 – O sistema RADARSAT-1.....	17
2.4.1 – Características gerais	18
2.4.2 – Modos de imageamento	18
2.4.3 – Órbitas utilizadas na aquisição de dados	21
2.5 – Emprego de dados meteo-oceanográficos como suporte à interpretação de exsudações de óleo no mar.....	22
2.5.1 – Mapa de temperatura da superfície do mar (TSM).....	22
2.5.2 – Mapa de temperatura do topo de nuvens (TTN)	24
2.5.3 – Mapa de intensidade do campo de vento.....	26

2.5.4 – Mapa de altura significativa de ondas	28
2.5.5 – Mapa de concentração de clorofila-a.....	30
CAPÍTULO 3 – O PROJETO DE MONITORAMENTO SISTEMÁTICO POR SATÉLITE DA EXSUDAÇÃO PETROLÍFERA DE CANTARELL	33
3.1 – Aquisição e processamento digital das imagens RADARSAT-1	37
CAPÍTULO 4 – DESCOBERTA DE CONHECIMENTO E REGRAS DE ASSOCIAÇÃO ..	40
4.1 – O processo de descoberta de conhecimento	40
4.2 – Regras de associação	44
4.2.1 – Medidas de interesse objetivas	48
4.2.2 – Medidas de interesse subjetivas.....	50
4.2.3 – Aplicativo CBA na extração de regras	50
CAPÍTULO 5 – RESULTADOS: PROCESSO DE KDD UTILIZANDO REGRAS DE ASSOCIAÇÃO	52
5.1 – Pré-processamento da base de dados.....	53
5.1.1 – Níveis Taxonômicos e análise exploratória das exsudações de óleo	54
5.2 – Extração de padrões (CBA).....	67
5.3 – Pós-processamento e análise das relações interessantes obtidas como resultado	71
5.4 – Análise do potencial de uso do novo conhecimento.....	82
CAPÍTULO 6 – CONCLUSÕES E RECOMENDAÇÕES.....	84
REFERÊNCIAS BIBLIOGRÁFICAS	86

LISTA DE FIGURAS

Figura 1.1 – Área de estudo, com níveis batimétricos e localização da exsudação petrolífera de Cantarell (Fonte: Miranda <i>et. al</i> , 2004).....	6
Figura 1.2 – Vista aérea da exsudação de óleo do Campo de Cantarell. Como fator de escala, notar a plataforma petrolífera no canto superior direito da foto (Fonte: Mendoza <i>et al.</i> , 2003).	7
Figura 1.3 – Os cinco componentes do processo de KDD incluídos na metodologia proposta (modificado de Rezende <i>et al.</i> , 2003).....	8
Figura 2.1 – Diagrama expandido do espectro eletromagnético em relação à transmitância atmosférica (modificado de Sabins, 1997). H ₂ O, CO ₂ e O ₃ referem-se a gases na atmosfera que absorvem a energia eletromagnética.....	11
Figura 2.2 – Visão esquemática do pulso do radar (modificado de Sabins, 1997).	13
Figura 2.3 – Geometria básica de imageamento do radar de abertura sintética – SAR (modificado de Sabins, 1997).	14
Figura 2.4 – Relação entre os ângulos de incidência, visada e depressão em uma superfície plana (modificado de Sabins,1997).....	14
Figura 2.5 – Tipos básicos de interação do pulso de radar com a superfície do mar (modificado de Sabins,1997).	17
Figura 2.6 – Geometria de visada e modos de operação do satélite RADARSAT-1 (modificado de RADARSAT <i>International</i> , 1996).....	19
Figura 2.7 – Configuração das órbitas descendente (com visada para oeste) e ascendente (com visada para leste) do RADARSAT-1 (modificado de RADARSAT <i>International</i> , 1996).....	21
Figura 2.8 – Mapa de temperatura da superfície do mar, obtido a partir do sensor AVHRR a bordo do satélite NOAA-17, em 27 de julho de 2006. O retângulo em preto está representando o frame da imagem SAR da Figura 2.9 (Roriz, 2006).	23
Figura 2.9 – Imagem do satélite RADARSAT-1, no modo SCN1 (órbita ascendente), adquirida no Golfo do México, em 27 de julho de 2006 (Roriz, 2006).....	24
Figura 2.10 – Mapa de temperatura do topo de nuvem obtido a partir de uma imagem do sensor AVHRR, a bordo do satélite NOAA-15, adquirida em 05 de julho de 2004. O	

retângulo em preto está representando o frame da imagem SAR da Figura 2.11 (Roriz, 2006).....	25
Figura 2.11 – Imagem do satélite RADARSAT-1, no modo SCN1 (órbita ascendente), adquirida no Golfo do México, em 06 de julho de 2004 (Roriz, 2006).....	26
Figura 2.12 – Mapa de intensidade do campo de vento obtido pelo sensor SeaWinds, a bordo do satélite QuickSCAT, em 27 de janeiro de 2004. As setas estão indicando as intensidades e as direções calculadas do vento; o retângulo em preto está representando o frame da imagem SAR da Figura 2.13 (Roriz, 2006).	27
Figura 2.13 – Imagem do satélite RADARSAT-1, no modo SCN1 (órbita ascendente), adquirida em 27 de janeiro de 2004 (Roriz, 2006).	28
Figura 2.14 – Mapa da altura significativa de ondas, obtido pelo altímetro a bordo do satélite TOPEX-POSEIDON, em 13 de março de 2005. O retângulo em preto está representando o frame da imagem SAR da Figura 2.15 (Roriz, 2006).	29
Figura 2.15 – Imagem do satélite RADARSAT-1 no modo SCN2 (órbita ascendente), adquirida em 13 de março de 2005 (Roriz, 2006).	30
Figura 2.16 – Mapa de concentração de clorofila-a, confeccionado a partir do processamento de uma imagem do satélite MODIS, obtida em 14 de julho de 2004. O retângulo em preto está representando o frame da imagem SAR da Figura 2.17 (Roriz, 2006).....	31
Figura 2.17 – Imagem do satélite RADARSAT-1 no modo SCN1 (órbita descendente), adquirida em 14 de julho de 2004 (Roriz, 2006).....	32
Figura 3.1 – Mapa temático das exsudações de Cantarell, confeccionado a partir da composição de 9 (nove) imagens do satélite RADARSAT-1, no modo W1, adquiridas durante o ano de 2000 (Miranda <i>et al.</i> , 2004). As curvas batimétricas estão expressas em metros.....	35
Figura 3.2 – Mapa temático das exsudações de Cantarell, confeccionado a partir da composição de 19 (nove) imagens do satélite RADARSAT-1, nos modos de imageamento SCN1, W1 e W2, adquiridas durante o ano de 2001 (Miranda <i>et al.</i> , 2004). As curvas batimétricas estão expressas em metros.	35
Figura 3.3 – (A) imagem RADARSAT-1 no modo de operação W1, adquirida em julho de 2000; (B) polígono resultante da classificação USTC para a exsudação de Cantarell (Mendoza <i>et al.</i> , 2003).	36

Figura 3.4 – Esquema das principais etapas propostas e realizadas pelo LABSAR para aquisição e processamento digital das imagens RADARSAT-1.	37
Figura 4.1 – Etapas do processo de Descoberta de Conhecimento em Base de Dados (Rezende <i>et. al.</i> , 2003).	42
Figura 4.2 – Tela inicial do aplicativo CBA utilizado para extração das regras de associação (Liu <i>et al.</i> , 1998).	51
Figura 5.1 – Representação do nível taxonômico superior das exsudações de óleo.	55
Figura 5.2 – Representação geral dos níveis taxonômicos das exsudações de óleo na Baía de Campeche, Golfo do México.	56
Figura 5.3 – Representação dos sub-níveis do nível taxonômico superior Localização. ...	57
Figura 5.4 – Gráfico de distribuição da quantidade de polígonos interpretados como exsudação nas imagens RADARSAT-1.	58
Figura 5.5 – Representação dos sub-níveis do nível taxonômico superior Contexto Temporal.	58
Figura 5.6 – Representação dos sub-níveis do nível taxonômico superior Forma.	59
Figura 5.7 – Representação dos sub-níveis do nível taxonômico superior Batimetria do Centróide do Polígono.	61
Figura 5.8 – Representação dos sub-níveis do nível taxonômico superior Características do Imageamento.	62
Figura 5.9 – Representação dos sub-níveis do nível taxonômico superior Agrupamento Associado.	63
Figura 5.10 – Representação dos sub-níveis do nível taxonômico superior Condições Meteo-oceanograficas.	65
Figura 5.11 – Gráfico do suporte em relação à quantidade de pares de itens.	74
Figura 5.12 – Gráfico da derivada primeira dos pares de itens em relação ao suporte (d Quantidade de pares / d Suporte).	74
Figura 5.13 – Derivada primeira dos pares de itens em relação ao suporte, para valores de suporte maiores que 50%.	75
Figura 5.14 – Exemplo de um polígono, de um total de dezenove, que representa a regra rara “Compactação= alta → Perímetro= baixo”.	81

LISTA DE TABELAS

Tabela 2.1 – Bandas espectrais, intervalos de comprimento de onda e frequência utilizados em sistemas de radar (modificado de Sabins, 1997).	12
Tabela 2.2 – Características do satélite RADARSAT-1 (Fonte: RADARSAT <i>International</i> , 1996).	18
Tabela 2.3 – Características dos diferentes modos de imageamento do satélite RADARSAT-1 (Fonte: RADARSAT <i>International</i> , 1996).	20
Tabela 3.1 – Síntese das fases do projeto de monitoramento de exsudações e derrames operacionais de óleo no Golfo do México, Baía de Campeche.	34
Tabela 4.1 – Resumo das fases e etapas envolvidas pelo processo de KDD.	42
Tabela 5.1 – Níveis taxonômicos superior, médio e inferior das exsudações de óleo na Baía de Campeche, Golfo do México.	55
Tabela 5.2 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Localização, expressa em termos de distribuição no tempo.	57
Tabela 5.3 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Contexto Temporal.	59
Tabela 5.4 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Forma.	60
Tabela 5.5 – Estatística dos sub-níveis do nível taxonômico superior Forma.	60
Tabela 5.6 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Batimetria do Centróide do Polígono.	61
Tabela 5.7 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Características do Imageamento.	62
Tabela 5.8 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Agrupamento Associado.	64
Tabela 5.9 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Condições Meteo-oceanograficas. Os níveis de discretização foram estabelecidos pelo especialista do domínio.	66
Tabela 5.10 – Estatística básica dos sub-níveis numéricos do nível taxonômico superior Condições Meteo-oceanograficas.	67

Tabela 5.11 – Frequência de ocorrência dos subníveis taxonômicos que serviram de atributos de entrada no aplicativo CBA.	67
Tabela 5.12 – Conteúdo do arquivo de entrada do CBA (exsudação.NAMES).	70
Tabela 5.13 – Pares de itens do arquivo de saída do CBA (Extensão *. ITM) com suporte maior ou igual a 0,1%. A linha em vermelho (76%) é ilustrada na Figura 5.13....	72
Tabela 5.14 – Lista de regras de associação intraníveis.....	78
Tabela 5.15 – Lista de regras de associação interníveis.....	79
Tabela 5.16 – Lista com a comparação entre as regras de associação correspondentes aos subníveis Cantarell e Outras do nível taxonômico Localização.....	80
Tabela 5.17 – Lista de regras de associação raras.....	80

CAPÍTULO 1

INTRODUÇÃO

O petróleo é um produto de grande importância mundial, pois é difícil determinar algo que dele não dependa como fonte de energia ou matéria prima. Embora conhecido desde os primórdios da civilização, só passou a ser utilizado comercialmente no século 18, após Edwin Drake, em 1859, perfurar o primeiro poço exploratório. Os produtos derivados a partir de seu refino e processamento petroquímico são: gasolina, gás, óleos combustíveis, óleos lubrificantes, asfalto, parafina, querosene, nafta e plásticos, entre outros.

O óleo e o gás natural são encontrados em superfície e subsuperfície, tanto na terra quanto costa afora. A diferença de densidade entre os hidrocarbonetos e a água do mar permite seu surgimento na superfície do oceano. Os hidrocarbonetos possuem densidade menor que a da água e, por esse motivo, migram verticalmente, a partir do assoalho marinho, formando assim as exsudações (*seepages*). As feições geológicas que permitem a migração do petróleo através da coluna sedimentar são falhas ou discordâncias aflorantes. Uma exsudação, portanto, é definida como um local na superfície terrestre onde hidrocarbonetos líquidos ou gasosos emanam naturalmente e podem ser observados.

Atualmente, não há dúvida no que diz respeito à capacidade do uso das imagens de radar na detecção e monitoramento de diversos eventos marinhos em áreas extensas. O monitoramento com o uso de radar de manchas de óleo (*oil slicks*) na superfície do mar subsidia a exploração petrolífera *offshore*, bem como facilita a compreensão da dinâmica temporal e distribuição espacial (localização) das exsudações.

A presença de óleo provoca a redução da rugosidade na superfície do mar, fenômeno passível de ser detectado por imagens orbitais de radares de abertura sintética (SAR), tais como o RADARSAT-1. Esse fenômeno é visível nas imagens como feições escuras e de textura lisa. Contudo, outros fatores podem produzir feições semelhantes às causadas pelo petróleo, que são classificadas, portanto, como falsos alvos. Essa discriminação é auxiliada pela análise de dados meteo-

oceanográficos complementares, obtidos concomitantemente, tanto quanto possível, à aquisição da imagem RADARSAT-1.

A persistência das manchas de petróleo ao longo do tempo e aproximadamente na mesma localização geográfica valida fortemente a classificação da feição analisada como representativa de exsudação, indicando a existência de rotas ativas de migração (Miranda *et al.*, 2004). Já os vazamentos operacionais, associados à poluição por petróleo, são provenientes de eventos de origem não identificada (derramamento ilegal de óleo - IOD) ou freqüentemente associados a vertimentos a partir de facilidades de produção e transporte *offshore* tais como dutos, navios e plataformas. Os dois últimos aparecem como pontos de forte retorno nas imagens de radar.

O LABSAR (Laboratório de Sensoriamento Remoto por Radar Aplicado à Indústria do Petróleo), situada na COPPE/UFRJ, realiza desde 2002 o monitoramento sistemático de exsudações e derramamentos de petróleo na Baía de Campeche, no sul do Golfo do México, através de imagens do satélite RADARSAT-1. Esse laboratório desenvolveu metodologia própria para a identificação de tais alvos. Primeiramente, as feições das imagens de satélite, classificadas como exsudações ou vazamentos operacionais de óleo, são individualizadas em polígonos através do uso do algoritmo intitulado *Unsupervised Semivariogram Textural Classifier* (USTC) e da avaliação de especialistas. Em seguida, para cada um desses polígonos, é obtido um conjunto de dados referentes à suas características geométricas (forma) e batimétricas, contexto temporal, modo de imageamento RADARSAT-1 e dados meteorológicos obtidos no momento da aquisição da imagem de radar (temperatura da superfície do mar, temperatura de topo de nuvem, velocidade do vento, altura da onda e concentração de clorofila-a). Por fim, esses dados são armazenados em uma base sobre manchas relacionadas a exsudações de óleo (*seepage slicks*), de forma que as linhas representam os polígonos (transações) e as colunas seus atributos (itens). Conseqüentemente, esse monitoramento resulta em uma extensa quantidade de dados que pode esconder conhecimento estratégico para apoio à tomada de decisão.

Nesse contexto, visando à obtenção de informações desconhecidas e relevantes, a presente dissertação contempla o processo de descoberta de conhecimento em base de dados (KDD – *Knowledge Discovery in Databases*). Este se dá por meio de estratégias automatizadas para a análise de grandes bases de dados e tem como objetivo a extração de padrões implícitos, potencialmente úteis e que

acrescentem novos conhecimentos aos já existentes. Assim sendo, a utilização de KDD auxilia a identificação de padrões claros e compreensíveis, de modo a dar suporte a procedimentos de tomada de decisão e de planejamento a médio e longo prazo.

Segundo Fayyad *et al.* (1996a), o processo de KDD possui várias etapas não triviais, que são interativas e iterativas, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis, a partir de grande quantidade de dados. Este processo incorpora as seguintes etapas: pré-processamento, extração de padrões (também conhecido por *Data Mining* - DM) e pós-processamento. Na etapa de pré-processamento, é realizada a preparação dos dados, ou seja, a captação, organização e tratamento das informações. Já na mineração de dados, é realizada a busca por conhecimento útil ao contexto da aplicação. Na etapa de pós-processamento, acontece o tratamento do conhecimento adquirido, objetivando viabilizar sua utilização.

A pesquisa aqui desenvolvida constitui o primeiro exercício de KDD utilizando a base de dados da Baía de Campeche, no qual pretende-se investigar apenas os fenômenos de origem natural. Assim, as transações referentes aos polígonos classificados como vazamento operacional foram desconsideradas, visto que o objetivo do processo de extração de padrões desse estudo envolve somente os dados relativos às exsudações. Vale ressaltar que a referida base de dados não foi produzida com o intuito de ser usada em uma aplicação de KDD. Portanto, a etapa de pré-processamento, onde os dados são tratados e preparados para as etapas subsequentes, foi realizada de maneira criteriosa, pois possui fundamental relevância no processo de descoberta de conhecimento.

A mineração de dados inclui várias técnicas; no entanto, a pesquisa em tela, tem por foco a aplicação da técnica de Regras de Associação. Tal abordagem foi apresentada inicialmente por Agrawal *et al.* (1993) e objetiva descobrir regras que descrevam dependências significativas entre os itens (atributos) das transações que ocorrem de forma simultânea. A utilização de regras como linguagem para representação do conhecimento é de fácil compreensão até mesmo por usuários que não possuem experiência em mineração de dados. Uma regra de associação é uma atividade descritiva que representa uma implicação do tipo “*se antecedente, então conseqüente*”. Isso significa que, quando uma transação contém o(s) item(ns) do antecedente, então provavelmente também contém aquele(s) do conseqüente.

Com a técnica de Regras de Associação, é possível encontrar todas as associações existentes nas transações de uma base de dados, podendo resultar em um grande número de regras, sendo que poucas são realmente interessantes ao usuário. Medidas objetivas, como suporte e confiança, são utilizadas em algoritmos de geração de regras de associação como filtro para a caracterização de regras fortes ou de eventos raros. O suporte pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto o antecedente quanto o conseqüente. Por outro lado, a confiança é a probabilidade de que uma transação satisfaça o conseqüente, dado que ela satisfaz o antecedente. O emprego dessas restrições conduz à redução no volume de regras geradas e, conseqüentemente, na diminuição do tempo de processamento e análise do resultado final. Assim, a alteração desses parâmetros inclui ou exclui novas regras ao conjunto de regras resultantes.

Utilizou-se o aplicativo *Classification Based in Association* (CBA), proposto por Liu *et al.* (1998), para a geração das regras de associação na etapa de mineração de dados. Essa ferramenta apresenta em sua arquitetura interna o algoritmo *Apriori* (Agrawal e Srikant, 1994) e utiliza as medidas objetivas de suporte e confiança no processo de geração das regras. Contudo, por tratar-se de uma tarefa descritiva de mineração de dados, o conjunto de regras gerado pelo CBA deve ser replicado no caso da utilização de outro aplicativo com a mesma finalidade.

Dada a grande quantidade de regras geradas com o CBA, foi utilizada, além do suporte e confiança, uma das principais medidas de interesse objetivos de avaliação do conhecimento encontradas na literatura, denominada *lift*. Tal procedimento visou avaliar a qualidade das regras e diminuir a dificuldade na detecção daquelas realmente úteis para o usuário final.

1.1 – Motivação para realização da pesquisa

Na parte sul do Golfo do México, ocorre uma das maiores províncias petrolíferas do mundo, na qual se inclui a proeminente exsudação de Cantarell. Nesta região, a atividade pesqueira é intensa e o ecossistema predominante são os delicados manguezais. Devido à grande fragilidade ambiental, é de fundamental relevância a realização de estudos que auxiliem na adoção de práticas apropriadas de gerenciamento dos riscos de derramamentos de óleo na zona costeira. Nesse contexto, o LABSAR realiza um projeto de monitoramento na Baía de Campeche, que

faz uso de avançadas técnicas de processamento de imagens aplicadas aos dados RADARSAT-1 (Mendoza *et al.*, 2003). Um dos objetivos desse projeto de monitoramento é a detecção de exsudações de óleo na superfície do mar, notadamente de Cantarell. Tais feições são indicativas da existência de sistemas petrolíferos ativos conectando rotas de migração de óleo ao fundo do mar. O petróleo oriundo das exsudações pode eventualmente atingir o litoral e danificar equipamentos de pesca.

Com o propósito de auxiliar o processo de geração de conhecimento como parte do citado projeto de monitoramento, faz-se necessário o uso de técnicas adequadas sobre a grande massa de dados por ele gerados. Tal procedimento possibilita a obtenção de conhecimento a partir da identificação de padrões nas relações entre os atributos das exsudações de óleo. Assim, a presente dissertação utiliza a tarefa de KDD denominada de regras de associação como suporte ao levantamento do conhecimento embebido na base de dados da Baía de Campeche.

1.2 – Objetivos

O objetivo principal desse estudo é propor uma metodologia para identificação de regras de associação interessantes na base de dados de exsudações da Baía de Campeche, a partir de medidas objetivas e do conhecimento tácito fornecido por um especialista do domínio da aplicação. Pretende-se explicitar relações e/ou dependências morfológicas, ambientais, batimétricas, climáticas e operacionais entre os dados acima citados, que são derivados de imagens do satélite RADARSAT-1 e de outros sistemas orbitais de sensoriamento remoto. Desta forma, será possível contribuir com conhecimentos inéditos e interessantes da região, como subsídio à gestão ambiental.

Os objetivos específicos são os seguintes:

- Consolidar, de maneira estruturada, os conhecimentos objetivos referentes à área de estudo;
- Obter conhecimentos subjetivos sobre o domínio da aplicação;
- Avaliar o resultado do aplicativo CBA;
- Obter uma lista de conhecimentos inéditos e/ou interessantes;
- Analisar o potencial de uso do novo conhecimento.

1.3 – Localização da área de estudo

Os dados utilizados neste estudo foram adquiridos na região da Baía de Campeche, localizada na porção meridional do Golfo do México, em cujo litoral se situa Ciudad *del Carmen* (Figura 1.1). A área se notabiliza pela presença da exsudação de óleo extremamente ativa do campo de Cantarell (o maior do hemisfério ocidental). Outras exsudações são também encontradas na Baía de Campeche, em áreas com maior lâmina d'água.

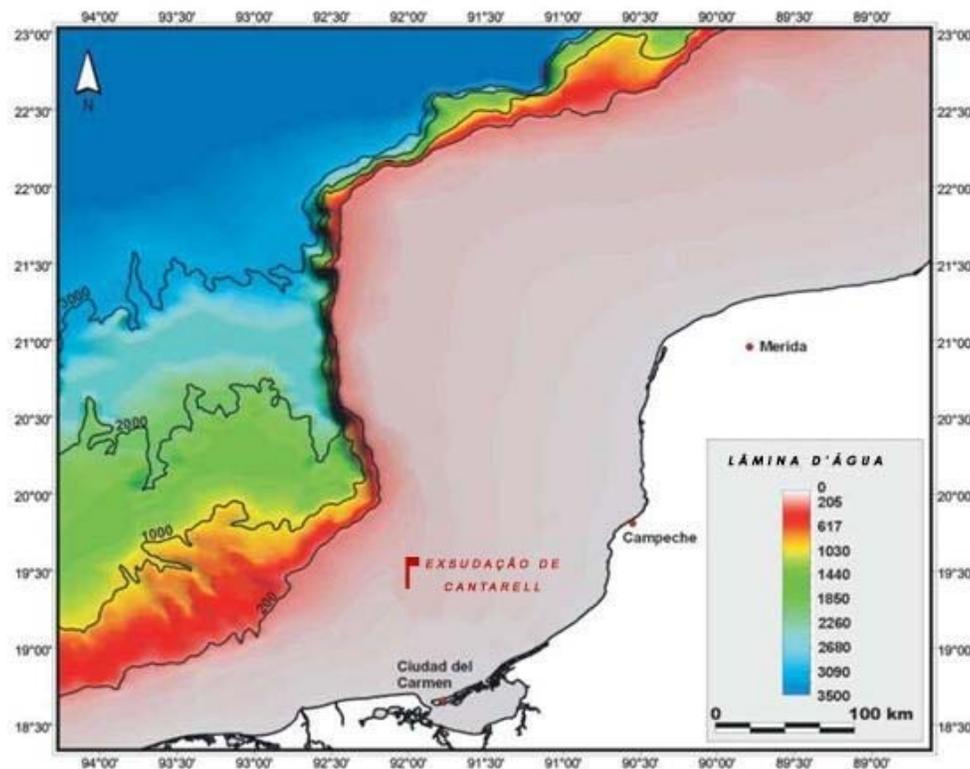


Figura 1.1 – Área de estudo, com níveis batimétricos e localização da exsudação petrolífera de Cantarell (Fonte: Miranda *et. al*, 2004).

Segundo Miranda *et al.* (2004), a produção de óleo no campo de Cantarell é da ordem de 1,9 milhões de barris por dia. Em tal região, as atividades pesqueiras são intensas e bem desenvolvidas e o ecossistema delicado é representado principalmente por manguezais. Assim, o entendimento da dinâmica espacial e temporal das exsudações de óleo nesse ambiente é fundamental para a adoção de práticas apropriadas de monitoramento ambiental, como subsídio à gestão dos recursos naturais ali presentes.

A companhia estatal Petróleos Mexicanos S.A. (PEMEX) é responsável pelas atividades de exploração, produção, refino e comercialização dos hidrocarbonetos líquidos e gasosos contidos no subsolo do México, como também pela distribuição dos derivados de petróleo. A companhia ocupa o décimo sétimo lugar no ranking mundial (Anuário Estatístico - PEMEX, 2007).

A atividade exploratória na porção mexicana do Golfo do México começou em 1971, logo após um pescador, Sr. Cantarell, relatar a existência da prolífica exsudação de óleo na Baía de Campeche. Posteriormente, o complexo de campos descoberto em associação a tal exsudação recebeu seu nome. É interessante notar que relatos históricos já mencionavam, desde épocas pré-colombianas, ocorrência do fenômeno de exsudação na área, que eram denominadas no dialeto nativo de *chapopoteras*.

A exsudação petrolífera de Cantarell ocorre em águas rasas (40m de profundidade), em uma província de tectônica salina, associada à geração ativa de petróleo, proveniente, principalmente, de rochas geradoras de idade Titoniana (Jurássico Superior). No local, as falhas e os corpos salinos (domos) penetram os sedimentos e criam caminhos de migração, da rocha geradora ou reservatório, até o assoalho marinho (Miranda *et al.*, 2004). Na Figura 1.2, encontra-se uma fotografia panorâmica obtida a partir de helicóptero, no qual é possível constatar a intensa atividade da exsudação do Campo de Cantarell.



Figura 1.2 – Vista aérea da exsudação de óleo do Campo de Cantarell. Como fator de escala, notar a plataforma petrolífera no canto superior direito da foto (Fonte: Mendoza *et al.*, 2003).

1.4 – Visão geral da metodologia proposta

A metodologia proposta para este estudo fundamenta-se na estrutura do processo de KDD apresentada por Resende *et al.* (2003). Tal abordagem tem início na fase de identificação do problema, seguida pelas seguintes etapas: pré-processamento, extração de padrões (CBA), pós-processamento (medidas de interesse objetivas e subjetivas) e análise do potencial de uso do novo conhecimento gerado. O fluxo de realização das etapas é iterativo; logo, existe a possibilidade de repetições integrais ou parciais do processo na busca de resultados cada vez mais satisfatórios para o especialista do domínio da aplicação.

O especialista do domínio conhece profundamente o assunto e o ambiente do contexto da aplicação onde é realizado o processo de KDD. Portanto, este usuário fornece suporte ao analista de KDD em todas as etapas do processo, o que torna sua participação ativa e de suma importância.

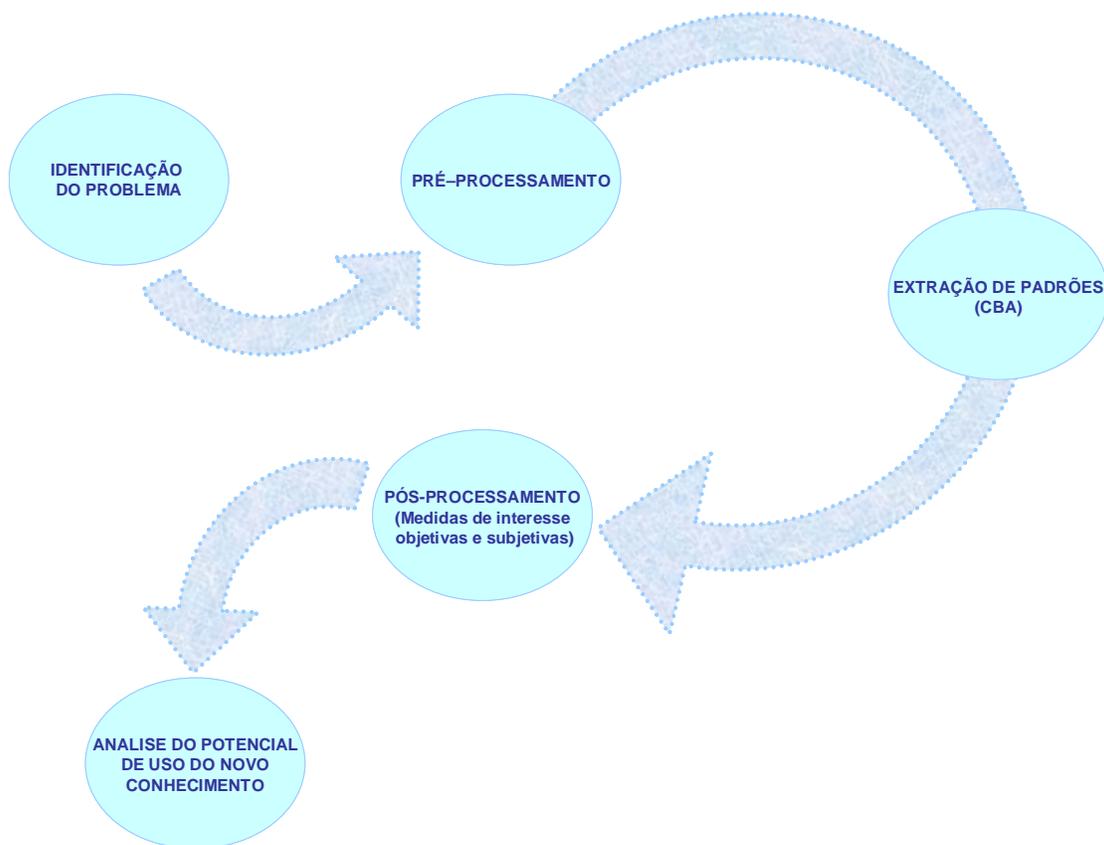


Figura 1.3 – Os cinco componentes do processo de KDD incluídos na metodologia proposta (modificado de Rezende *et al.*, 2003).

Os cinco componentes da metodologia proposta, apresentados na Figura 1.3, são sumariados a seguir:

A – Identificação do problema (exsudações petrolíferas na Baía de Campeche):

- Obtenção de conhecimento do domínio da aplicação;
- Aquisição da base de dados;
- Definição do objetivo e das metas a serem alcançados no processo de KDD;
- Identificação do conhecimento a ser extraído;

B – Pré-processamento:

- Identificação dos atributos relevantes;
- Tratamento dos dados inconsistentes;
- Redução do número de valores dos atributos (discretização);
- Preparação dos dados para a geração de regras;

C – Extração de padrões (CBA):

- Análise da configuração para as medidas de suporte e confiança;
- Geração das regras de associação, através do aplicativo CBA;
- Estabelecimento de critérios para a identificação do domínio de regras fortes;

D – Pós-processamento:

- Identificação do conhecimento subjetivo do especialista do domínio;
- Avaliação da qualidade das regras, através de uma medida de interesse objetiva (*lift*);
- Identificação das regras inéditas, inesperadas e/ou interessantes;

E – Análise do potencial do uso do novo conhecimento.

CAPÍTULO 2

DETECÇÃO DE EXSUDAÇÕES DE ÓLEO NA SUPERFÍCIE DO MAR UTILIZANDO SENSORIAMENTO REMOTO POR RADAR

2.1 – Conceitos básicos

O sensoriamento remoto por radar permite a aquisição de informações sobre áreas de interesse (alvos), indicando sua distância à antena (range) e a resposta do sinal de retorno (amplitude e fase). O termo radar é o acrônimo da expressão (RAdio Detection And Ranging). Os sistemas de radar transmitem e recebem pulsos de energia eletromagnética na faixa espectral de microondas (Sabins, 1997).

Os radares imageadores são sistemas ativos, que produzem sua própria fonte de irradiação. Tal tecnologia, portanto, é capaz de gerar imagens durante o dia ou à noite, sem necessidade de considerar o sol com uma fonte de energia. Além disso, na região das microondas, a transmitância atmosférica é alta, o que contribui para a aquisição de dados de radar sob condições meteorológicas adversas (Figura 2.1).

O princípio básico do radar imageador consiste na transmissão da radiação eletromagnética (REM) na direção da superfície terrestre e na gravação de intensidade, tempo de retardo e fase da energia retroespalhada pelo alvo na superfície (Soler, 2002). É através da REM retornada à antena que a informação sobre o alvo é recebida pelo sensor. Pelo tempo de retorno, calcula-se a distância e a posição entre o alvo e a antena de radar. Quanto maior for o retorno dos pulsos emitidos, maior será a intensidade ou brilho observado na imagem de radar. Informações de amplitude e fase do sinal dependem de fatores como as características do alvo e a configuração do sensor.

O nível de cinza da imagem de radar representa a intensidade da energia retroespalhada pelo alvo que é registrada pelo sensor. Este nível de cinza é usualmente denominado DN, do inglês *Digital Number*. Cabe ao sensor captar a energia retroespalhada em intervalos específicos de comprimento de onda, codificá-la e enviar tais informações sob a forma de dados brutos para estações terrestres, onde

serão convertidos em imagens digitais. Desse modo, os *pixels* da imagem, após seu processamento, indicam usualmente a intensidade do sinal de retorno. Sistemas mais sofisticados também capturam a informação da fase da REM recebida pela antena.

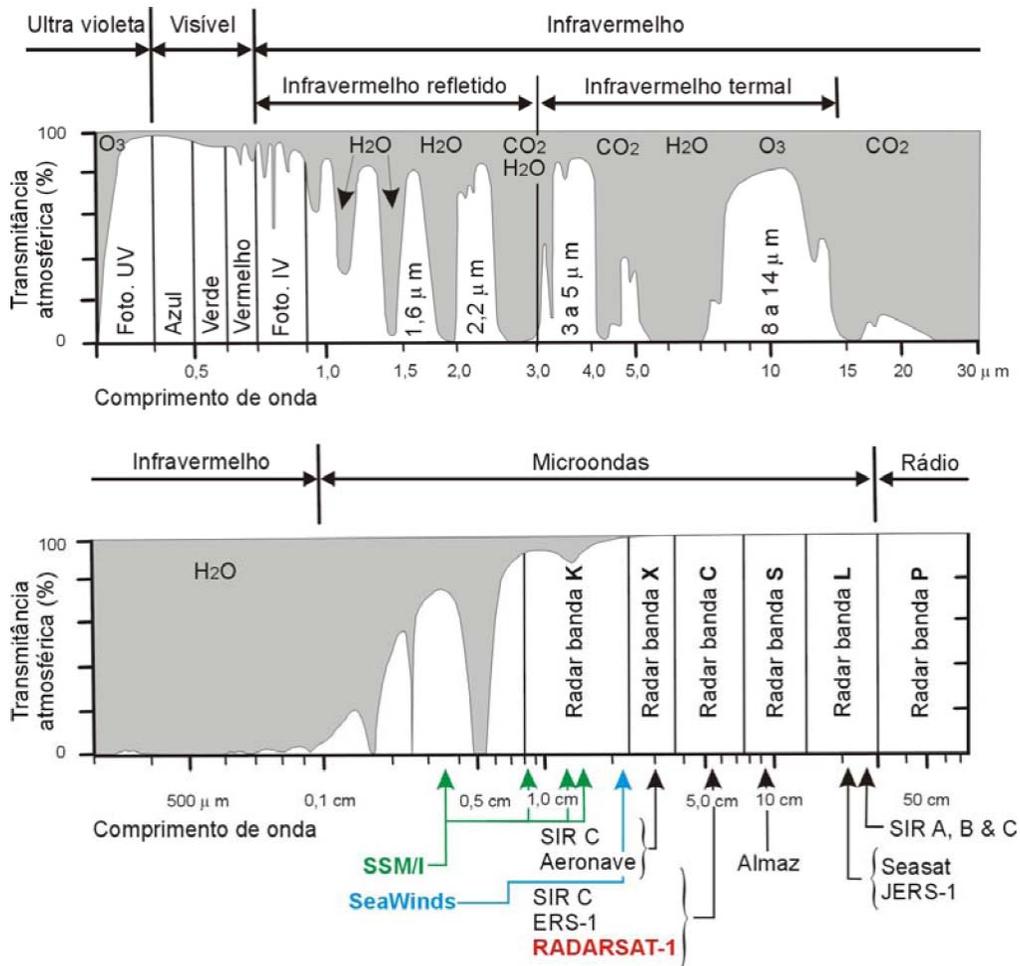


Figura 2.1 – Diagrama expandido do espectro eletromagnético em relação à transmitância atmosférica (modificado de Sabins, 1997). H₂O, CO₂ e O₃ referem-se a gases na atmosfera que absorvem a energia eletromagnética.

A imagem de radar é constituída pela integração dos pulsos retroespalhados pelo alvo. O sinal de retorno sofre influência das características do sistema sensor, dentre as quais destacam-se (A) o comprimento de onda ou freqüência, (B) a polarização e (C) a geometria de aquisição da imagem.

Os intervalos de comprimento de onda e freqüência são representados por letras, que correspondem às bandas de radar. Quanto maior o comprimento de onda,

menor é a frequência correspondente, e menor é a quantidade de REM transmitida. Os sistemas orbitais de radar atualmente em operação utilizam bandas X, C e L (Tabela 2.1). O satélite canadense RADARSAT-1, empregado no monitoramento de exsudação de Cantarell, trabalha com a banda C.

A onda eletromagnética se propaga no vácuo à velocidade da luz. A conversão de comprimento de onda (λ) para frequência (f) está descrita na equação

$$\lambda = \frac{c}{f}, \quad \text{Equação 2.1}$$

onde:

c = velocidade da luz (3×10^8 m/s);

f = frequência em hertz (Hz);

λ = comprimento de onda (m).

Tabela 2.1 – Bandas espectrais, intervalos de comprimento de onda e frequência utilizados em sistemas de radar (modificado de Sabins, 1997).

Banda de radar	Comprimento de onda (cm)	Frequência (GHz)
Ka	0,8 – 1,1	40,0 – 26,5
Ks	1,1 – 1,7	26,5 – 18,0
Ku	1,7 – 2,4	18,0 – 12,5
* X	2,4 – 3,8	12,5 – 8,0
* C	3,8 – 7,5	8,0 – 4,0
S	7,5 – 15,0	4,0 – 2,0
* L	15,0 – 30,0	2,0 – 1,0
P	30,0 – 100,0	1,0 – 0,3

* bandas utilizadas por sistemas orbitais de radar hoje em operação.

O pulso de REM transmitido pelo radar consiste na conjugação de campos elétricos e magnéticos segundo um padrão ondulatório harmônico (ondas espaçadas regularmente no tempo). Tal oscilação é caracterizada pelo comprimento de onda, que mede a distância entre dois picos máximos dos citados campos. Já a periodicidade no tempo é medida pela frequência, que é expressa pelo número de cristas de onda que passam por um determinado ponto do espaço (Figura 2.2).

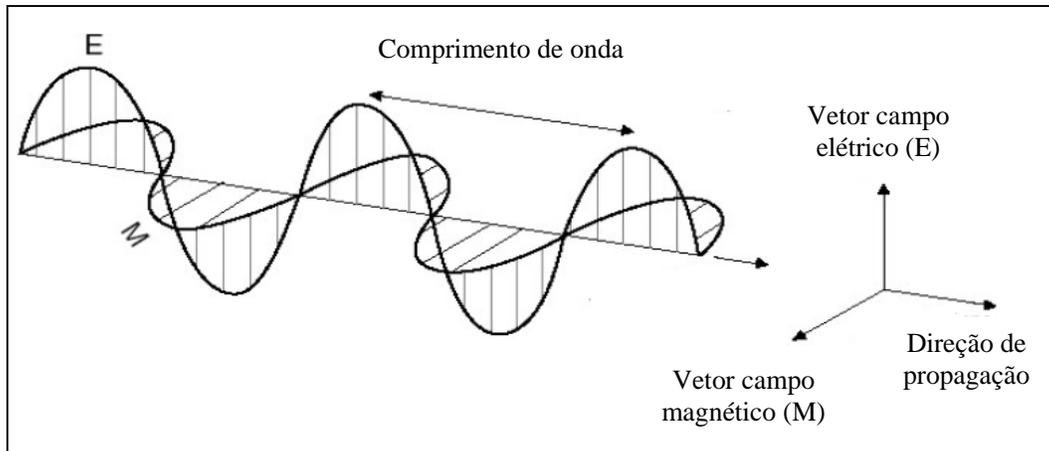


Figura 2.2 – Visão esquemática do pulso do radar (modificado de Sabins, 1997).

O conceito de polarização diz respeito à orientação dos campos elétricos e magnéticos da REM em relação ao eixo da antena. A polarização transmitida e recebida pelo sistema de radar pode ser horizontal (com o campo elétrico paralelo ao eixo da antena) ou vertical (com o campo elétrico perpendicular ao eixo da antena). A polarização é denominada paralela quando os campos elétricos transmitidos e recebidos pelo sensor se orientam na mesma direção. Neste caso, existem duas formas de polarização: HH (transmissão e recepção horizontal) e VV (transmissão e recepção vertical). Além disso, a polarização é denominada cruzada quando os campos elétricos transmitidos e recebidos pelo sensor são ortogonais, o que resultam em duas formas adicionais de polarização: HV (transmissão horizontal e recepção vertical) e VH (transmissão vertical e recepção horizontal).

A geometria de aquisição de um sistema de radar de abertura sintética (SAR) é ilustrada na Figura 2.3.

O ângulo de depressão é definido como ângulo agudo entre o pulso de radar e um plano horizontal, sendo o ângulo de visada seu complemento. O ângulo de incidência é formado entre o pulso do radar e a linha perpendicular à superfície da terra. Esse ângulo varia ao longo da faixa imageada, quanto mais larga for a faixa, maiores serão os intervalos de ângulos de incidência ao longo da mesma. Em regiões de área plana (sem relevo), como é o caso das regiões oceânicas, os ângulos de incidência e depressão são complementares (Figura 2.4). O ângulo de incidência é um dos principais fatores que influenciam o retroespalhamento dos alvos.

1. Superfícies lisas (reflexão especular): quase toda energia incidente é refletida com um ângulo de reflexão igual e oposta ao ângulo de incidência. Neste caso, a antena de radar recebe pouco ou nenhum sinal de retorno, fazendo com que feições com essas características apareçam com muito baixo valor de DN nas imagens;
2. Superfícies com rugosidade intermediária (transição entre reflexão especular e espalhamento difuso): uma porção da energia incidente é refletida em várias direções, enquanto a outra parte é refletida especularmente. As feições com superfícies de rugosidade intermediária apresentam valores baixos a médios de DN;
3. Superfície totalmente rugosa (espalhamento difuso): toda a energia incidente é espalhada difusamente de maneira uniforme em várias direções. Assim, feições com essas características aparecem com valores médios a altos de DN.

A intensidade do sinal de retorno pode também ser intensificada pela presença de alvos com geometrias específicas, que provocam múltiplas reflexões do sinal transmitido. Os refletores de canto (*corner reflectors*) são formados quando duas ou três superfícies voltadas para o radar formam ângulos retos. Como consequência do forte retorno na imagem de radar a partir de tais diedros e triedros, o pixel correspondente aparece saturado, ou seja, com valores muito altos de DN. Embarcações e plataformas na superfície do oceano são exemplos de refletores de canto (Roriz, 2006).

Os critérios para definição do espalhamento proposto por Rayleigh são aplicados quando as dimensões das irregularidades superficiais são comparáveis ao comprimento de onda do sinal do radar. Como os comprimentos de onda na faixa das microondas são centimétricos, as irregularidades das superfícies detectadas se apresentam na mesma escala (Sabins, 1997). Uma superfície pode ser considerada lisa (reflexão especular) quando

$$h < \frac{\lambda}{8 \cos \theta} , \quad \text{Equação 2.2}$$

onde:

h = altura média das variações da superfície (cm);

λ = comprimento de onda (cm);

θ = ângulo de incidência (graus).

Peake & Oliver (1971) modificaram o critério de Rayleigh, definindo critérios para alvos lisos, intermediários e rugosos, conforme descrito abaixo:

• Superfície lisa:
$$h < \frac{\lambda}{25 \cos \theta} ;$$
 Equação 2.3

• Superfície rugosa:
$$h > \frac{\lambda}{4,4 \cos \theta} ;$$
 Equação 2.4

• Superfície intermediária:
$$\frac{\lambda}{25 \cos \theta} \leq h \leq \frac{\lambda}{4,4 \cos \theta} .$$
 Equação 2.5

2.3 – O efeito de redução do espalhamento Bragg pela presença de óleo

O retorno do sinal de radar na superfície do mar é descrito pelo modelo de espalhamento Bragg. Esse fenômeno ocorre devido à rugosidade local causada pela presença de pequenas ondas induzidas pelo vento (na escala de milímetro a centímetros), denominadas ondas capilares de gravidade ou ondas Bragg. A teoria formulada por Bragg estabelece que, para uma superfície aleatória, dividida em seus componentes espectrais, a energia retroespalhada dominante origina-se das componentes que estiverem em ressonância com a onda incidente (Vesecky, 1995).

No caso das imagens de radar, que possuem comprimentos de onda centimétricos, a superfície do mar apresenta valores médios a alto de DN, caracterizando uma reflexão difusa (Figura 2.5).

Manchas de óleo na superfície do mar tendem a amortecer (suavizar) as ondas Bragg, fazendo com que o retorno para a antena seja muito menor onde elas ocorrem. A estabilidade local provocada pela presença de óleo reduz a amplitude das cristas das ondas, produzindo superfícies menos rugosas e relativamente lisas em reação ao oceano. Essas áreas lisas aparecem com valores baixos de DN na imagem de radar, devido à reflexão especular do alvo associada à presença de óleo (Figura 2.5).

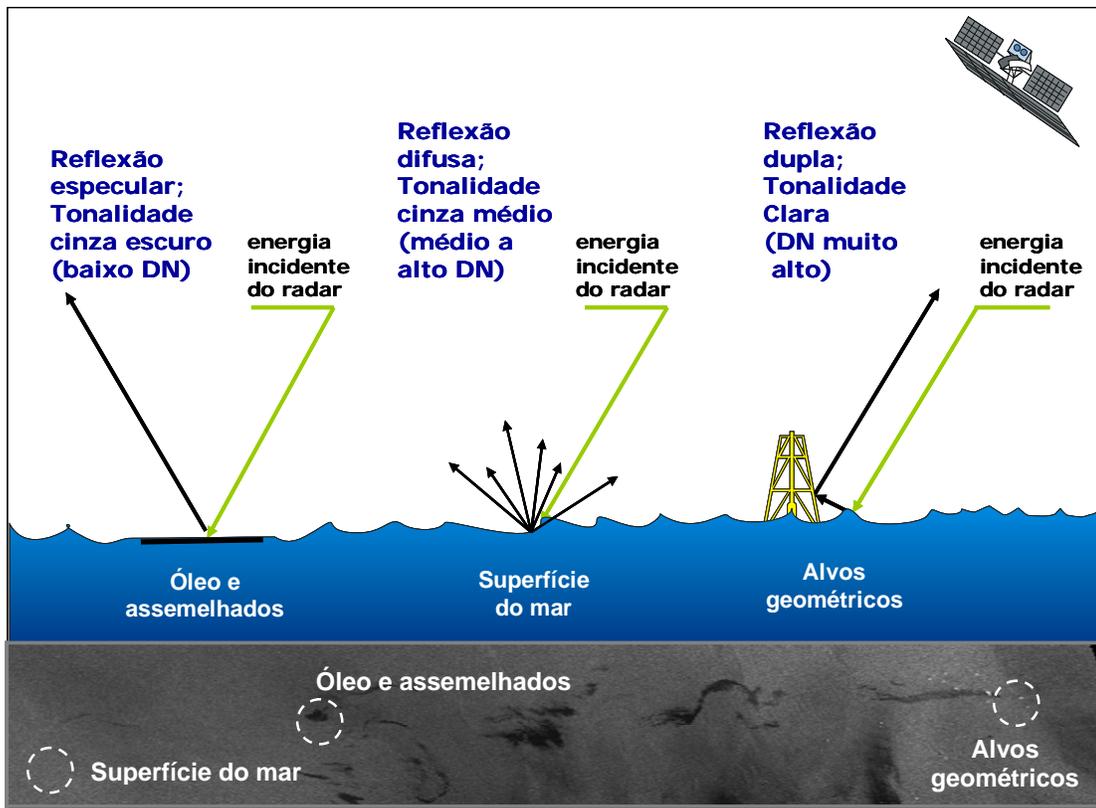


Figura 2.5 – Tipos básicos de interação do pulso de radar com a superfície do mar (modificado de Sabins,1997).

Entretanto, assim como o óleo, outros fenômenos provocam a redução da rugosidade da superfície do mar, amortecendo as ondas capilares. Esses fenômenos também produzem áreas com baixos valores de DN nas imagens SAR, que são consideradas, portanto, como falsos alvos. A identificação dos falsos alvos é subsidiada pela análise de dados meteo-oceanográficos, que dão suporte à interpretação de exsudações de óleo no mar utilizando radares orbitais.

2.4 – O sistema RADARSAT-1

O RADARSAT-1 é um sofisticado sistema de sensoriamento remoto por radar desenvolvido pela Agência Espacial Canadense (CSA) e lançado em novembro de 1995, com propósito de monitorar mudanças ambientais e recursos naturais.

A capacidade de imagear grandes áreas, manter alta resolução espacial e operar dia e noite, em quaisquer condições meteorológicas, faz esta tecnologia

particularmente atraente para aplicações que requerem aquisições confiáveis e regulares, como, por exemplo, um programa de monitoramento de atividades petrolíferas (Miranda *et al.*, 2004).

2.4.1 – Características gerais

O radar de abertura sintética (SAR) a bordo do RADARSAT-1 opera na banda C, com frequência de aproximadamente 5,3 GHz e polarização HH (transmissão e recepção horizontal). A Tabela 2.2 sintetiza as principais características de tal satélite.

Tabela 2.2 – Características do satélite RADARSAT-1 (Fonte: RADARSAT *International*, 1996).

Geometria	Circular, síncrona com o sol
Altitude	798 km
Inclinação	98,6 ^o
Período	100,7 minutos
Repetição do ciclo	24 dias
Órbitas por dia	14
Frequência	5,3 GHz
Comprimento de onda	5,6 cm (Banda C)
Polarização	HH (Horizontal-Horizontal)

A fim de servir a usuários comerciais, os dados são transmitidos em tempo real a estações de recepção em terra (*Ground Receiving Stations*). Porém, quando a área imageada encontra-se fora da cobertura desses centros de recepção, o sistema é capaz de gravar aproximadamente 10 minutos de dados SAR com alta qualidade, até atingir uma região abrangida por uma das diversas estações espalhadas pelo globo terrestre (CSA, 2008).

2.4.2 – Modos de imageamento

O sistema RADARSAT-1 conta com sete modos de operação distintos, em virtude da grande variedade de feixes de transmissão (Tabela 2.3). Os modos são caracterizados pelas dimensões da faixa imageada, com variação de 50km a 500km, bem como pela resolução espacial, que pode atingir valores de 8 m a 100m. O

RADARSAT-1 foi o sistema de cobertura global pioneiro na utilização de diferentes geometrias de aquisição (Figura 2.6).

Para obter contraste adequado nas imagens das feições representativas da presença de óleo na superfície oceânica, é importante considerar o ângulo de incidência. Segundo Miranda *et al.* (2004), ângulos de incidência baixos, entre 10° e 40°, melhoram o contraste para este tipo de aplicação. Os seguintes modos de operação se enquadram nesta característica: *Extended Low 1*, *Standard 1* até *4*, *Wide 1 e 2* e *ScanSAR Narrow 1*.

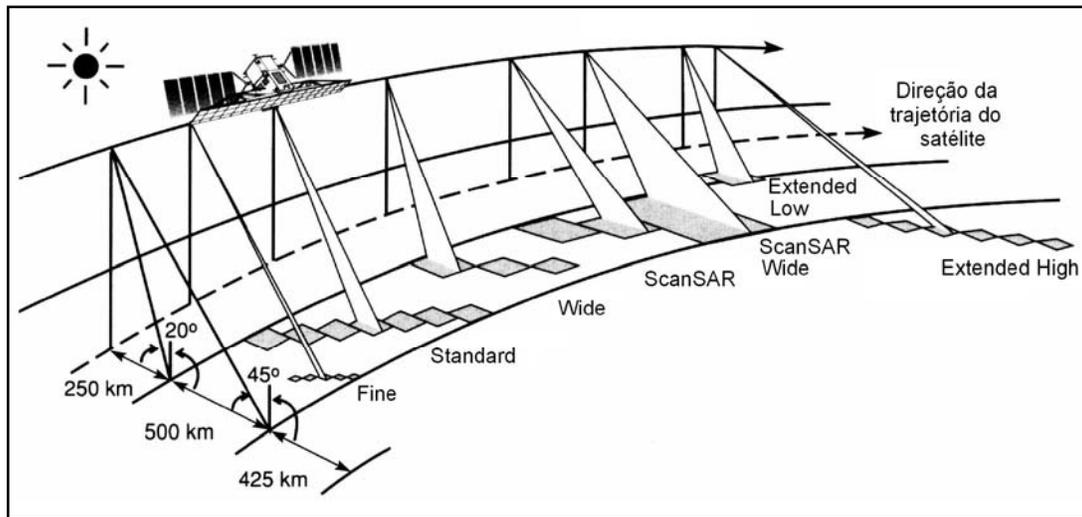


Figura 2.6 – Geometria de visada e modos de operação do satélite RADARSAT-1 (modificado de RADARSAT *International*, 1996).

Tabela 2.3 – Características dos diferentes modos de imageamento do satélite RADARSAT-1 (Fonte: RADARSAT *International*, 1996).

Modos de imageamento	Posição do feixe	Ângulo de incidência (graus)	Área nominal (km)	Resolução nominal (metros)
Fino (<i>Fine</i>)	<i>F1 near</i>	36.4 - 39.6	50 X 50	8
	<i>F1</i>	36.8 - 39.9		
	<i>F1 far</i>	37.2 - 40.3		
	<i>F2 near</i>	38.8 - 41.8		
	<i>F2</i>	39.2 - 42.1		
	<i>F2 far</i>	39.6 - 42.5		
	<i>F3 near</i>	41.1 - 43.7		
	<i>F3</i>	41.5 - 44.0		
	<i>F3 far</i>	41.8 - 44.3		
	<i>F4 near</i>	43.1 - 45.5		
	<i>F4</i>	43.5 - 45.8		
	<i>F4 far</i>	43.8 - 46.1		
	<i>F5 near</i>	45.0 - 47.2		
	<i>F5</i>	45.3 - 47.5		
	<i>F5 far</i>	45.6 - 47.8		
Padrão (<i>Standard</i>)	<i>S1</i>	20 - 27	100 X 100	25
	<i>S2</i>	24 - 31		
	<i>S3</i>	30 - 37		
	<i>S4</i>	34 - 40		
	<i>S5</i>	36 - 42		
	<i>S6</i>	41 - 46		
	<i>S7</i>	45 - 49		
Extenso (<i>Wide</i>)	<i>W1</i>	20 - 31	165 X 165	30
	<i>W2</i>	31 - 39	150 X 150	
	<i>W3</i>	39 - 45	150 X 150	
ScanSAR Estreito (<i>ScanSAR Narrow</i>)	<i>SCN1</i>	20 - 40	300 X 300	50
	<i>SNB</i>	31 - 46		
ScanSAR Extenso (<i>Scan SAR Wide</i>)	<i>SW1</i>	20 - 49	500 X 500	100
Alta extensão (<i>Extended High</i>)	<i>H1</i>	49 - 52	75 X 75	25
	<i>H2</i>	50 - 53		
	<i>H3</i>	52 - 55		
	<i>H4</i>	54 - 57		
	<i>H5</i>	56 - 58		
	<i>H6</i>	57 - 59		
Baixa Extensão (<i>Extended Low</i>)	<i>EXL1</i>	10 - 23	170 X 170	35

2.4.3 – Órbitas utilizadas na aquisição de dados

A órbita descrita pelo sistema RADARSAT-1 é circular em torno da Terra, quasi-polar e síncrona ao sol, com período orbital de cerca de 100,7 minutos. A inclinação da órbita é de $98,6^{\circ}$ em relação ao Equador. Em média, o satélite completa um ciclo de revista em torno da terra em 24 dias, para um mesmo modo de aquisição. Por possuir vários modos de imageamento, o RADARSAT-1 pode ser configurado de maneira a permitir um ciclo de revista mais freqüente.

Esse sistema possui a bordo um radar com visada lateral para direita, podendo adquirir imagens tanto em órbita ascendente, com visada para leste, quanto em órbita descendente, com visada a oeste (Figura 2.7). O satélite atravessa o Equador às 18:00 h, em horário local na órbita ascendente e às 06:00 h, em horário local na órbita descendente (Soler, 2002).

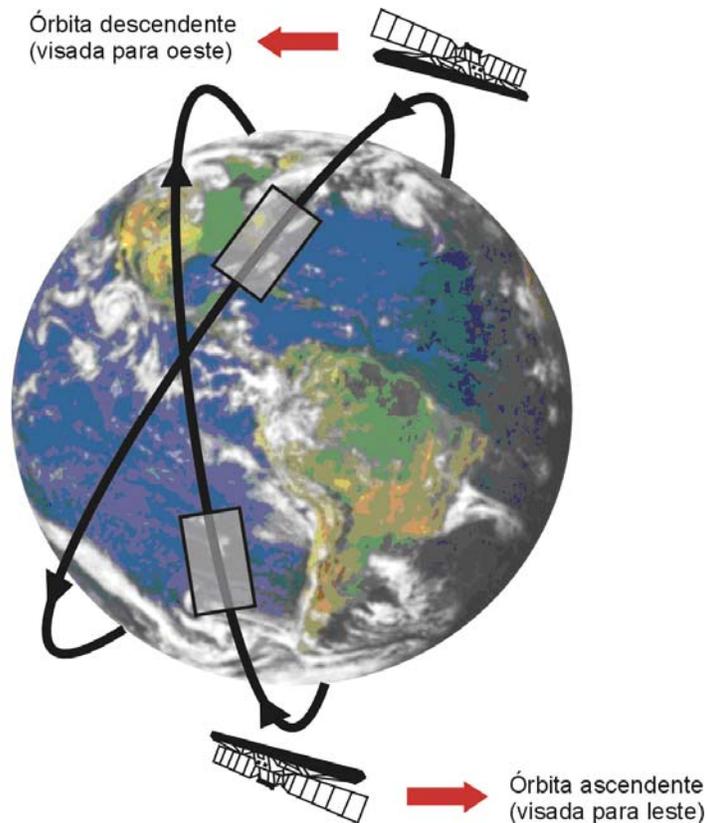


Figura 2.7 – Configuração das órbitas descendente (com visada para oeste) e ascendente (com visada para leste) do RADARSAT-1 (modificado de RADARSAT International, 1996).

2.5 – Emprego de dados meteo-oceanográficos como suporte à interpretação de exsudações de óleo no mar

Estudos constando da aplicação de dados RADARSAT-1 para a detecção de películas de óleo no oceano têm se mostrado eficazes (Bentz e Miranda, 2001; Bentz, 2006). No entanto, a identificação de óleo na superfície do mar é uma tarefa complexa devido à presença de inúmeros fenômenos que também atenuam as ondas capilares, superficiais originando feições escuras nas imagens de radar. Essas feições semelhantes às causadas por óleo são classificadas, portanto, como falsos alvos. Para auxiliar na interpretação das imagens SAR, são utilizadas informações ambientais obtidas através do processamento de dados de vários sensores orbitais, a fim de caracterizar as diversas feições existentes na imagem e identificar, dentre as manchas escuras, aquelas que têm sua origem associada ao óleo (Silva Júnior *et al.*, 2003). Tais informações devem ser adquiridas o mais próximo possível do período de tomada da imagem de radar. Conforme descrito a seguir, os principais parâmetros meteo-oceanográficos que auxiliam nessa discriminação são: temperatura da superfície do mar, temperatura de topo de nuvem, velocidade do vento, altura da onda e concentração de clorofila-a. Eles propiciam o diagnóstico objetivo e eficiente das características meteo-oceanográficas no momento da passagem do RADARSAT-1 (Silva Júnior *et al.*, 2003). Por clorofila-a entende-se um grupo de pigmentos fotossintético encontrado nos cloroplastos das plantas (incluindo-se algas, cianofíceas e procariontes). Esses pigmentos apresentam diferentes respostas à radiação incidente, absorvendo os comprimentos de onda do azul (433nm) e vermelho (686nm), e refletindo o verde (550nm). Dados como temperatura do mar (TSM) e da concentração da clorofila-a da camada superficial do mar podem ser obtidos através dessas imagens a partir da aplicação de algoritmos da cor do oceano (Kampel *et al.*, 2005).

2.5.1 – Mapa de temperatura da superfície do mar (TSM)

Dados de temperatura da superfície do mar (TSM) são obtidos a partir do sensor *Advanced Very High Resolution Radiometer* – AVHRR, a bordo dos satélites da série *National Oceanic and Atmospheric Administration* (NOAA 12, 14, 15, 16, 17 e 18). Os satélites dessa série possuem órbita polar e altitude que varia de 830 a 870 km. Seus sensores estão distribuídos pelos canais do visível e

infravermelho, em cinco bandas. O sistema de varredura é transversal, gerando imagens de aproximadamente 2.300 km de largura.

O sensor AVHRR mede a quantidade de radiação emitida pela superfície dos oceanos, o que permite o cálculo de sua temperatura. A partir do mapa de TSM é possível a identificação de regiões oceânicas de alto gradiente térmico, que apresentam alterações na sua rugosidade superficial. Assim, esses mapas oferecem ao intérprete uma visão sinótica da temperatura da superfície do mar, que favorece a identificação de importantes feições oceânicas, tais como zonas de ressurgência, frentes oceânicas, meandros, vórtices e padrões de correntes (Miranda *et al.*, 2004). Essas feições são utilizadas na identificação de falsos alvos durante a análise de dados SAR para detecção de óleo na superfície do mar.

Na Figura 2.8, pode-se observar o mapa de temperatura da superfície do mar, obtido a partir do sensor AVHRR a bordo do satélite NOAA-17. Notam-se variações no gradiente termal, temperaturas entre 24°C e 29°C, principalmente na área próxima à costa na parte superior da imagem. Este fenômeno justifica as mudanças na rugosidade da superfície do mar nessa região observadas na imagem SAR (Figura 2.9).

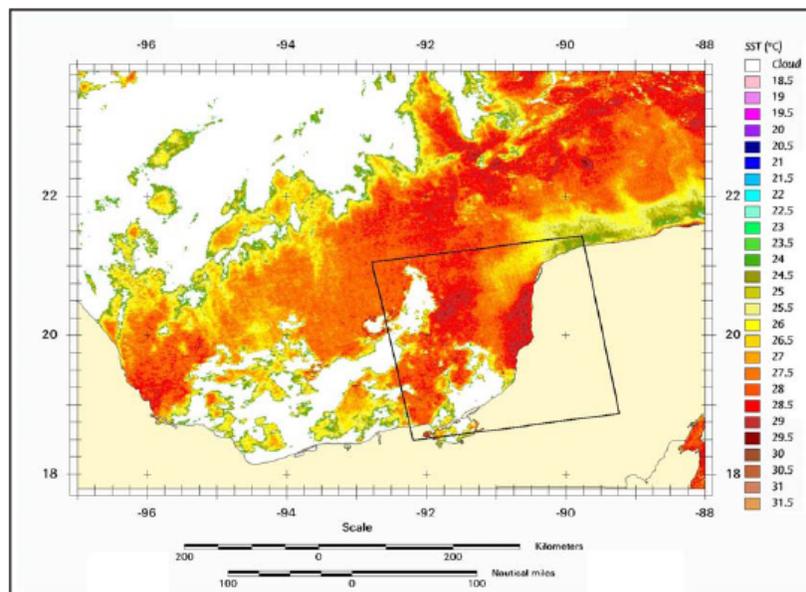


Figura 2.8 – Mapa de temperatura da superfície do mar, obtido a partir do sensor AVHRR a bordo do satélite NOAA-17, em 27 de julho de 2006. O retângulo em preto está representando o frame da imagem SAR da Figura 2.9 (Roriz, 2006).

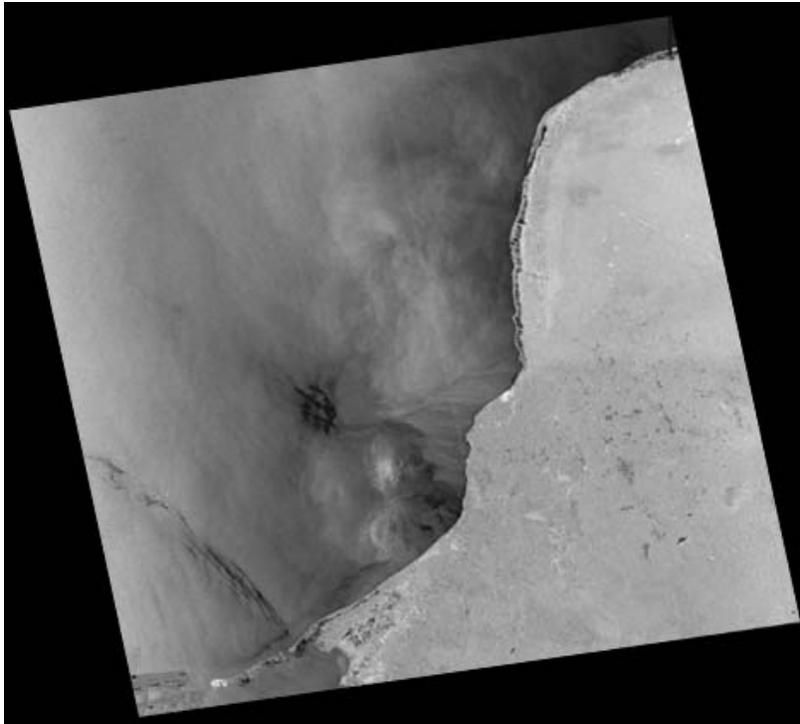


Figura 2.9 – Imagem do satélite RADARSAT-1, no modo SCN1 (órbita ascendente), adquirida no Golfo do México, em 27 de julho de 2006 (Roriz, 2006).

2.5.2 – Mapa de temperatura do topo de nuvens (TTN)

Para se observar o padrão de distribuição de nuvens e suas respectivas temperaturas, diversos sensores são utilizadas, principalmente aqueles a bordo dos satélites das séries NOAA e GOES (*Geostationary Operational Environmental Satellite*). Os satélites GOES são geoestacionários e se posicionam a uma altitude de aproximadamente 35.800 km. Sua grande vantagem em relação aos demais é a alta resolução temporal, o que permite a obtenção de imagens da mesma área na superfície da Terra a cada 30 minutos (Roriz, 2006). O Sensor *Imager* a bordo deste satélite registra a radiação proveniente da Terra em 5 bandas espectrais, cujo comprimento de onda varia do visível ao infravermelho termal.

Os mapas de TTN são utilizados para identificar os locais onde células de chuva, que se desenvolvem em formações do tipo *Cumulus Nimbus*, estão potencialmente presentes (Miranda *et al.*, 2004). Nuvens com temperatura do topo extremamente baixas, da ordem de -40°C , são associadas a células de chuva

forte. A ocorrência desse fenômeno resulta no alisamento da superfície do mar, com o amortecimento das ondas capilares em razão da turbulência.

Na Figura 2.10, pode-se observar o mapa de temperatura de topo de nuvens confeccionado a partir dos dados do sensor AVHRR (NOAA-15). Uma ampla área no mapa apresenta temperatura de -60°C , que está associada à presença de uma célula de chuva de grandes dimensões. Tal fato justifica a presença de grande variação na rugosidade da superfície do mar, que se observa no canto direito da imagem do satélite RADARSAT-1 da Figura 2.11.

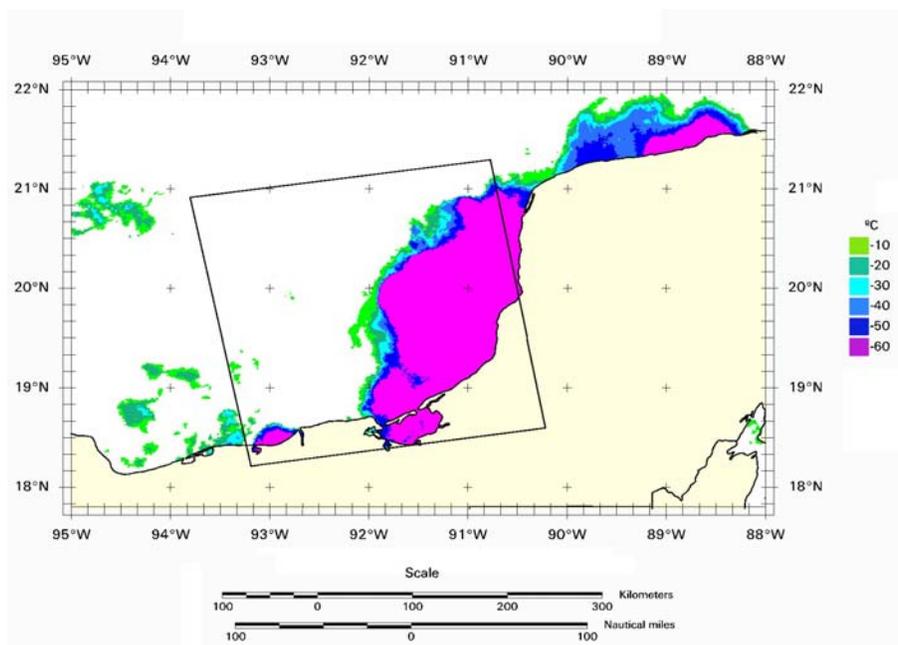


Figura 2.10 – Mapa de temperatura do topo de nuvem obtido a partir de uma imagem do sensor AVHRR, a bordo do satélite NOAA-15, adquirida em 05 de julho de 2004. O retângulo em preto está representando o frame da imagem SAR da Figura 2.11 (Roriz, 2006).

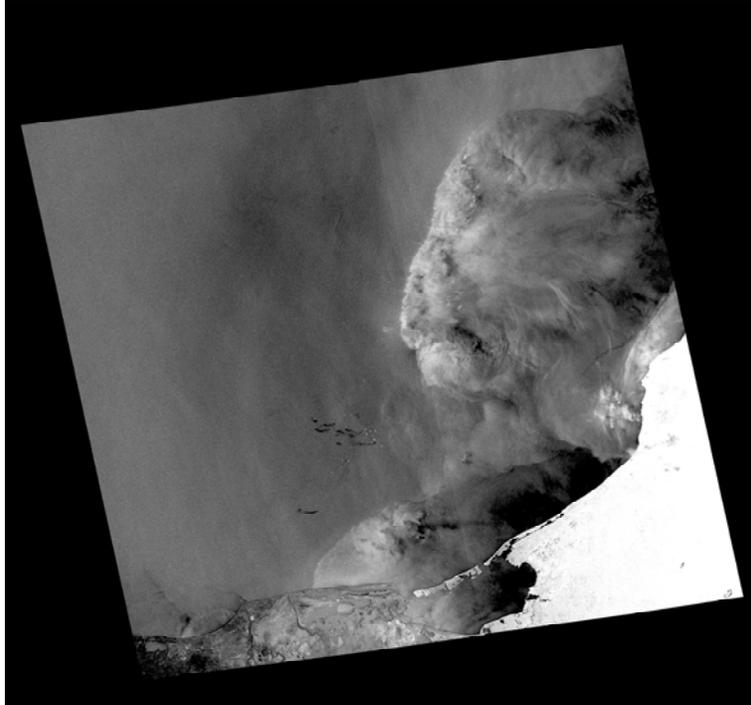


Figura 2.11 – Imagem do satélite RADARSAT-1, no modo SCN1 (órbita ascendente), adquirida no Golfo do México, em 06 de julho de 2004 (Roriz, 2006).

2.5.3 – Mapa de intensidade do campo de vento

Para a obtenção informações sobre a intensidade e a direção do campo de vento, pode ser utilizado o sensor SeaWinds (*Sea-viewing Wide Field of View*), a bordo do satélite QuikSCAT, que foi lançado em julho de 1999. Esse sensor é um escaterômetro que opera em oito bandas na faixa das microondas, adquirindo dados em uma faixa de varredura de 1.800 km no terreno, cobrindo, com isto, 90% da superfície do planeta em um dia (Miranda *et al.*, 2004).

Variações na intensidade do campo de vento influenciam fortemente as condições do estado do mar. Ventos fracos, com velocidade abaixo de 3 m/s, provocam redução da rugosidade da superfície do oceano, praticamente inibindo a produção de ondas Bragg, o que resulta em áreas escuras nas imagens SAR, associadas à baixa intensidade do vento. Por outro lado, devido à agitação do mar em áreas de vento muito intenso, com velocidade acima de 8 m/s, não se observa o contraste de rugosidade entre áreas com e sem manchas de óleo. Tal fato, também dificulta a interpretação de feições associadas à presença de óleo nas imagens

SAR. De acordo com Miranda *et al.* (2004), o intervalo ideal de velocidade do vento para a detecção de óleo no oceano situa-se entre 3 e 8 m/s.

Na Figura 2.12, pode-se observar o mapa de intensidade do campo de vento calculado a partir do sensor SeaWinds. Na porção central e noroeste do frame da imagem SAR, nota-se uma área com ventos baixos, entre 1 e 3 m/s. Tal fato justifica o baixo retorno do pulso do radar nessa região (Figura 2.13), associada às condições de ventos baixos, os quais não são propícios à formação das ondas Bragg.

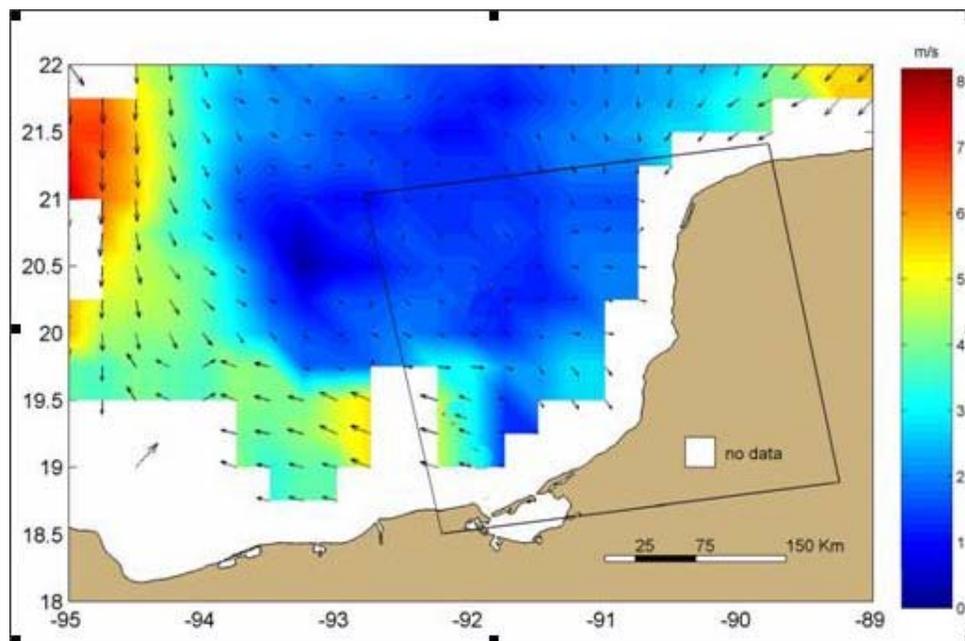


Figura 2.12 – Mapa de intensidade do campo de vento obtido pelo sensor SeaWinds, a bordo do satélite QuickSCAT, em 27 de janeiro de 2004. As setas estão indicando as intensidades e as direções calculadas do vento; o retângulo em preto está representando o frame da imagem SAR da Figura 2.13 (Roriz, 2006).

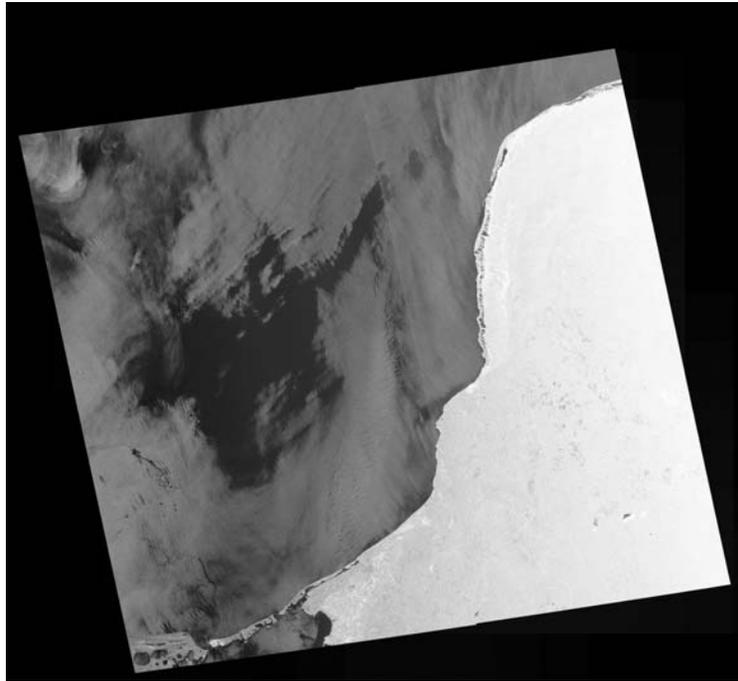


Figura 2.13 – Imagem do satélite RADARSAT-1, no modo SCN1 (órbita ascendente), adquirida em 27 de janeiro de 2004 (Roriz, 2006).

2.5.4 – Mapa de altura significativa de ondas

Informações sobre a altura significativa de ondas podem ser obtidas a partir de um altímetro a bordo do satélite TOPEX TOSEIDON, lançado em 1992. Esse satélite faz parte de projeto conjunto da NASA e do *Centre National d'Etudes Spatiales*. Opera na faixa de microondas, com uma frequência entre 5-15 GHz, nas bandas Ku e C do espectro eletromagnético. A largura da faixa de coleta de dados é de 315 km, com acurácia na ordem de centímetros na medição de alturas significativas de onda.

O principal objetivo de um altímetro é determinar as elevações e depressões presentes na área imageada. Estas variações de altura caracterizam a topografia dinâmica do oceano a qual, por sua vez, está intimamente relacionada com a circulação oceânica superficial (Mata e Garcia, 1996). O altímetro a bordo do satélite TOPEX TOSEIDON adquire um grande número de informações sobre a dinâmica dos oceanos, dentre elas, a altura significativa da onda. O estado de mar ideal para a detecção de exsudações de óleo é caracterizado por ondas com

alturas menores que 1,5 metros (Miranda *et al.*, 2004). Nessa situação, considera-se que o valor do ângulo de incidência do pulso de radar não sofre interferência das variações locais na superfície do oceano.

Na Figura 2.14, pode-se observar o mapa da altura significativa de ondas obtido pelo altímetro a bordo do satélite TOPEX POSEIDON. O forte gradiente observado na porção central do mapa está associado a variações buscadas na altura significativa de onda entre 0,6 e 2,3 metros. Esse fato pode ocasionar variações locais no ângulo de incidência do pulso do radar na imagem RADARSAT-1 (Figura 2.15), que causam diferenças no sinal de retorno da superfície do oceano.

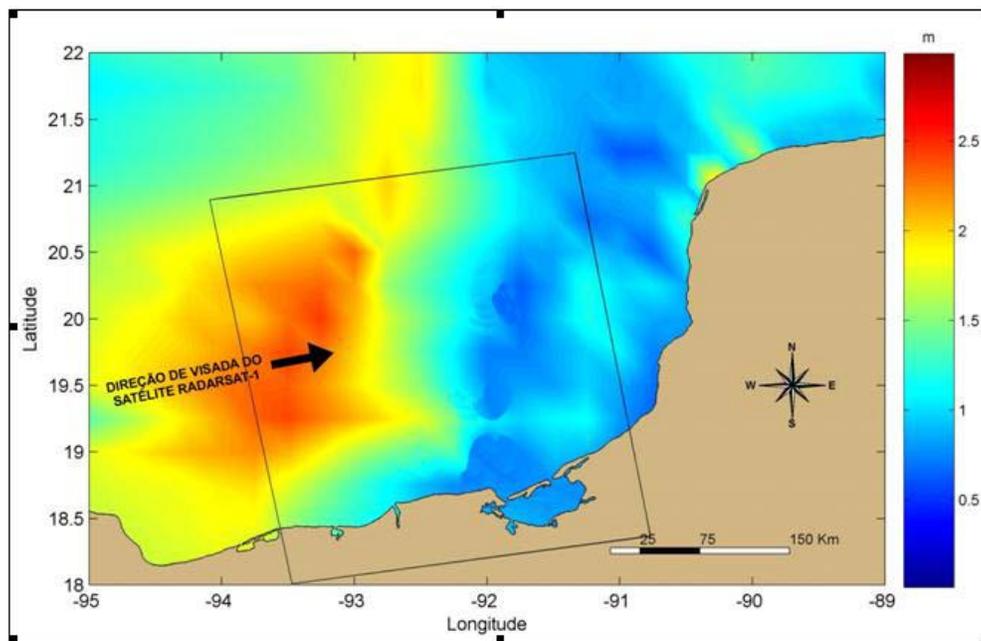


Figura 2.14 – Mapa da altura significativa de ondas, obtido pelo altímetro a bordo do satélite TOPEX-POSEIDON, em 13 de março de 2005. O retângulo em preto está representando o frame da imagem SAR da Figura 2.15 (Roriz, 2006).

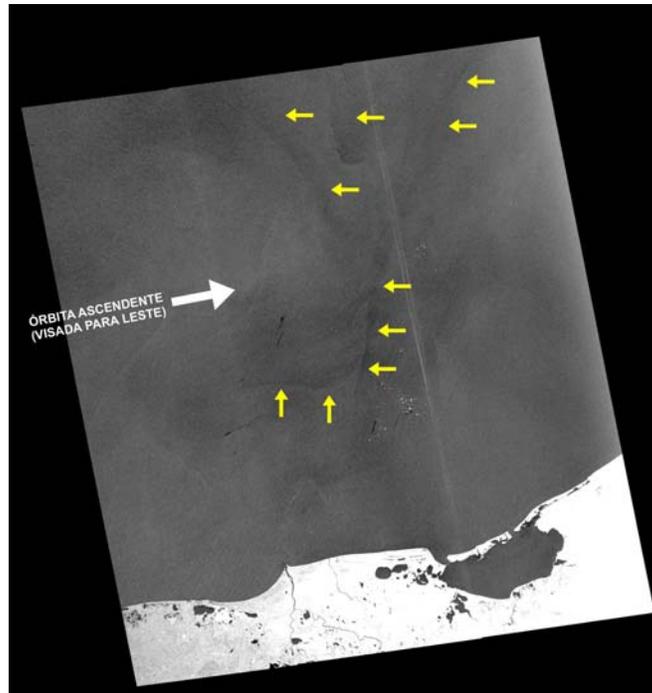


Figura 2.15 – Imagem do satélite RADARSAT-1 no modo SCN2 (órbita ascendente), adquirida em 13 de março de 2005 (Roriz, 2006).

2.5.5 – Mapa de concentração de clorofila-a

Informações sobre produtividade primária relacionada à clorofila-a podem ser obtidas a partir do sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*), que se encontra a bordo dos satélites TERRA e ACQUA da plataforma *Earth Observing System* (EOS), pertencentes à NASA. O satélite TERRA, lançado 1999, passa pelo Equador às 10:30 h da manhã, enquanto o ACQUA, lançado 2002, passa às 13:30 h da tarde. O sensor MODIS visualiza toda a superfície da Terra no período de 1 a 2 dias, com uma varredura de 2.330 km, adquirindo dados em 36 bandas espectrais distintas.

Para identificar regiões na superfície do mar com alta concentração de clorofila-a, mapas gerados a partir do sensor MODIS podem ser prontamente utilizados. Altas concentrações de algas ou de material fito-planctônico reduzem a amplitude das ondas capilares com isso confundindo a interpretação de feições relacionadas à presença de óleo nas imagens SAR.

Na Figura 2.16, confeccionado a partir do processamento de uma imagem do satélite MODIS, pode-se observar o mapa da concentração de clorofila-a. Nota-se alta concentração de clorofila-a na área próxima à costa, por volta de $11,0 \text{ mg/m}^3$. Tal fato justifica o baixo retorno do sinal de radar na região próxima à costa, a nordeste da imagem SAR (Figura 2.17).

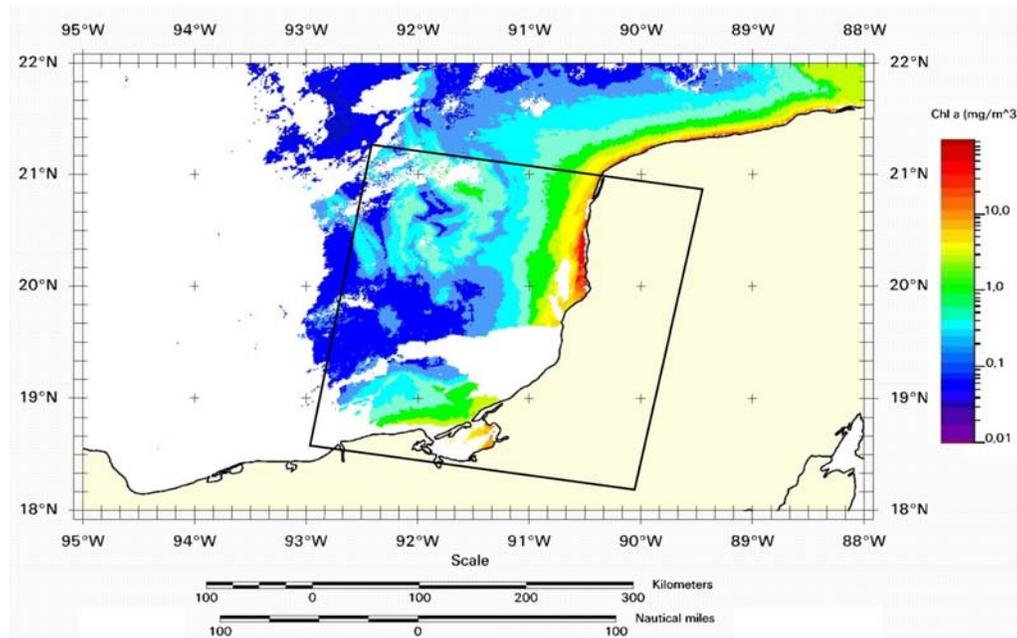


Figura 2.16 – Mapa de concentração de clorofila-a, confeccionado a partir do processamento de uma imagem do satélite MODIS, obtida em 14 de julho de 2004. O retângulo em preto está representando o frame da imagem SAR da Figura 2.17 (Roriz, 2006).

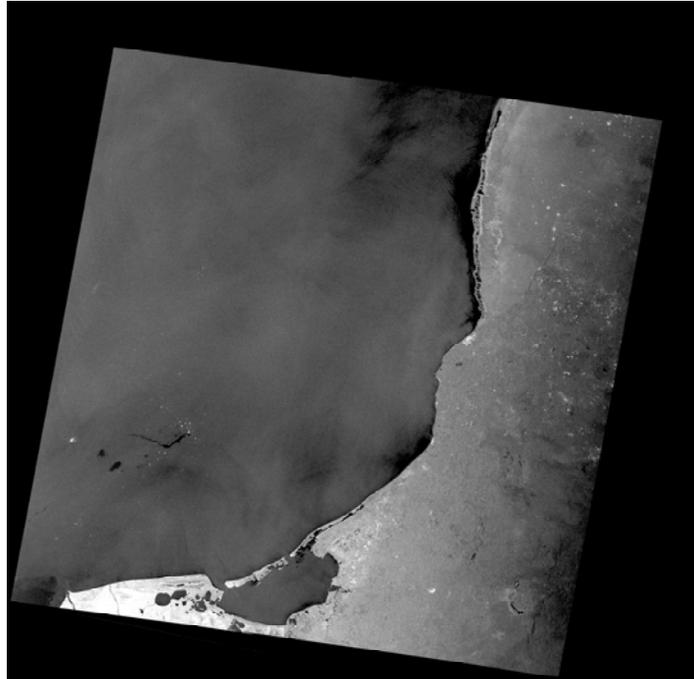


Figura 2.17 – Imagem do satélite RADARSAT-1 no modo SCN1 (órbita descendente), adquirida em 14 de julho de 2004 (Roriz, 2006).

CAPÍTULO 3

O PROJETO DE MONITORAMENTO SISTEMÁTICO POR SATÉLITE DA EXSUDAÇÃO PETROLÍFERA DE CANTARELL

Conforme já enfatizado no Capítulo 1, nas porções costeiras a sul e sudoeste do Golfo do México (território mexicano), ocorre uma das maiores províncias petrolíferas do mundo, na qual se inclui a proeminente exsudação de Cantarell. Essa região é altamente sensível à presença de óleo, devido a atividades pesqueiras intensas e bem desenvolvidas e ao ecossistema delicado, representado principalmente por manguezais. Em virtude da grande fragilidade ambiental, é de fundamental relevância a adoção de práticas apropriadas de monitoramento, como subsídio à gestão dos recursos naturais ali presentes.

Neste contexto, o entendimento da dinâmica temporal e distribuição espacial das exsudações e o acompanhamento dos vazamentos operacionais de óleo são de extrema importância para auxiliar na avaliação do risco ambiental na operação dos recursos petrolíferos do complexo de Cantarell. Para caracterização de feições associadas à presença de óleo nessa região, freqüentemente coberta por nuvens, é altamente recomendável a utilização de sensoriamento remoto utilizando radares de abertura sintética (SAR), tais como o RADARSAT-1. Por essa razão, no ano de 2000, a companhia estatal mexicana PEMEX Exploração e Produção (PEP) viabilizou um projeto monitoramento sistemático de exsudações e derramamentos de petróleo na Baía de Campeche, a sul do Golfo do México, através de imagens do citado satélite. Esse estudo foi realizado pelo Laboratório de Sensoriamento Remoto por Radar Aplicado à Indústria do Petróleo (LABSAR), situado na COPPE/UFRJ, em parceria com a empresa canadense RADARSAT *Internacional Inc.* (RSI) e com a PEP.

A seleção das imagens RADARSAT-1 de resolução plena para esse estudo foi auxiliada pela análise de informações meteo-oceanográficas adquiridas próximo ao momento de obtenção dos dados SAR. Assim, em 2000, foram processadas 12 (doze) imagens RADARSAT-1 de resolução plena no modo de imageamento W1, a partir de 36 possibilidades de produtos preliminares com resolução degradada (*quicklooks*). A exsudação de Cantarell aparece ativa em 9 (nove) delas, ou seja, em 92% das imagens adquiridas (Figura 3.1). Essa etapa do projeto demonstrou a capacidade da

metodologia adotada para detecção de óleo na superfície do mar. No ano de 2001, o trabalho foi realizado em caráter pré-operacional. Essa etapa foi importante para a adequação da metodologia às demandas operacionais da PEMEX. Neste ano, foram obtidas 20 (vinte) imagens RADARSAT-1, nos modos SCN1, W1 e W2, a partir de 60 *quicklooks*. Na ocasião foi identificada atividade em Cantarell em 19 (dezenove) produtos (Figura 3.2). Após grande êxito das duas primeiras etapas, entre os anos de 2002 e 2006, o projeto foi realizado em caráter operacional, dividido em dois módulos. O primeiro orientado para o monitoramento de derrames e exsudações de óleo; o segundo visou ao aprofundamento técnico-científico da metodologia empregada. Tal estudo contribuiu para eximir a PEP de multas por possíveis vazamentos de óleo na área da exsudação de Cantarell, demonstrando que o grande volume de óleo no mar esta associado a processos atuais. O projeto foi renovado em seguida, para o período de 2007 a 2011 (Tabela 3.1).

Tabela 3.1 – Síntese das fases do projeto de monitoramento de exsudações e derrames operacionais de óleo no Golfo do México, Baía de Campeche.

Título do projeto	Fase	Período de execução	Modos de Imageamento	N° de imagens RADARSAT-1
<i>Diseño de un sistema local de detección de emanaciones naturales y accidentales de hidrocarburos empleando imágenes de radar del satélite RADARSAT-1, en la Región Marina Noreste</i>	piloto	2000	W1	12 imagens
<i>Programa regional para la detección de emanaciones naturales y derrames de petróleo utilizando imágenes del satélite RADARSAT-1 en el Golfo de México</i>	pré-operacional	2001	SCN1, W1 e W2	20 imagens
<i>Monitoreo de emanaciones naturales y derrames de petróleo a través de imágenes del satélite RADARSAT-1 en el Golfo do México</i>	operacional	2002 até 2006	EXL1, SCN1, SNB e W1	276 imagens
<i>Monitoreo de emanaciones naturales y derrames de petróleo a través de imágenes del satélite RADARSAT-1 en el Golfo do México</i>	operacional (renovação)	2007 até 2011	EXL1, SCN1, SNB e W1	Previsão de processamento de 332 imagens

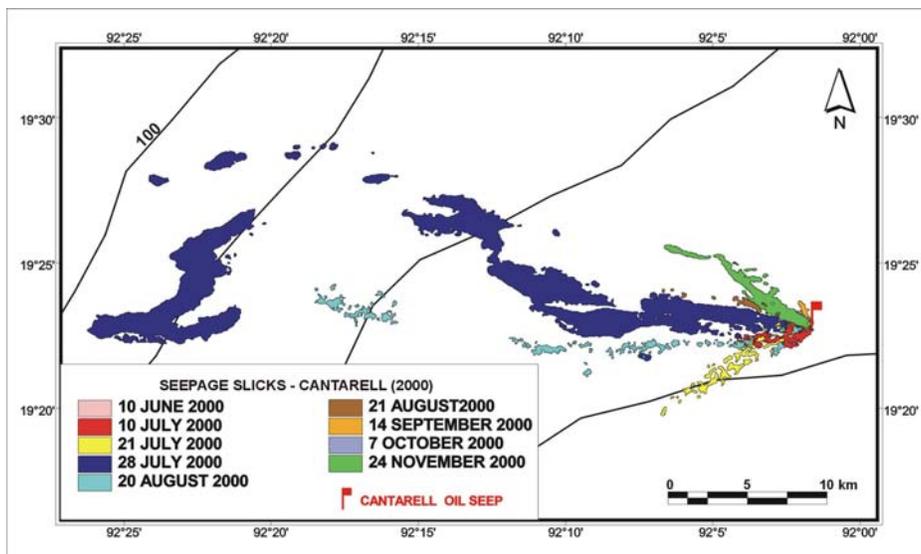


Figura 3.1 – Mapa temático das exsudações de Cantarell, confeccionado a partir da composição de 9 (nove) imagens do satélite RADARSAT-1, no modo W1, adquiridas durante o ano de 2000 (Miranda *et al.*, 2004). As curvas batimétricas estão expressas em metros.

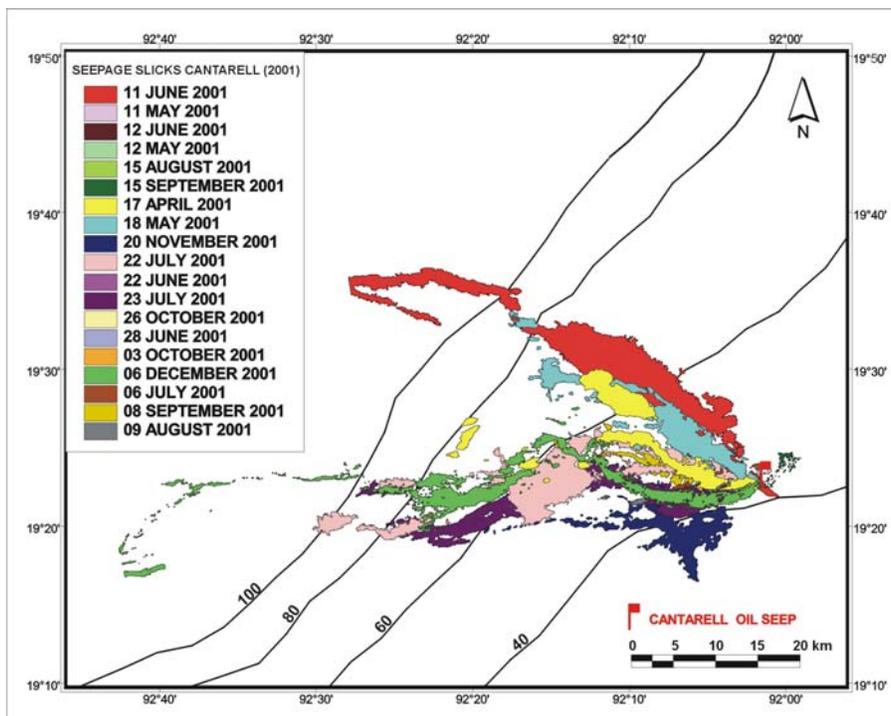


Figura 3.2 – Mapa temático das exsudações de Cantarell, confeccionado a partir da composição de 19 (nove) imagens do satélite RADARSAT-1, nos modos de imageamento SCN1, W1 e W2, adquiridas durante o ano de 2001 (Miranda *et al.*, 2004). As curvas batimétricas estão expressas em metros.

O algoritmo de classificação textural não supervisionado, denominado USTC (*Unsupervised Semivariogram Textural Classifier*), utilizado por Miranda *et al.* (2004), consiste em um classificador determinístico, que extrai informações radiométricas e texturais das imagens SAR, realçando as características da superfície. Esse método viabiliza a discriminação das feições correspondentes às manchas de óleo presentes nas imagens RADARSAT-1. O resultado da classificação é um bitmap, onde a cor azul clara está relacionada à superfície rugosa do mar e áreas de forte retorno nas plataformas e navios, enquanto que a cor vermelha está associada a manchas de óleo e a outras feições de baixo retorno de radar presentes na superfície do mar (Figura 3.3).

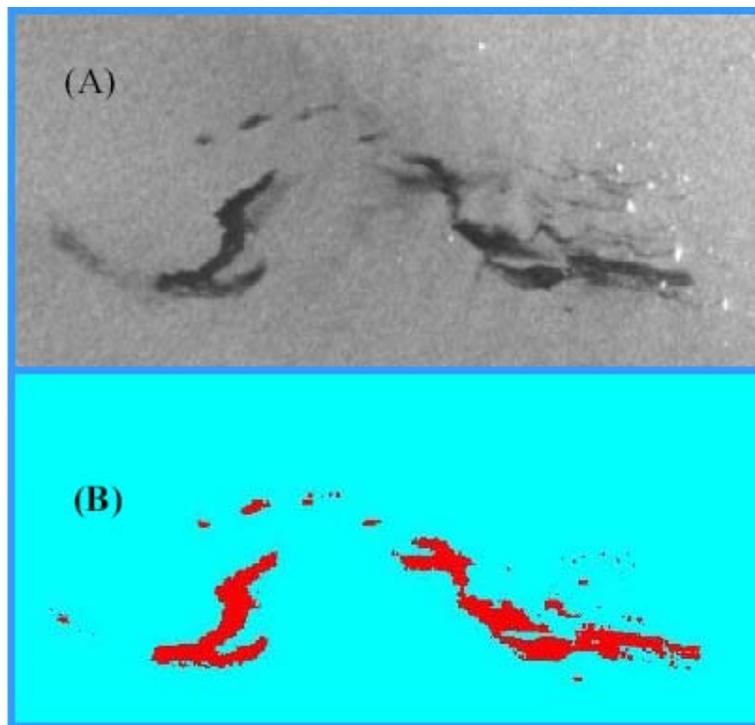


Figura 3.3 – (A) imagem RADARSAT-1 no modo de operação W1, adquirida em julho de 2000; (B) polígono resultante da classificação USTC para a exsudação de Cantarell (Mendoza *et al.*, 2003).

3.1 – Aquisição e processamento digital das imagens RADARSAT-1

As etapas aqui descritas foram propostas e realizadas pelo LABSAR, geralmente com periodicidade semanal. Os resultados compõem o relatório entregue à PEMEX com periodicidade quadrimestral, como parte do projeto operacional realizado nos anos de 2002 a 2006.

As imagens RADARSAT-1 foram selecionadas para processamento digital a partir de produtos com resolução degradada, denominados *quicklooks*. Estes foram submetidos à análise visual, objetivando selecionar os que contivessem um maior número de feições interpretadas como possíveis exsudações ou vazamentos operacionais de óleo. Dados meteorológicos e oceanográficos também foram considerados nessa escolha. Somente após a conclusão da análise dos *quicklooks*, foram solicitadas imagens com resolução plena. Tais produtos foram analisados com auxílio de técnicas de processamento digital de imagens, notadamente do algoritmo de classificação textural USTC, como suporte ao conhecimento dos especialistas. O resultado desses processamentos são polígonos, cujos atributos compõem o banco de dados em estudo (inclusive os dados meteo-oceanográficos), os quais refletem o esforço de monitoramento de exsudações e vazamentos operacionais de óleo na Baía de Campeche, Golfo do México (Figura 3.4). A origem de todos esses polígonos foi validada posteriormente pela empresa estatal mexicana PEMEX; eles são, portanto, considerados como expressão da verdade.



Figura 3.4 – Esquema das principais etapas propostas e realizadas pelo LABSAR para aquisição e processamento digital das imagens RADARSAT-1.

Etapas da construção da base de dados:

- Aquisição dos produtos com resolução degradada (*quicklooks*) do satélite RADARSAT-1, nos modos de operação SCN1, W1, W2, SNB, EXL1, FR2 e SCWB;
- Análise visual dos *quicklooks*, objetivando selecionar aquele que contenha o maior número de feições interpretadas como possíveis exsudações ou vazamentos de óleo, bem como os que apresentem visualmente condições meteo-oceanográficas mais favoráveis;
- Escolha da imagem RADARSAT-1 de resolução plena, com base nos resultados do item anterior;
- Obtenção das informações meteo-oceanográficas mais próximas da data de aquisição da imagem RADARSAT-1 selecionada. Essas informações compreendem os seguintes mapas: temperatura de superfície do mar, temperatura de topo de nuvem, velocidade do vento, altura da onda e concentração de clorofila-a;
- Processamento digital da imagem RADARSAT-1 de resolução plena:
 - No caso de imagens FR2, W1, W2 e EXL1, reescalamento dos números digitais de 16 bits para 8 bits;
 - Filtragem para a redução do ruído *speckle* e realce de contraste;
 - Classificação através do método USTC;
 - Ajuste da PCT (*PseudoColor Table*) mais adequada e agregação das classes obtidas no item anterior para a delimitação de polígonos no formato raster representando superfícies lisas (*targets*);
 - Vetorização automática dos polígonos no formato raster através da utilização de um algoritmo específico e posterior discriminação das feições interpretadas como exsudações de óleo ou vazamentos operacionais;
- Integração das informações geológicas, geofísicas e batimétricas com os dados de sensoriamento remoto em um banco de dados digitais geo-referenciados residente em ambiente SIG (Sistema de Informações Geográficas);
- Hierarquização das exsudações de óleo, com o objetivo de obter o fator de confiabilidade geral, a partir de critérios tectônicos, temporais e ambientais;
- Identificação de agrupamentos (*clusters*) de polígonos interpretados como exsudações petrolíferas com origem comum na superfície do mar;
- Validação com a PEMEX da origem dos vazamentos operacionais interpretados como relacionados a instalações conhecidas de produção e transporte *offshore*;

- Construção de uma tabela no formato .xls (Excel) contendo todas as variáveis da base de dados a serem consideradas no processo de descoberta de conhecimento.

Materiais utilizados para interpretação dos polígonos:

- *Quicklooks* e Imagens de resolução plena do satélite RADARSAT-1;
- Softwares de geoprocessamento: Geomatica PCI 10.0.3 e ARCVIEW GIS 9.2;
- Método de classificação USTC (*Unsupervised Semivariogram Textural Classifier*), utilizado por Miranda *et al.* (2004);
- As informações de TSM são obtidas a partir dos dados do sensor *Advanced Very High Resolution Radiometer (AVHRR)*, a bordo dos satélites da *National Oceanic and Atmospheric Administration (NOAA)*;
- Para a temperatura de topo de nuvem, o sensor utilizado é o *Imager*, a bordo do *Geostationary Operational Environmental Satellite (GOES-8)*, assim como o AVHRR da série NOAA;
- Para a obtenção de dados de vento, o sensor utilizado foi o *Seawinds*, a bordo do satélite *Quickscat*;
- Para a altura de onda, utilizou-se uma combinação dos dados fornecidos pelos radares altímetros dos satélites TOPEX/POSEIDON e *European Remote Sensing Satellite (ERS-2)*;
- Para a aquisição de informações sobre a concentração de clorofila-a, utilizou-se o sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*), instrumento a bordo do satélite ACQUA.

CAPÍTULO 4

DESCOBERTA DE CONHECIMENTO E REGRAS DE ASSOCIAÇÃO

As regras de associação descrevem padrões de relacionamento entre itens de uma base de dados. Uma de suas aplicações típicas é uma análise de transações de compras efetuadas por famílias em supermercados (Gonçalves, 2005). Na presente dissertação, pretende-se utilizar essa metodologia na solução de um problema científico, ou seja, na identificação de regras de associação interessantes em uma base de dados sobre exsudações de óleo no Golfo do México.

4.1 – O processo de descoberta de conhecimento

É crescente o interesse em pesquisas que apóiam o desenvolvimento de tecnologias de extração de conhecimento a partir de dados. As utilizações de métodos essencialmente manuais tornam essa tarefa geralmente lenta, subjetiva e financeiramente dispendiosa, praticamente inviável quando aplicada a um grande volume de dados. Devido à grande disponibilidade de dados armazenados nos meios computacionais das organizações, faz-se necessário o desenvolvimento de tecnologias automáticas e inteligentes que explorem o grande potencial das informações para explicitar conhecimento contido nesses dados. O campo de pesquisa que trata desse assunto é denominado KDD (do inglês *Knowledge Discovery in Databases*), ou seja, Descoberta de Conhecimento em Bases de Dados. A tecnologia incorporada pelo KDD auxilia na tarefa de analisar, interpretar e apresentar o conhecimento implícito em grandes volumes de dados, tornando claros e compressíveis os padrões neles embebidos.

O processo de KDD tem por objetivo a identificação de padrões interessantes, previamente desconhecidos, que revele conhecimento útil ao domínio da aplicação em um formato de fácil interpretação (Fayyad *et al.*, 1996a). As áreas de aplicação do KDD são inúmeras, tanto no campo acadêmico como no corporativo (Han e Kamber, 2001). O conhecimento descoberto através desse processo pode indicar riscos de negócio a evitar e oportunidades a serem aproveitadas, trazendo potenciais vantagens competitivas às organizações.

O KDD é um campo de pesquisa multidisciplinar que incorpora técnicas utilizadas em diversas áreas de conhecimento como Aprendizado de Máquina, Inteligência Artificial e Estatística. As técnicas utilizadas por esse processo podem complementar os resultados de outras ferramentas de análise de dados, tais como, *Data Warehousing* e *Online Analytical Processing (OLAP)*.

A literatura atual oferece várias definições para KDD e Mineração de Dados, termos esses que alguns autores consideram equivalentes. No presente trabalho, o termo KDD, é considerado o processo completo de descoberta de conhecimento, sendo a mineração de dados uma de suas etapas.

Segundo Fayyad *et al.* (1996a), KDD é definido como o "processo não trivial de várias etapas, interativas e iterativas, que objetivam a identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grande quantidade de dados". O processo é composto de vários passos envolvendo a definição do escopo da aplicação, a preparação do conjunto de dados, a busca por padrões através de técnicas de mineração, bem como avaliação desses padrões e sua consolidação na forma de conhecimento. A expressão não trivial alerta para a complexidade normalmente presente na execução do processo, que é considerado interativo ou semi-automático, pois depende da interação do homem como responsável pelo seu controle. Por sua vez, o termo iterativo refere-se à possibilidade de repetição integral ou parcial de todo o processo.

De acordo com Rezende *et al.* (2003), o processo de KDD pode ser agrupado em um ciclo com três grandes etapas: pré-processamento, extração de padrões e pós-processamento (Figura 4.1). Além dessas etapas, os autores também consideram uma fase anterior ao ciclo, referente à identificação do problema, e outra posterior, referente à utilização do conhecimento obtido. Na Tabela 4.1, é apresentada resumidamente a descrição das fases e etapas envolvidas em aplicações na área de KDD. As etapas do processo possuem um fluxo de seqüências bem ordenadas, fato que pode sugerir erroneamente que exista uma trajetória linear. Em uma dada etapa, podem ser detectados problemas ocorridos em alguma etapa anterior. Portanto, os resultados de uma determinada etapa podem acarretar mudanças em quaisquer das etapas anteriores ou, ainda, o recomeço de todo o processo (Fayyad *et al.*, 1996b). Essa busca por refinamentos sucessivos faz do KDD um processo eminentemente iterativo.

Dentre as tarefas de extração de conhecimento, as que mais se destacam são: classificação, regressão, regras de associação, *sumarização* e *clustering*. A escolha da tarefa depende do cenário, problema e resultado que se deseja obter. Segundo Fayyad *et al.* (1996a), essas tarefas podem ser agrupadas em atividades preditivas e descritivas. A atividade preditiva, ou supervisionada, envolve a utilização de algumas variáveis da base de dados para prever valores desconhecidos ou futuros de outras variáveis de interesse (atributo meta ou classe). As principais tarefas de predição são classificação e regressão. Já a atividade descritiva, ou não-supervisionada, procura por padrões que descrevam o comportamento intrínseco existente nos dados. As principais tarefas descritivas são regras de associação, *clustering* e *sumarização*. A tarefa de regras de associação é o objeto de estudo deste trabalho, o que justifica sua descrição detalhada na seção 4.2.



Figura 4.1 – Etapas do processo de Descoberta de Conhecimento em Base de Dados (Rezende *et. al.*, 2003).

Tabela 4.1 – Resumo das fases e etapas envolvidas pelo processo de KDD.

ETAPA	DESCRIÇÃO
Identificação do problema	Esta fase envolve a compreensão do domínio da aplicação e a definição de objetivos e metas a serem alcançados com o processo, bem como a seleção dos conjuntos de dados a serem utilizados durante o processo.

Pré-processamento	Etapa de integração e limpeza dos dados, onde se resolvem problemas como os de redundância, inconsistência e ausência de valores. Pode-se realizar, também, a redução de dados, permitindo um número menor de variáveis. Esta etapa tem por finalidade a adaptação dos dados para etapa seguinte.
Extração de padrões	Nesta etapa, é realizada a escolha, configuração e execução de algoritmo(s) e ou aplicativo(s) de extração de padrões (para por exemplo, a geração de regras de associação).
Pós-processamento	Simplificação e avaliação dos novos padrões extraídos, identificando os realmente interessantes ao usuário final.
Utilização do conhecimento	Análise de como o conhecimento obtido pode ser utilizado pelo usuário final de forma a auxiliar no processo de tomada de decisão.

A etapa de extração de padrões, apesar de importante, representa menos de 10% do esforço total de uma aplicação de KDD (Brachman e Anand, 1996). O gargalo do processo de KDD está, portanto, nas etapas de pré-processamento e pós-processamento, as quais representam a maior parte do tempo empregado pelo analista em sua realização.

O sucesso do processo de extração de padrões depende em grande parte da qualidade da interação entre as diferentes categorias de usuários. Os usuários do processo podem ser agrupados da seguinte forma:

- **Especialista do domínio** – Usuário que conhece profundamente o domínio da aplicação onde será realizado o processo de KDD. Este usuário fornece suporte ao analista em todas as etapas do processo, pois sua participação é imprescindível na identificação do conhecimento (padrão) considerado interessante na etapa de pós-processamento;
- **Analista** – Usuário especialista em KDD responsável por coordenar as ações do processo em cada etapa. É o executor do processo e, para tanto, interage ativamente com o especialista do domínio da aplicação e o usuário final;
- **Usuário final** – Usuário ou grupo de usuários que se beneficiarão do conhecimento extraído pelo processo como apoio à tomada de decisão. Este usuário geralmente tem maior atuação na etapa de utilização do conhecimento, mas também pode atuar na definição dos objetivos e avaliação do conhecimento extraído.

O usuário final não necessariamente precisa ter conhecimento aprofundado sobre todo o domínio da aplicação. Já o analista deve ter algum grau de conhecimento, mesmo que geral, do domínio da aplicação para poder direcionar de forma bem sucedida, o processo de KDD. Cabe ressaltar que a atuação pode se dar em mais de uma classe, como é o caso do especialista do domínio também ser o usuário da final.

4.2 – Regras de associação

As regras de associação, introduzidas por Agrawal *et al.* (1993), são consideradas uma das mais populares tarefas de mineração de dados. Os motivos dessa preferência estão relacionados à sua grande aplicabilidade em inúmeros problemas no mundo dos negócios e ao fato das regras de associação serem de fácil compreensão, mesmo por aqueles usuários não experientes em mineração de dados (Hipp *et al.*, 2000). É uma atividade de mineração de dados descritiva, que consiste em encontrar padrões que explicitem dependências significativas entre eventos que ocorrem juntos (Agrawal *et al.*, 1993).

Os padrões obtidos podem se apresentar em dois aspectos: o nível estrutural, onde são especificadas variáveis localmente dependentes de outras, e o nível quantitativo, que especifica a força das dependências, utilizando algum critério numérico (Fayyad *et al.*, 1996a). Portanto, regras de associação fornecem um modo conveniente para identificar e representar dependências entre atributos em uma base de dados. Uma regra determina o quanto a presença de um conjunto de itens (atributos) nas transações (registros) de uma base de dados implica na presença de outro conjunto distinto de itens nas mesmas transações.

As associações são geralmente apresentadas em forma de regras do tipo “se X então Y ”, ou simplesmente, $X \rightarrow Y$. Esta implicação representa que, quando uma transação contém o conjunto de itens X , então provavelmente também contém Y . Os elementos X e Y são itens ou conjuntos de itens (*itemset*) distintos que representam, respectivamente, o antecedente e o conseqüente da regra. Seja I um conjunto de itens da base de dados, a regra de associação é uma implicação $X \rightarrow Y$, em que $X \subset I$, $Y \subset I$ e $X \cap Y = \emptyset$ (Agrawal e Srikant, 1994).

Tanto o antecedente quanto o conseqüente de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens. Um *itemset* é definido com um conjunto de itens; e a quantidade de itens pertencentes a um *itemset* é denominada comprimento do conjunto. Assim, *k-itemset* é um conjunto de itens de comprimento *k*. O procedimento para minerar as regras de associação em uma base de dados consiste em encontrar todos os *k-itemset* freqüentes presentes em um conjunto de transações, e então definir as regras de associação que satisfaçam tanto um limiar de suporte mínimo e uma confiança mínima (Agrawal *et al.*, 1993).

Os parâmetros de suporte e confiança correspondem às medidas objetivas básicas de significância de uma regra de associação (variam de 0% a 100%). Alterações em seus valores significam incluir ou excluir novas regras ao conjunto de regras resultante. Essas medidas são as mais empregadas tanto na avaliação do conhecimento na etapa de pós-processamento como na seleção de *itemsets* durante o processo de geração das regras. Esta última abordagem também é conhecida como modelo suporte/confiança. A utilização desse modelo restringe o universo de regras geradas, tendo como conseqüência um menor esforço computacional no tempo de processamento dos algoritmos. Quanto maior o percentual de confiança e suporte definidos pelo usuário, menor e mais específico será o número de regras extraídas. Não existe um valor ótimo para confiança e suporte mínimos; ele vai depender do problema que está sendo analisado e do objetivo a ser atingido.

Uma regra de associação é considerada forte quando satisfaz um suporte mínimo e uma confiança mínima (Han e Kamber, 2001). O suporte indica a porcentagem de ocorrências da associação dentre o montante total de registros e a confiança indica a porcentagem de todas as ocorrências do antecedente onde o item conseqüente está associado (Levy, 1999). As definições e equações apresentadas abaixo facilitam a compreensão dessas duas medidas.

Suporte: quantifica a freqüência de um item, conjunto de itens ou de uma regra na base de dados. O **suporte de um item ou conjunto de itens *Z*** representa a porcentagem de transações da base de dados que contem *Z*. Seja $n(Z)$ o número de transações nas quais *Z* ocorre e *N* o número total de transações considerada. A equação do suporte do item pode ser representada por:

$$\text{sup}(Z) = \frac{n(Z)}{N} \times 100 . \quad \text{Equação 4.1}$$

O **suporte da regra** corresponde à sua relevância estatística, representando a probabilidade percentual de ocorrer X e Y em simultâneo ($X \cup Y$). A equação do suporte da regra é dada por:

$$\text{sup}(X \rightarrow Y) = \frac{n(X \cup Y)}{N} \times 100 . \quad \text{Equação 4.2}$$

Confiança: indica a percentagem de transações que contêm X em conjunto com Y , em relação ao número total de transações que contêm X . A confiança mede a validade da implicação, pode ser representada pela seguinte equação:

$$\text{conf}(X \rightarrow Y) = \frac{n(X \cup Y)}{n(X)} \times 100 . \quad \text{Equação 4.3}$$

Portanto, usando como parâmetro de entrada um valor de suporte igual a 5%, a ferramenta de geração de regras de associação apresentará apenas 5% de todos os itens sob análise que apareçam juntos com maior frequência. Da mesma forma, um nível de 50% de confiança estabelece esse grau de garantia para a associação dos itens. A partir dos valores mínimos de suporte (*Sup_min*) e confiança (*Conf_min*) especificados pelo usuário, os procedimentos para geração de regras de associação podem ser divididos em duas etapas (Agrawal *et al.*, 1993);

1. Encontrar todos os *k-itemsets* (conjunto de k itens) que possuem suporte maior ou igual ao *Sup_min*, denominados *k-itemsets* frequentes.
2. Utilizar os conjuntos frequentes de tamanho maior ou igual a dois para gerar todas as regras que tenham confiança maior ou igual à *Conf_min*.

A primeira etapa é computacionalmente intensiva, devido ao número exponencial de possíveis conjuntos frequentes em relação ao número total de itens. Os algoritmos utilizados nessa etapa efetuam múltiplas passagens pelos dados. Muitos algoritmos foram propostos na literatura visando ao aprimoramento no desempenho da identificação e contagem dos *itemsets*, dentre eles: *AIS* (Agrawal *et al.*, 1993), *Apriori* e *AprioriTid* (Agrawal e Srikant, 1994), *SETM* (Houtsma e Swami, 1995), *Partition* (Savasare *et al.*, 1995), *DHP* – Direct Hashing and Pruning (Park, *et al.*, 1995), *Opus* (Webb, 1995), *ECLAT* (Zaki *et al.*, 1997), *DIC* – *Dynamic Set Counting* (Brin *et al.*, 1997), *Closet* (Pei *et al.*, 2000), *FP-Growth*, (Han *et al.*, 2000; Wang *et al.*,

2002), *Charm* (Zaki e Hsiao, 2002). Os estudos de Hipp *et al.* (2000) e Mannila *et al.* (1994) apresentam um exame dos principais algoritmos e uma comparação de seu desempenho. Contudo, ainda que diferentes, esses algoritmos teoricamente devem sempre gerar resultados idênticos para um mesmo conjunto de entrada: suporte mínimo, confiança mínima e base de dados (Zeng *et al.*, 2001).

Dentre os principais algoritmos para identificação e contagem dos *itemsets*, destaca-se o algoritmo *Apriori*, proposto por Agrawal e Srikant (1994), pelo pioneirismo e por ter sido base de muitas outras propostas subseqüentes. Esse algoritmo foi o primeiro a reduzir significativamente o esforço necessário para identificação dos conjuntos freqüentes.

Para encontrar todos os *k-itemsets* freqüentes contidos em uma base de dados, o algoritmo *Apriori* gera um conjunto de *k-itemsets* candidatos e então percorre a base de dados para determinar se os mesmos são freqüentes. Durante a primeira passagem, o algoritmo procura por todos os conjuntos de itens freqüentes com um (1) item (*itemset* freqüente de comprimento igual a um), denominados L_1 . O L_1 é empregado para encontrar o L_2 (*itemset* freqüente de comprimento igual a dois), que é utilizado para achar o L_3 e assim sucessivamente, até que nenhum *k-itemset* possa ser mais encontrado. Em cada passo, os conjuntos freqüentes da interação anterior são utilizados na geração dos *itemsets* candidatos e a base de dados é percorrida para sua contagem. Em seguida, consideram-se somente os *itemsets* candidatos efetivamente freqüentes, ou seja, aqueles que possuem suporte maior ou igual ao *Sup_min* especificado. Finalmente, a saída do algoritmo é composta pela união dos conjuntos L_k de *k-itemsets* freqüentes. Esta solução é utilizada como entrada de um algoritmo que gera regras de associação. O aspecto fundamental do *Apriori* é que qualquer subconjunto de um *itemset* freqüente deve também ser freqüente.

É na segunda etapa do procedimento que as regras de associação são geradas. Um dos algoritmos mais simples com essa finalidade foi proposto por Agrawal e Srikant (1994). Esse algoritmo é executado para *k-itemsets* freqüentes, com $k \geq 2$. Primeiramente, são gerados os subconjuntos não vazios de um *itemset* freqüente. Logo depois, são geradas as regras utilizando os subconjuntos criados, sendo consideradas somente aquelas que possuem confiança maior ou igual à confiança mínima especificada pelo usuário (*Conf_min*).

O Número de regras de associação resultante desse processo é geralmente muito grande, mesmo quando a técnica é aplicada em bases de dados relativamente pequenas (Melanda, 2005). Essa grande quantidade de regras dificulta a identificação daquelas que representam, para o usuário final, conhecimento útil, interessante e/ou inovador, tornando o pós-processamento uma etapa desafiante. Fica evidente, portanto, a necessidade da utilização de métodos para evitar a produção de muitas regras e para selecionar as que se apresentam mais interessantes para os usuários finais. Neste contexto, através do emprego de medidas de interesse objetivas de avaliação de conhecimento na etapa de pós-processamento, é possível analisar e avaliar as regras sob vários aspectos. Estas análises se apóiam na identificação de regras realmente interessantes dentre o grande volume daquelas geradas.

4.2.1 – Medidas de interesse objetivas

A aplicação de medidas de avaliação do conhecimento é uma das principais técnicas de suporte para se apreciar as regras obtidas no processo anteriormente descrito. A aplicação dessas medidas ocorre nas etapas de extração de padrões e pós-processamento (identificação de regras interessantes ou relevantes), permitindo a avaliação da qualidade e do desempenho de uma regra auxiliando na identificação daquelas que são de fato relevantes e úteis dentre as muitas que podem ser mineradas. A avaliação da qualidade está relacionada ao grau de interesse e de compreensibilidade, enquanto a avaliação do desempenho quantifica a fidelidade com que as regras representam os dados.

As medidas de avaliação podem ser classificadas como objetivas ou subjetivas. As medidas objetivas identificam estatisticamente a força da regra, visto que consideram apenas os dados e a estrutura dos padrões. Para as medidas subjetivas, a determinação da força da regra considera sua estrutura e a da base de dados, assim como o conhecimento do especialista do domínio. Segundo Sinoara (2006), a combinação de medidas de avaliação do conhecimento permite a identificação de regras de associação interessantes ao especialista de maneira mais eficiente do que o uso de apenas um tipo de medida. Portanto, devido à dificuldade de captação do conhecimento subjetivo do usuário, optou-se nesse estudo pelo uso de algumas medidas objetivas, subsidiadas pela interpretação do especialista do domínio da aplicação para seleção das regras interessantes.

Medidas objetivas são independentes do domínio da aplicação e dependem exclusivamente da estrutura dos padrões e dos dados utilizados no processo de extração de conhecimento (Silberschatz e Tuzhilin, 1996). As medidas de interesse objetivas são índices estatísticos utilizados para selecionar regras interessantes dentre as muitas que podem ser descobertas por um algoritmo de mineração de regras de associação. A utilização dessas medidas como filtro na etapa de extração de padrões e pós-processamento, além de reduzir o número de regras, facilita sua interpretação por parte do especialista. Trabalhos que sintetizam algumas medidas objetivas de interesse foram publicados por Hilderman e Hamilton (1999), Tan *et al.* (2002) e Melanda (2005).

O suporte e a confiança são exemplos de medidas de interesse objetivas. No entanto, diversas outras medidas propostas na literatura objetivam avaliar a qualidade, medir a dependência entre os itens, como também o desempenho e interessabilidade da regra de associação. Vale ressaltar que somente a significância estatística não garante que uma regra é interessante. Nesse contexto, a participação do especialista do domínio da aplicação é imprescindível no fornecimento de conhecimento e demandas específicas, que determinam a identificação daquelas que são realmente interessantes.

Neste estudo propõe-se o uso conjunto dos índices objetivos de suporte, confiança e *lift* (Brin *et al.*, 1997) no processo de pós-processamento de regras de associação. A utilização destas medidas permite que usuários possam realizar análises alternativas sobre uma mesma regra, pois cada uma das medidas é capaz de destacar uma característica a respeito da associação (Gonçalves, 2005).

A medida de interesse *lift* (Brin *et al.*, 1997), também conhecida como *interest*, é uma das mais utilizadas na avaliação de dependência. O *lift* é a razão entre confiança da regra e o suporte do conseqüente, indicando o quanto mais freqüente torna-se conseqüente quando o antecedente ocorre. Dada a regra $A \rightarrow B$, o valor do *lift* é calculado por:

$$lift(A \rightarrow B) = \frac{Conf(A \rightarrow B)}{Sup(B)} . \quad \text{Equação 4.4}$$

Esta medida varia entre 0 e ∞ , possuindo interpretação bastante simples: quanto maior o valor do *lift*, mais interessante é a regra, pois A aumentou (“*lifted*”) B

numa maior taxa (Gonçalves, 2005). São abaixo apresentados os intervalos de valores possíveis do *lift* em relação ao grau de dependência entre os itens de uma regra:

- Se $lift(A \rightarrow B) = 1$, então A e B são independentes;
- Se $lift(A \rightarrow B) > 1$, então A e B são positivamente dependentes;
- Se $lift(A \rightarrow B) < 1$, então A e B são negativamente dependentes.

O índice *lift* consegue destacar com facilidade a dependência positiva entre conjuntos de itens que possuem suporte baixo, sendo simétrico em relação aos lados da regra, ou seja, $lift(A \rightarrow B) = lift(B \rightarrow A)$.

4.2.2 – Medidas de interesse subjetivas

As medidas de interesse objetivas identificam quantitativamente a força das regras de associação. No entanto, uma regra pode possuir valores altos para determinadas medidas objetivas e não ser subjetivamente interessante para o especialista que as examina. Desse modo, medidas de interesse subjetivas qualificam o interesse de uma regra de associação para um determinado usuário (Gonçalves, 2005).

4.2.3 – Aplicativo CBA na extração de regras

O desenvolvimento de ferramentas de Mineração de Dados tem como objetivo principal fornecer aos tomadores de decisão das organizações, que são usuários geralmente não especialistas nessas técnicas abordagens intuitivas e amigáveis (Sinoara, 2006). O aplicativo de extração de regras de associação CBA (*Classification Based Association*), desenvolvido por Liu *et al.* (1998), gera regras baseadas no algoritmo *Apriori*. Por constituir uma tarefa descritiva de mineração de dados, o conjunto de regras gerado pelo CBA deve ser replicado no caso da utilização de qualquer outro aplicativo com a mesma finalidade. Esse programa é de domínio público, podendo ser obtido pela internet gratuitamente.

O CBA utiliza dois arquivos de entrada: o arquivo de dados e o arquivo de definições. O primeiro apresenta a extensão *.DATA, onde os dados selecionados para a aplicação devem estar limpos e compostos somente por valores discretos. Esse aplicativo utiliza bases de dados do tipo itens-transações, onde cada linha representa

uma transação composta de vários itens, não havendo restrição quanto à quantidade de registros e variáveis. Já o segundo arquivo possui a extensão *.NAMES e contém o nome dos atributos e de seus possíveis valores. Ambos os arquivos são definidos no formato texto, devendo possuir o mesmo nome antes da extensão.

As opções de configuração padrão podem ser alteradas pelo usuário, tais como a definição de valores mínimos de suporte e confiança o número limite de regras a serem geradas e o número de condições (valor mínimo igual a 2). Após a identificação dos arquivos de entrada com extensão *.DATA e *.NAMES, da escolha por “Tabela de Regras de Associação” (*Table Assoc Rules*) na tela principal e da definição das opções de configuração (*Mine: Single Sup*), as regras são apresentadas pelo aplicativo.

O programa CBA possui vários arquivos de saída; porém, destacam-se os arquivos com extensão *.ITM e *.ARL. O primeiro demonstra o percentual de ocorrência dos *k-itensets* gerados e o segundo é composto por todas as regras obtidas com seus respectivos valores de suporte e confiança.

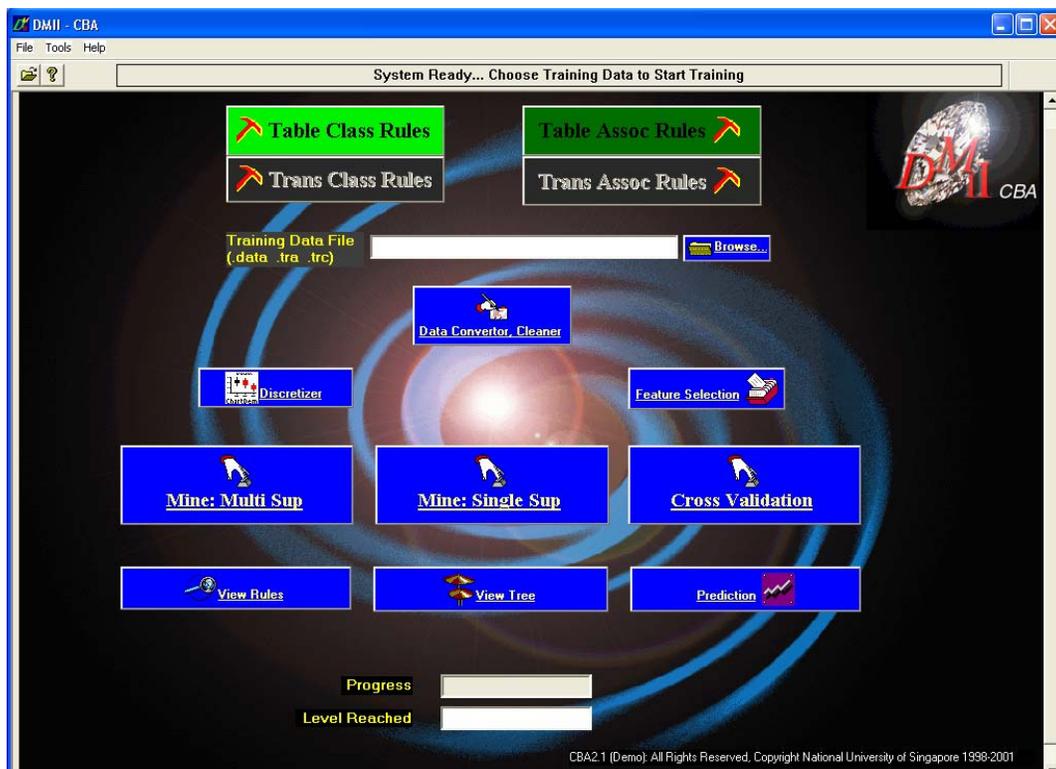


Figura 4.2 – Tela inicial do aplicativo CBA utilizado para extração das regras de associação (Liu *et al.*, 1998).

CAPÍTULO 5

RESULTADOS: PROCESSO DE KDD UTILIZANDO REGRAS DE ASSOCIAÇÃO

A metodologia proposta para este estudo fundamenta-se na estrutura do processo de KDD abordada por Resende *et al.* (2003), que tem início na fase de identificação do problema, seguida pelo ciclo de etapas: pré-processamento, extração de padrões (CBA) e pós-processamento (aplicação de medidas de interesse objetivas) e utilização do conhecimento. O fluxo de realização das etapas é iterativo; logo, existe a possibilidade de repetições integrais ou parciais do processo na busca de resultados satisfatórios. O especialista do domínio conhece profundamente o assunto e o ambiente do contexto da aplicação onde é realizado o processo de KDD. Portanto, este usuário fornece suporte ao analista de KDD em todas as etapas do processo, o que torna sua participação ativa é de suma importância.

O LABSAR (Laboratório de Sensoriamento Remoto por Radar Aplicado à Indústria do Petróleo) é a unidade de pesquisa da COPPE/UFRJ que, dentre outras atividades, também realiza o monitoramento de exsudações e derramamentos de petróleo na Baía de Campeche, sul do Golfo do México, através de dados do satélite RADARSAT-1. Os resultados da análise dessas imagens de sensoriamento remoto alimentam uma rica base de dados, que é o objeto de estudo desse trabalho. Tal base de dados é composta por variáveis representativas de polígonos interpretados em imagens do satélite RADARSAT-1, obtidas, na região investigada, no período de janeiro de 2002 a dezembro de 2006. As áreas de textura lisa, classificadas como exsudações ou vazamentos operacionais de óleo, foram individualizadas em polígonos, através do uso do algoritmo USTC e da avaliação de especialistas. Como resultado desse procedimento, foram detectados 1.540 (mil quinhentos e quarenta) polígonos classificados como exsudações de óleo. Para cada um deles, foi obtido um conjunto de variáveis numéricas ou categóricas (classes), apresentando como principal característica a propriedade espaço-temporal. O aspecto temporal está configurado pela série histórica de imagens sistematicamente adquiridas na área de estudo, as quais são utilizadas como insumo para a interpretação dos polígonos que compõem o banco de dados.

O objetivo principal da mineração dessa base de dados é expressar relações e/ou dependências morfológicas, ambientais, climáticas, temporais e operacionais entre os polígonos interpretados como exsudações de óleo nas imagens do satélite RADARSAT-1 da Baía de Campeche, Golfo do México. Desta forma, pretende-se contribuir para a aquisição de conhecimentos inéditos e interessantes sobre o fenômeno, como subsídio à gestão dos recursos petrolíferos na região.

5.1 – Pré-processamento da base de dados

A referida base de dados não foi produzida com o intuito de ser usada em um estudo de mineração. Portanto, a etapa de pré-processamento, onde os dados são tratados e preparados para as etapas subseqüentes, foi realizada de maneira criteriosa, pois possui fundamental relevância no processo de descoberta de conhecimento.

A base de dados disponibilizada para mineração, fornecida em planilha Excel, não apresentou ausência de valores, mas requereu correção de alguns erros de digitação. Não foi necessário o uso de nenhuma técnica de limpeza e redução dos dados. A base possui 1.540 transações no formato atributo-valor, de forma que as linhas representam as transações e as colunas seus atributos (itens). Os atributos numéricos (inteiros e reais) sofreram redução de volume através do processo de discretização de valores, conforme detalhado abaixo. Durante esta etapa, o especialista do domínio forneceu os intervalos considerados (limites inferior e superior) e os nomeou. Como resultado do processo, a base passou a ter um total de 98 valores possíveis para os 18 atributos discretos.

No processo de discretização, ocorre a substituição de um atributo contínuo (inteiros ou reais) por um discreto, através do seu particionamento em intervalos. Assim, cada intervalo corresponde a um valor discreto do atributo (Glymour *et al.*, 1997). O método de discretização, portanto, reduz o número de atributos e, conseqüentemente, o espaço de busca. Os métodos de discretização podem ser classificados em supervisionados ou não-supervisionado, locais ou globais, e parametrizados ou não-parametrizados (Félix, 1998). Dentre as vantagens desse método, podem ser citados o melhor entendimento do conhecimento descoberto e a redução do tempo de processamento pelo algoritmo de extração das regras. Em Srikant e Agrawal (1996) propuseram o método de discretização para lidar com

atributos quantitativos, sendo que a dificuldade encontrada neste enquadramento é saber em quantos intervalos devem se dividir os dados. Contudo, detalhes importantes sobre as informações extraídas podem ser perdidos em razão do particionamento dos atributos, o que, conseqüentemente, reduz a qualidade do conhecimento descoberto.

No particionamento das variáveis desse trabalho, foram utilizados elementos lingüísticos ao invés de intervalos, objetivando facilidade na compreensão do resultado obtido. Como efeito, uma regra que utilize elementos lingüísticos é nitidamente mais compreensível que outra que trabalhe com comparações numéricas, já que grande parte do conhecimento humano está armazenada em forma lingüística (Guillaume, 2001). Após a aplicação do método de discretização dos atributos, os mesmos foram agrupados em níveis taxonômicos. Em seguida, a base foi adaptada para utilização do aplicativo CBA de extração de padrões.

5.1.1 – Níveis Taxonômicos e análise exploratória das exsudações de óleo

As taxonomias refletem uma visão, coletiva ou individual e arbitrária, de como os atributos podem ser hierarquicamente classificados. Eventualmente, múltiplas taxonomias podem estar presentes simultaneamente, refletindo a existência de diferentes pontos de vista ou a possibilidade de classificações distintas para o mesmo conjunto de atributos (Melanda, 2005). As regras com taxonomias são denominadas Regras de Associação Generalizadas e buscam identificar associações no mesmo nível de taxonomia ou em níveis diferentes (Skikant e Agrawal, 1996).

Todos os atributos discretizados foram agrupados em níveis taxonômicos (múltiplas camadas hierárquicas), conforme pode ser visualizado na Figura 5.1. O nível taxonômico superior, aqui chamado de Exsudações de Óleo, possui sete grupos: localização, contexto temporal, forma, batimetria do centróide do polígono, características do imageamento, agrupamento associado e condições meteo-oceanográficas. O nível médio, subgrupo do nível superior, e o nível inferior estão apresentados na tabela 5.1. A apresentação geral dos níveis taxonômicos da base de dados estudada é mostrada na Figura 5.2.

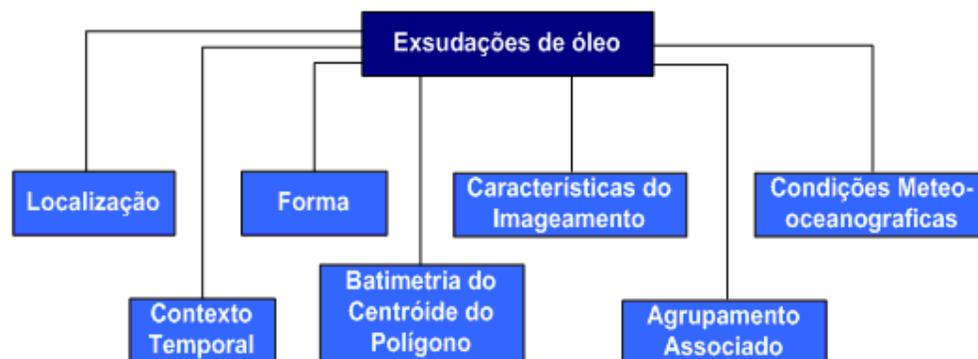


Figura 5.1 – Representação do nível taxonômico superior das exsudações de óleo.

Tabela 5.1 – Níveis taxonômicos superior, médio e inferior das exsudações de óleo na Baía de Campeche, Golfo do México.

SUPERIOR	MÉDIO	INFERIOR
Localização	Cantarell	-
	Outros	
Contexto temporal	Ano	2002, 2003, 2004, 2005, 2006
	Estação	inverno, primavera, verão, outono
Forma	Área	baixa, intermediária, alta
	Perímetro	
	Compactação	
Batimetria do centroide do polígono	Rasa	-
	Profunda	
	Ultra-profunda	
Características de imageamento	Modo	SCN1, SNB, W1, EXL1
	Órbita	ascendente, descendente
Agrupamento associado	0 (nenhum) ... 44	-
Condições meteo-oceanográficas	Velocidade do vento (Mínima)	baixa, intermediária, alta, sem dados
	Velocidade do vento (Máxima)	
	Altura da onda (Mínima)	baixa, intermediária, alta
	Altura da onda (Máxima)	
	TSM (Mínimo)	baixa, intermediária, alta, sem dados
	TSM (Máximo)	
	TTN favorável à chuva	sim, não
	Alta concentração de clorofila	sim, não, sem dados

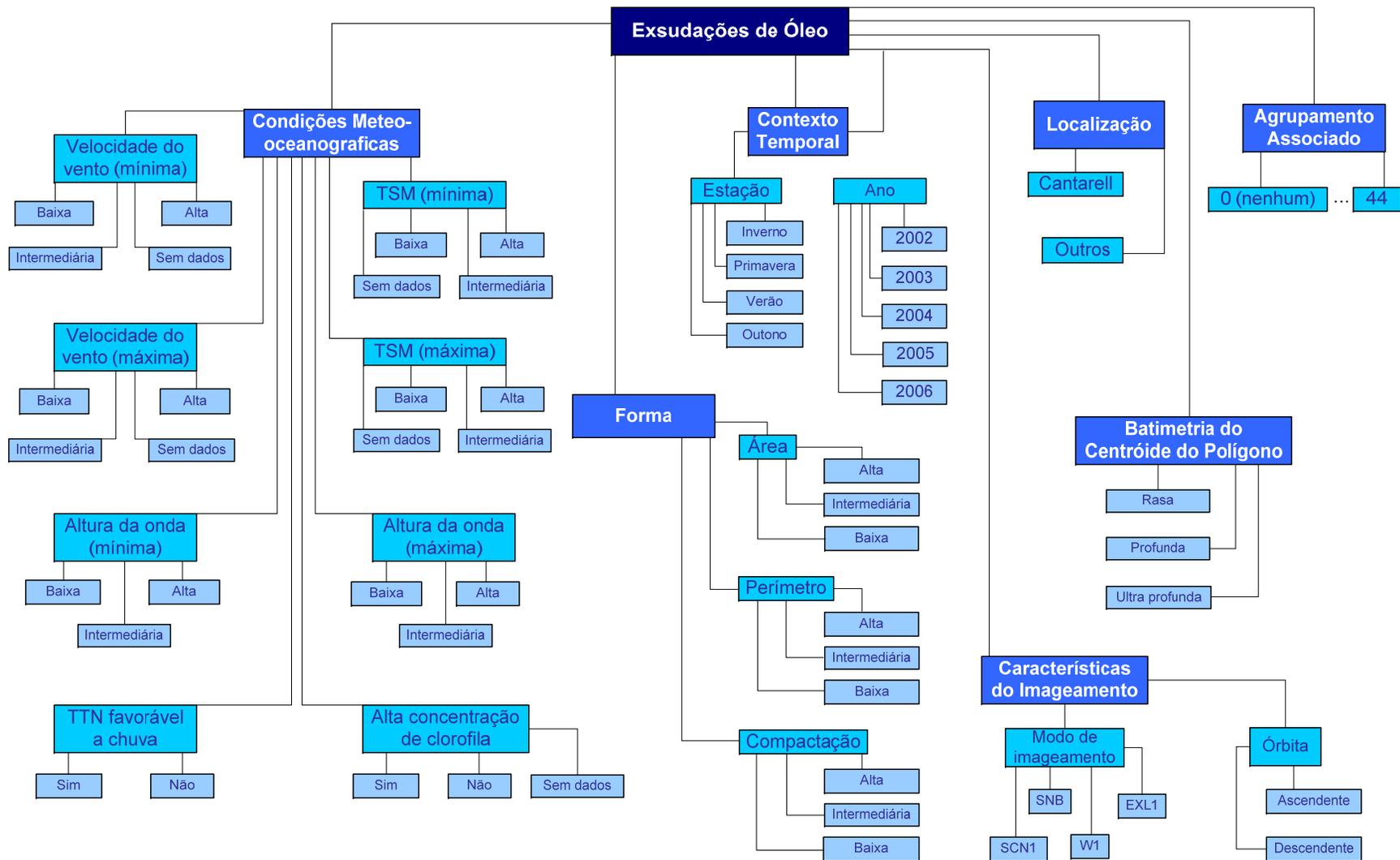


Figura 5.2 – Representação geral dos níveis taxonômicos das exsudações de óleo na Baía de Campeche, Golfo do México.

Na análise exploratória das variáveis representadas na Figura 5.2 foi calculada uma série de medidas de localização (Mínimo, Média, Máximo, 1º Quartil, Mediana, 3º Quartil,) e de dispersão (Desvio Padrão). O objetivo destas medidas é explicitar as características dos dados. Estas variáveis e medidas estão explicitadas e comentadas na seqüência.

1) Localização das exsudações

Este nível diferencia os polígonos relacionados à exsudação de Cantarell daqueles que ocorrem em outros pontos da Baía de Campeche, Golfo do México.



Figura 5.3 – Representação dos sub-níveis do nível taxonômico superior Localização.

Tabela 5.2 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Localização, expressa em termos de distribuição no tempo.

Localização das exsudações			
Ano	Outras	Cantarell	Total
2002	260	40	300
2003	188	49	237
2004	83	51	134
2005	423	53	476
2006	340	53	393
Total	1294	246	1540

Foram identificados 1.540 (mil quinhentos e quarenta) polígonos classificados nas imagens RADARSAT-1 como exsudação, no período compreendido entre 2002 e 2006.

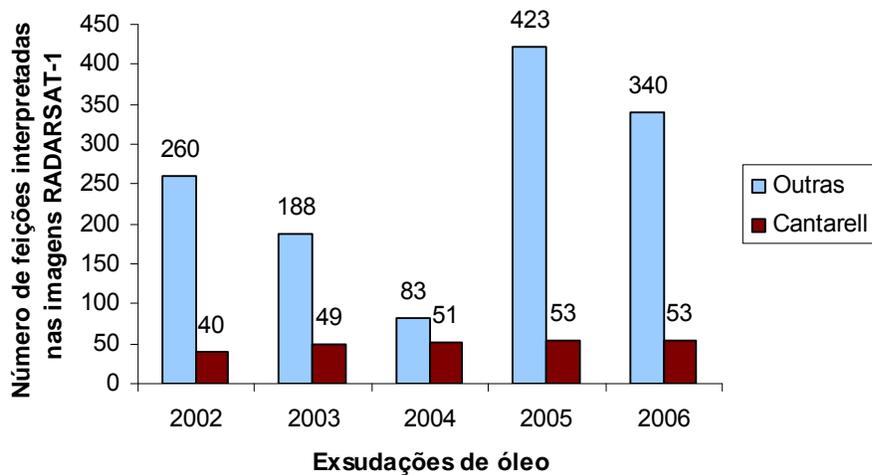


Figura 5.4 – Gráfico de distribuição da quantidade de polígonos interpretados como exsudação nas imagens RADARSAT-1.

2) Contexto Temporal

Este nível determina o ano de aquisição da imagem RADARSAT-1 cuja interpretação gerou os polígonos de exsudação de óleo, como também caracteriza a estação do ano em que o dado de sensoriamento remoto foi obtido (Figura 5.5 e Tabela 5.3).

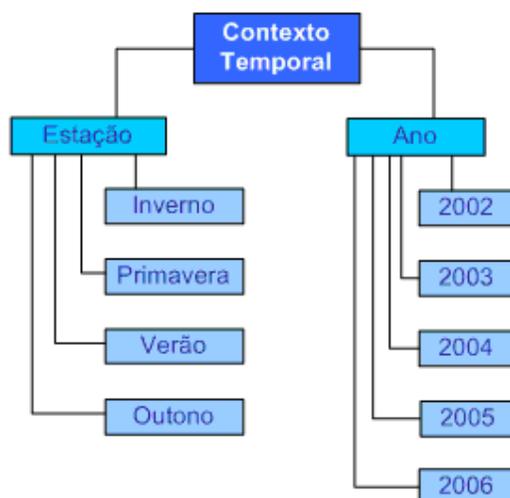


Figura 5.5 – Representação dos sub-níveis do nível taxonômico superior Contexto Temporal.

Tabela 5.3 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Contexto Temporal.

Estações do ano				
Ano	Inverno (21/dez a 20/mar)	Primavera (21/mar a 20/jun)	Verão (21/jun a 20/set)	Outono (21/set 20/mar)
2002	48	63	113	76
2003	69	39	67	62
2004	21	17	76	20
2005	115	190	101	70
2006	95	122	131	45
Total	348	431	488	273

3) Forma

Aspectos relativos à área, ao perímetro e à compactação do polígono interpretado como exsudação de óleo (Figura 5.6) pode ser úteis na determinação da influência em sua forma de fatores tais como as condições meteo-oceanográficas. O atributo Compactação expressa o quanto uma feição apresenta forma que se aproxima de um círculo (Bentz, 2006), sendo calculado por

$$C = \frac{4\pi \times \text{Área}}{\text{Perímetro}} \quad \text{Equação 5.1}$$

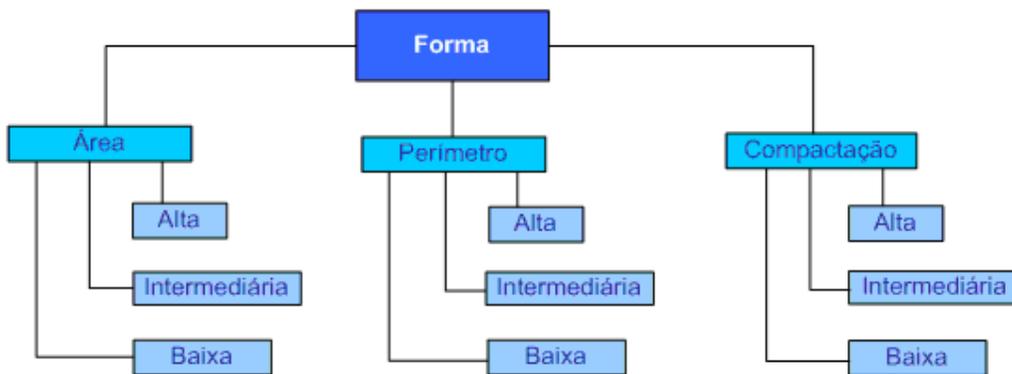


Figura 5.6 – Representação dos sub-níveis do nível taxonômico superior Forma.

Tabela 5.4 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Forma.

Área (km ²)				Perímetro (km)			
Ano	Baixa (< 0,21)	Intermediária (≥0,21 e <198,06)	Alta (≥198,06)	Ano	Baixo (< 6,3)	Intermediário (≥6,3 e <1586,82)	Alto (≥1586,82)
2002	18	281	1	2002	54	246	0
2003	18	219	0	2003	37	200	0
2004	1	132	1	2004	4	129	1
2005	41	433	2	2005	69	405	2
2006	83	310	0	2006	78	315	0
Total	161	1375	4	Total	242	1295	3

Compactação			
Ano	Baixa (< 0,0009)	Intermediária (≥0,0009 e <0,44)	Alta (≥0,44)
2002	0	293	7
2003	0	233	4
2004	1	132	1
2005	2	470	4
2006	2	388	3
Total	5	1516	19

Os resultados estão apresentados nas Tabelas 5.4 e 5.5. Verifica-se que a área das exsudações varia significativamente, com valores que vão de 0,01 km² a 495,80 km², sendo a média 6,15 km².

Tabela 5.5 – Estatística dos sub-níveis do nível taxonômico superior Forma.

Medidas	Área (km ²)	Perímetro (km)	Compactação
Moda	0,35	6,50	0,06
Desvio padrão	22,64	152,68	0,08
Variância	512,53	23311,74	0,01
Média	6,15	55,55	0,07
Mínimo	0,01	0,40	0,00
Máximo	495,80	2826,76	0,79
1° Quartil	0,45	9,05	0,02
Mediana	1,16	18,50	0,04
2° Quartil	3,54	40,78	0,08

4) Batimetria do centróide do polígono

Na base de dados investigada, a batimetria consiste na medida da profundidade do mar na região correspondente ao centróide do polígono interpretado como exsudação de óleo (ou seja, o ponto que constitui o seu centro topológico de massa). Esse parametro pode determinar a ocorrência de ambientes mais propícios a certos processos meteorológicos ou oceanográficos (Figura 5.7). Verifica-se que a maioria dos polígonos encontra-se em intervalos batimétricos correspondentes a águas profundas (Tabela 5.6). Vale registrar que a exsudação de Cantarell se situa a 40 metros de profundidade (águas rasas).

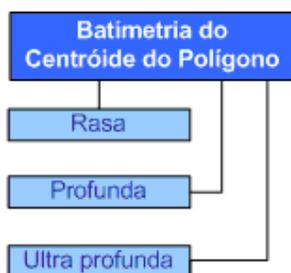


Figura 5.7 – Representação dos sub-níveis do nível taxonômico superior Batimetria do Centróide do Polígono.

Tabela 5.6 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Batimetria do Centróide do Polígono.

Batimetria do Centróide do Polígono (m)			
Ano	Rasa (≤ 200)	Profunda (>200 e ≤ 3.000)	Ultra profunda (>3.000)
2002	64	226	10
2003	69	166	2
2004	60	74	0
2005	88	379	9
2006	73	310	10
Total	354	1155	31

5) Características do imageamento

Os diversos modos de imageamento do satélite RADARSAT-1 apresentam capacidade diferenciada de proporcionar um contraste adequado nas imagens para as feições representativas da presença de óleo na superfície oceânica. Dentre aquelas

utilizadas no monitoramento sistemático do Golfo do México (Figura 5.8), o mais recomendado para esta aplicação é o Wide 1 (W1), em virtude de sua boa relação entre a área nominal de cobertura (165 x 165 Km), ângulos de incidência (20° - 30°) e resolução nominal (30 metros). Outro modo comumente utilizado com tal finalidade é o Scan Sar Narrow 1 (SCN1), com ampla área nominal de cobertura (300 x 300 Km) e ângulos de incidência adequados (20° - 40°), porém com menor resolução espacial (50 metros). As imagens Scan Sar Narrow 2 (SNB) apresentam características muito próximas da SCN1 (Tabela 2.3). Os dados do tipo Extended Low 1 (EXL1) são muito afetados pelo ruído speckle, que caracteriza as imagens de radar. Além disso, possuem valores muito baixos de ângulo de incidência no alcance próximo, o que torna o retorno da superfície do mar excessivamente rugoso.

O modo predominante de aquisição de imagens RADARSAT-1 no programa de monitoramento do Golfo do México é o Scan Sar Narrow 1 (SCN1), seguido pelo Scan Sar Narrow 2 (SNB), conforme a Tabela 6.7. Além disso, as aquisições na órbita ascendente (18:00, hora local) são mais numerosas que na órbita descendente (06:00, hora local).

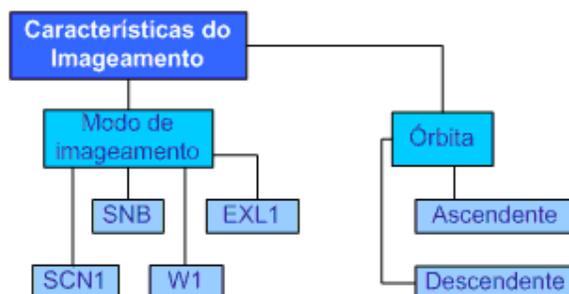


Figura 5.8 – Representação dos sub-níveis do nível taxonômico superior Características do Imageamento.

Tabela 5.7 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Características do Imageamento.

Modo de imageamento				
Ano	SCN1	SNB	W1	EXL1
2002	299	0	0	1
2003	203	33	0	1
2004	56	71	7	0
2005	304	168	3	1
2006	279	105	8	1
Total	1141	377	18	4

Órbita		
Ano	Ascendente	Descendente
2002	157	143
2003	153	84
2004	112	22
2005	301	175
2006	282	111
Total	1005	535

6) Agrupamento associado

Podem ser inferidas tendências regionais de distribuição no espaço e no tempo de sistemas petrolíferos ativos no Golfo de México, por meio de um estudo sistemático das feições representativas de exsudação de óleo nas imagens RADARSAT-1. Um agrupamento de tais feições é interpretado como um grupo de exsudações que compartilham o mesmo ponto de origem na superfície do mar. O ponto de origem foi qualitativamente definido na interseção dos polígonos que coexistem no espaço ou na interseção do prolongamento para frente de polígonos separados, porém convergentes. Como resultado deste procedimento, foram identificados 44 agrupamentos na área de estudo (Figura 5.9). A identificação de agrupamentos de exsudações pode ser considerada como um indicador forte de geração e migração de óleo naquela porção do Golfo do México.

Na Tabela 5.8, são listados os 44 agrupamentos de polígonos interpretados como exsudações de óleo. O valor zero indica a inexistência da tendência à formação de agrupamentos (564 polígonos). Os restantes 976 polígonos estão distribuídos nos 44 agrupamentos especificados.

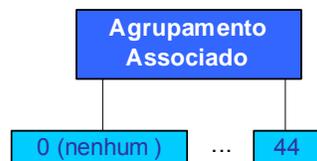


Figura 5.9 – Representação dos sub-níveis do nível taxonômico superior Agrupamento Associado.

Tabela 5.8 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Agrupamento Associado.

Nº	2002	2003	2004	2005	2006	Total
0	118	79	33	168	166	564
1	8	5	2	6	3	24
2	11	12	3	11	0	37
3	15	9	4	8	12	48
4	5	3	1	8	6	23
5	7	4	3	13	7	34
6	6	6	5	16	9	42
7	6	3	2	8	6	25
8	6	4	1	14	9	34
9	5	4	1	11	5	26
10	6	5	1	12	11	35
11	6	4	3	11	9	33
12	6	6	1	13	9	35
13	6	5	2	7	5	25
14	3	1	2	3	1	10
15	3	3	2	11	7	26
16	3	1	1	0	1	6
17	3	2	2	7	3	17
18	2	4	2	5	5	18
19	2	2	2	4	3	13
20	1	1	1	4	1	8
21	5	3	1	6	6	21
22	1	2	0	3	5	11
23	1	1	2	8	3	15
24	3	2	0	8	1	14
25	2	1	0	4	1	8
26	2	2	0	3	1	8
27	0	2	0	6	2	10
28	2	2	0	1	2	7
29	0	4	0	4	2	10
30	2	0	2	2	3	9
31	4	1	0	6	3	14
32	2	0	1	2	3	8
33	1	0	1	4	3	9
34	3	0	0	3	5	11
35	0	3	1	2	2	8
36	0	0	0	3	5	8
37	0	1	0	3	0	4
38	0	0	0	2	2	4
39	1	1	0	3	3	8
40	3	0	0	3	2	8
41	2	0	0	3	2	7
42	0	0	1	1	2	4
43	0	0	0	3	4	7
44*	38	49	51	53	53	244
Total	300	237	134	476	393	1540

* exudação de Cantarell.

7) Condições meteo-oceanográficas

As condições meteo-oceanográficas influenciam significativamente o sinal do radar retroespalhado pela superfície do oceano. Inúmeros fenômenos naturais, da mesma forma que o filme de óleo, atenuam as ondas capilares superficiais, gerando falsos alvos na imagem SAR. Para diminuir tais dúvidas, dados meteo-oceanográficos complementares foram adquiridos por sensores orbitais o mais próximo possível da aquisição das imagens RADARSAT-1. A interpretação de tais produtos gerou as informações incluídas na Figura 5.10. Os atributos discretizados são apresentados na Tabela 5.9, enquanto que as estatísticas básicas dessas variáveis são exibidas na Tabela 5.10.

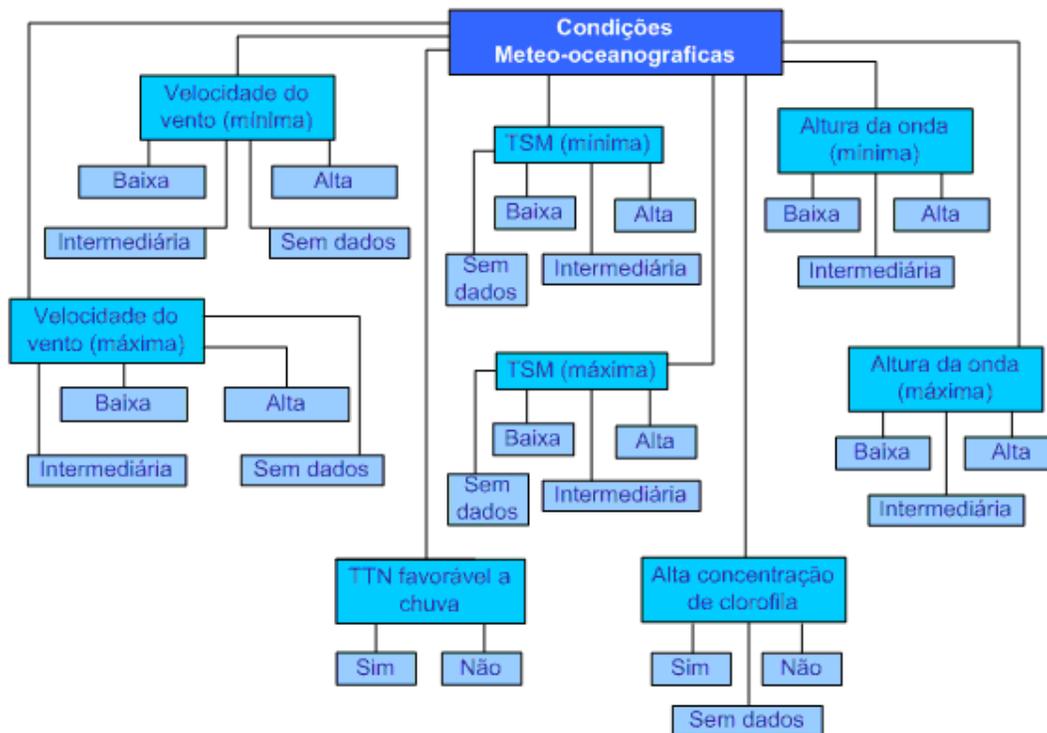


Figura 5.10 – Representação dos sub-níveis do nível taxonômico superior Condições Meteo-oceanográficas.

Tabela 5.9 – Quantidade de polígonos dos sub-níveis do nível taxonômico superior Condições Meteo-oceanograficas. Os níveis de discretização foram estabelecidos pelo especialista do domínio.

Velocidade do Vento (Mínima – m/s)				
Ano	Baixa (< 3,0)	Intermediária (≥3,0 e ≤8,0)	Alta (>8,0)	Sem dado
2002	137	101	0	62
2003	60	140	2	35
2004	79	55	0	0
2005	134	341	1	0
2006	159	234	0	0
Total	569	871	3	97

Velocidade do Vento (Máxima – m/s)				
Ano	Baixa (< 3,0)	Intermediária (≥3,0 e ≤8,0)	Alta (>8,0)	Sem dado
2002	0	170	68	62
2003	0	127	75	35
2004	0	104	30	0
2005	0	282	194	0
2006	0	310	83	0
Total	0	993	450	97

Altura da onda (Mínima – m)			
Ano	Baixa (≤ 1,5)	Intermediária (>1,5 e ≤2,0)	Alta (>2,0)
2002	271	25	4
2003	228	9	0
2004	133	1	0
2005	476	0	0
2006	383	10	0
Total	1491	45	4

Altura da onda (Máxima – m)			
Ano	Baixa (≤ 1,5)	Intermediária (>1,5 e ≤2,0)	Alta (>2,0)
2002	260	10	30
2003	138	75	24
2004	112	14	8
2005	414	28	34
2006	374	6	13
Total	1298	133	109

Temperatura da Superfície do Mar (Mínima - °C)				
Ano	Baixa (≤ 20)	Intermediária (>20 e ≤30)	Alta (>30)	Sem dado
2002	0	295	0	5
2003	0	219	0	18
2004	0	127	0	7
2005	2	473	0	1
2006	0	390	0	3
Total	2	1504	0	34

Temperatura da Superfície do Mar (Máxima - °C)				
Ano	Baixa (≤ 20)	Intermediária (>20 e ≤30)	Alta (>30)	Sem dado
2002	0	293	2	5
2003	0	210	9	18
2004	0	123	4	7
2005	0	428	47	1
2006	0	390	0	3
Total	0	1444	62	34

TTN favorável à chuva		
Ano	Sim (TTN < -40°C)	Não (TTN > -40°C)
2002	90	210
2003	54	183
2004	31	103
2005	142	334
2006	118	275
Total	435	1105

Alta concentração de clorofila-a			
Ano	Sim	Não	Sem dado
2002	17	226	57
2003	7	230	0
2004	38	96	0
2005	19	457	0
2006	3	385	5
Total	84	1394	62

Tabela 5.10 – Estatística básica dos sub-níveis numéricos do nível taxonômico superior Condições Meteo-oceanograficas.

Medidas	Velocidade do vento (Mínima – m/s)	Velocidade do vento (Máxima – m/s)	Altura da onda (Mínima - m)	Altura da onda (Máxima - m)	TSM (Mínima -°C)	TSM (Máxima - °C)
Moda	2,00	9,00	0,50	1,30	27,00	28,00
Desvio padrão	1,42	2,27	0,35	0,47	2,21	2,14
Variância	2,02	5,17	0,12	0,22	4,90	4,59
Mínimo	0,50	3,00	0,00	0,50	19,00	23,00
Média	3,26	7,46	0,69	1,26	25,47	27,65
Máximo	13,00	16,99	2,50	3,00	29,50	33,50
1º Quartil	2,00	5,50	0,50	1,00	24,00	26,00
Mediana	3,00	7,00	0,50	1,20	26,00	28,00
2º Quartil	4,00	9,00	0,90	1,50	27,00	29,50

5.2 – Extração de padrões (CBA)

Em atenção ao requisito de simplicidade imposto pelo especialista do domínio, foram consideradas na presente dissertação apenas as regras de associação geradas no CBA que envolvem dois itens, ou seja, um item no antecedente e um item no conseqüente, dos conjuntos freqüentes gerados de comprimento dois (*2-itemsets*). Pretende-se, assim, descrever padrões de relacionamento entre pares de itens da base de dados sobre exsudações de óleo no Golfo do México.

Tal abordagem pode determinar condições meteo-oceanográficas, de localização, contexto temporal, forma, batimetria, características de imageamento e tendência a agrupamento que costumam ocorrer em conjunto. Espera-se que uma transação que contem o item antecedente provavelmente apresente também aquele do conseqüente, considerando que a regra de associação atenda aos requisitos mínimos de suporte e confiança pré-estabelecidos. Estão apresentadas na Tabela 5.11 as freqüências de ocorrência dos subníveis taxonômicos que serviram de atributo de entrada do aplicativo CBA.

Tabela 5.11 – Freqüência de ocorrência dos subníveis taxonômicos que serviram de atributos de entrada no aplicativo CBA.

Variáveis	Atributos	Freqüência absoluta	Freqüência relativa (%)
Localização	Cantarell	246	15,97%
	Outros	1294	84,03%

Variáveis	Atributos		Frequência absoluta	Frequência relativa (%)
Contexto temporal	Ano	2002	300	19,48%
		2003	237	15,39%
		2004	134	8,70%
		2005	476	30,91%
		2006	393	25,52%
	Estação	inverno	348	22,60%
		primavera	431	27,99%
		verão	488	31,69%
outono		273	17,72%	
Forma	Área	baixa	161	10,45%
		intermediária	1375	89,29%
		alta	4	0,26%
	Perímetro	baixo	242	15,71%
		intermediário	1295	84,09%
		alto	3	0,20%
	Compactação	baixa	5	0,33%
		intermediária	1516	98,44%
alta		19	1,23%	
Batimetria do centroide do polígono	Rasa		354	22,99%
	Profunda		1155	75,00%
	Ultra-profunda		31	2,01%
Características de imageamento	Modo	SCN1	1141	74,09%
		SNB	377	24,48%
		W1	18	1,17%
		EXL1	4	0,26%
	Órbita	ascendente	1005	65,26%
		descendente	535	34,74%
Agrupamento associado	0 (nenhum)		564	36,62%
	1		24	1,56%
	2		37	2,40%
	3		48	3,12%
	4		23	1,49%
	5		34	2,21%
	6		42	2,73%
	7		25	1,62%
	8		34	2,21%
	9		26	1,69%
	10		35	2,28%
	11		33	2,14%
	12		35	2,27%
	13		25	1,62%
	14		10	0,65%
	15		26	1,69%

Variáveis	Atributos	Frequência absoluta	Frequência relativa (%)	
	16	6	0,39%	
	17	17	1,11%	
	18	18	1,17%	
	19	13	0,84%	
	20	8	0,52%	
	21	21	1,36%	
	22	11	0,71%	
	23	15	0,97%	
	24	14	0,91%	
	25	8	0,52%	
	26	8	0,52%	
	27	10	0,65%	
	28	7	0,46%	
	29	10	0,65%	
	30	9	0,58%	
	31	14	0,91%	
	32	8	0,52%	
	33	9	0,58%	
	34	11	0,71%	
	35	8	0,52%	
	36	8	0,52%	
	37	4	0,26%	
	38	4	0,26%	
	39	8	0,52%	
	40	8	0,52%	
	41	7	0,46%	
	42	4	0,26%	
	43	7	0,46%	
	44	244	15,84%	
Condições meteo-oceanográficas	Velocidade do vento (Mínima)	baixa	569	36,95%
		intermediária	871	56,56%
		alta	3	0,19%
		sem dados	97	6,30%
	Velocidade do vento (Máxima)	baixa	-	-
		intermediária	993	64,48%
		alta	450	29,22%
		sem dados	97	6,30%
	Altura da onda (Mínima)	baixa	1491	96,82%
		intermediária	45	2,92%
		alta	4	0,26%
	Altura da onda (Máxima)	baixa	1298	84,28%
intermediária		133	8,64%	
alta		109	7,08%	

Variáveis	Atributos		Frequência absoluta	Frequência relativa (%)
	TSM (Mínimo)	baixa	2	0,13%
		intermediária	1504	97,66%
		alta	-	-
		sem dados	34	2,21%
	TSM (Máximo)	baixa	-	-
		intermediária	1444	93,77%
		alta	62	4,02%
		sem dados	34	2,21%
	TTN favorável à chuva	sim	435	28,25%
		não	1105	71,75%
	Alta concentração de clorofila	sim	84	5,45%
		não	1394	90,52%
sem dados		62	4,03%	

Assim, após a discretização dos atributos na etapa de pré-processamento, os dados foram adaptados para utilização do aplicativo CBA de extração de padrões. No arquivo exsudação.DATA, empregado no CBA, os dados referentes às exsudações estão limpos e compostos somente pelos valores discretos. O arquivo exsudação.NAMES contém o nome dos atributos e de seus valores (Tabela 5.12). Porém para geração das regras também se faz necessária a definição das opções de configuração. Para o presente estudo foram especificadas as seguintes opções de configuração:

- Suporte_Min = 0,1
- Confiança_Min = 0
- Limite de regras = 800.000
- Número de condições = 2

Utilizou-se o menor valor de suporte e confiança, pois o objetivo foi de identificar todas as associações existentes. A configuração supra citada resultou na geração de 100.925 regras de associação para a base em estudo.

Tabela 5.12 – Conteúdo do arquivo de entrada do CBA (exsudação.NAMES).

Modo_imageamento:	SCN1, SNB, W1, EXL1.
Órbita:	ascendente, descendente.
Ano:	2002, 2003, 2004, 2005, 2006.
Estação:	inverno, primavera, verão, outono.
TTN_favorável_chuva:	sim, não.

Clorofila_alta:	sim, não, sem-dado.
Cluster:	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44.
Área:	baixa, intermediária, alta.
Perímetro:	baixo, intermediário, alto.
Compactação:	baixa, intermediária, alta.
Velocidade_vento_Mín:	baixa, intermediária, alta, sem-dado.
Velocidade_vento_Máx:	baixa, intermediária, alta, sem-dado.
Altura_onda_Mín:	baixa, intermediária, alta.
Altura_onda_Máx:	baixa, intermediária, alta.
TSM_Mín:	baixa, intermediária, alta, sem-dado.
TSM_Máx:	baixa, intermediária, alta, sem-dado.
Batimetria_centroide:	rasa, profunda, ultra-profunda.
Localização:	Outras, Cantarell.

Para facilitar a identificação e a localização das regras de associação envolvendo dois itens, utilizaram-se inicialmente pares de itens do arquivo de saída do CBA com extensão *. ITM (Tabela 5.13). Os pares de itens determinam a frequência de ocorrência de dois itens associados. Tal procedimento foi utilizado, pois o CBA gera regras com no mínimo até dois itens no antecedente e dois no conseqüente, não disponibilizando a visualização apenas das regras com um antecedente e um conseqüente.

5.3 – Pós-processamento e análise das relações interessantes obtidas como resultado

O modelo típico para mineração de regras de associação em base dados consiste em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo (Sup_Min) e a uma confiança mínima (Conf_Min), especificados pelo usuário (Gonçalves, 2005). Na presente dissertação, procurou-se definir o valor do Sup_Min pela observação da taxa de variação da quantidade de pares de itens com o suporte, ou seja, a derivada $d\text{Quantidade de pares} / d\text{Suporte}$. Tais gráficos estão apresentados nas Figuras 5.11 e 5.12.

A análise dos valores da derivada na Tabela 5.13 indica que há uma tendência de estabilização em um patamar correspondente ao intervalo -1 a -2, a partir do Suporte de 76%. Tal comportamento, visto em detalhe na Figura 5.13, permitiu

estabelecer que Sup_Min = 76% (linha em letras vermelhas). Esse procedimento foi adotado por recomendação do especialista do domínio, com a finalidade de tornar objetivo o critério de definição do Sup_Min.

Tabela 5.13 – Pares de itens do arquivo de saída do CBA (Extensão *. ITM) com suporte maior ou igual a 0,1%. A linha em vermelho (76%) é ilustrada na Figura 5.13.

Suporte (%)	Frequência de pares	Frequência de pares (Acumulado)	dQuantidade de pares / dSuporte)
0,1	1442	2388	-1442
1	226	946	-226
2	126	720	-126
3	55	594	-55
4	38	539	-38
5	35	501	-35
6	24	466	-24
7	29	442	-29
8	16	413	-16
9	18	397	-18
10	21	379	-21
11	8	358	-8
12	16	350	-16
13	11	334	-11
14	14	323	-14
15	28	309	-28
16	10	281	-10
17	8	271	-8
18	12	263	-12
19	9	251	-9
20	14	242	-14
21	10	228	-10
22	12	218	-12
23	5	206	-5
24	14	201	-14
25	13	187	-13
26	10	174	-10
27	13	164	-13
28	9	151	-9
29	3	142	-3
30	7	139	-7
31	9	132	-9
32	4	123	-4
33	2	119	-2
34	6	117	-6
35	4	111	-4
36	3	107	-3
37	1	104	-1
38	1	103	-0,5
40	1	102	-0,5

Suporte (%)	Frequência de pares	Frequência de pares (Acumulado)	dQuantidade de pares / dSuporte)
42	3	101	-3
43	1	98	-0,3
46	2	97	-2
47	2	95	-2
48	3	93	-3
49	1	90	-1
50	1	89	-0,5
52	1	88	-1
53	3	87	-3
54	3	84	-3
55	4	81	-4
56	3	77	-3
57	3	74	-3
58	1	71	-1
59	2	70	-2
60	1	68	-1
61	5	67	-2,5
63	5	62	-5
64	4	57	-4
65	1	53	-1
66	1	52	-1
67	1	51	-1
68	4	50	-4
69	1	46	-1
70	3	45	-3
71	4	42	-4
72	1	38	-1
73	3	37	-3
74	3	34	-1,5
76	3	31	-1,5
78	1	28	-1
79	2	27	-2
80	1	25	-1
81	3	24	-3
82	5	21	-5
83	2	16	-2
84	2	14	-1
86	2	12	-2
87	1	10	-1
88	2	9	-2
89	1	7	-1
90	1	6	-0,5
92	1	5	-1
93	1	4	-1
94	1	3	-1
95	1	2	-1
96	1	1	

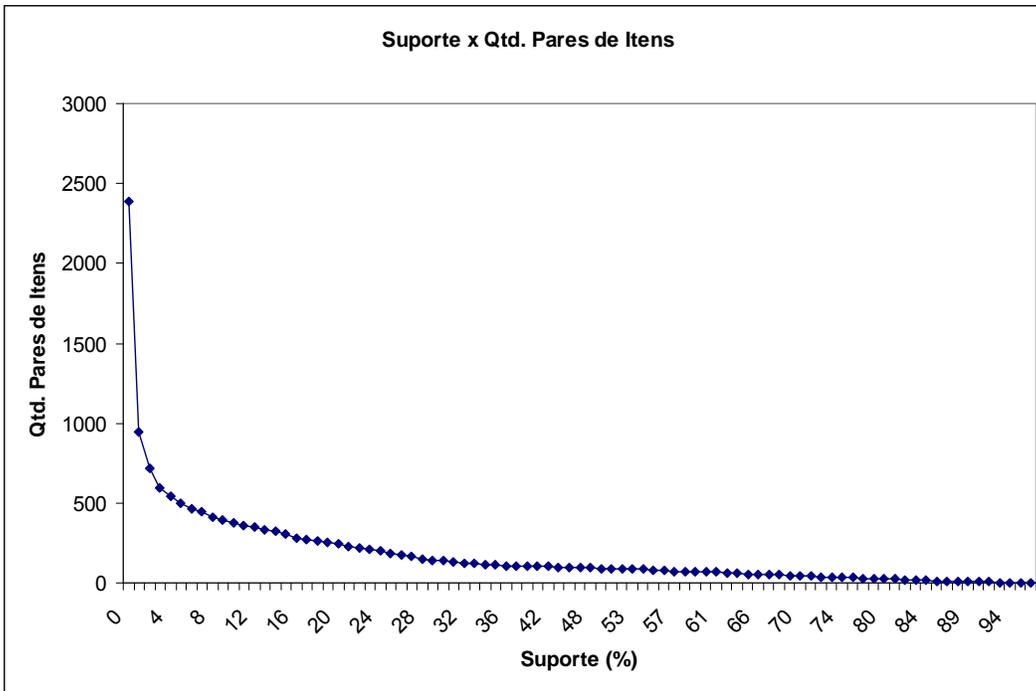


Figura 5.11 – Gráfico do suporte em relação à quantidade de pares de itens.

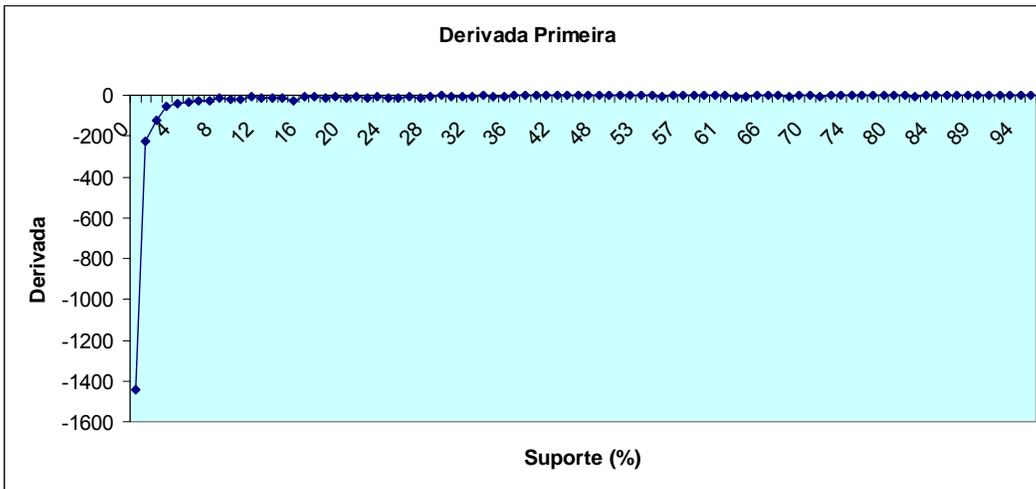


Figura 5.12 – Gráfico da derivada primeira dos pares de itens em relação ao suporte (d Quantidade de pares / d Suporte).

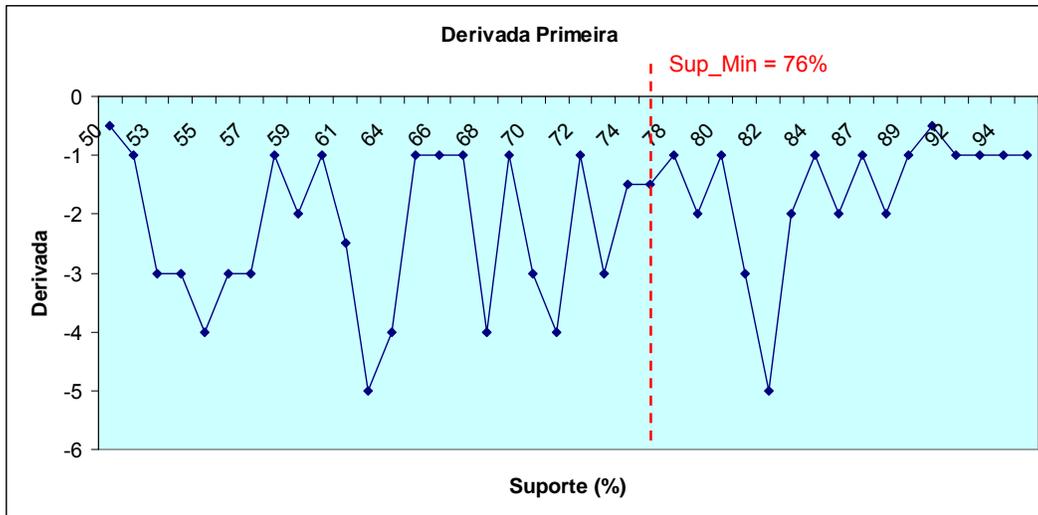


Figura 5.13 – Derivada primeira dos pares de itens em relação ao suporte, para valores de suporte maiores que 50%.

As relações resultantes estão assim apresentadas: (1) regras de associação obtidas exclusivamente entre os subníveis de um mesmo nível taxonômico (intraníveis; Tabela 5.14); (2) regras de associação obtidas entre níveis taxonômicos distintos (interníveis; Tabela 5.15). Para as regras intraníveis, o suporte varia de 94,98% a 70,04%, enquanto a confiança abrange o intervalo de 100,00% a 84,14%. Para as regras interníveis, o suporte varia de 96,10% a 76,10%, enquanto que a confiança abrange o intervalo de 98,41% a 83,70%.

Verifica-se na Tabela 5.14, referente às regras de associação intraníveis, que a regra 1 (TSM_Mín = intermediária → Altura_onda_Mín = baixa) apresenta os maiores valores, com suporte de 94,48% e confiança de 96,74%. Ou seja, em 94,48% das exsudações ocorrem TSM_Mín = intermediária e Altura_onda_Mín = baixa em simultâneo. Já confiança de 96,74% representa a probabilidade de ocorrer Altura_onda_Mín = baixa dado que TSM_Mín = intermediária ocorreu.

Na Tabela 5.15, que lista as regras de associação interníveis, a regra 1 (TSM_Mín = intermediária → Compactação = intermediária) apresenta os maiores valores, com suporte de 96,10% e uma confiança de 98,40%. Ou seja, em 96,10% das exsudações ocorrem TSM_Mín = intermediária e Compactação = intermediária em simultâneo. Já confiança de 98,40% representa a probabilidade de ocorrer Compactação = intermediária dado que TSM_Mín = intermediária ocorreu.

Se aplicada aos demais resultados das Tabelas 5.14 e 5.15, semelhante análise indica a presença de regras fortes, que exibem valores altos de suporte e confiança.

O grau de dependência das regras de associação mostradas nas Tabelas 5.14 e 5.15 é aquilatado pelo valor da medida de interesse denominada *lift*. Para as regras intraníveis, o valor do *lift* está majoritariamente situado entre 1,00 e 1,03 (existe apenas um caso de $lift = 1,10$); para as interníveis, o intervalo abrange 1,00 a 1,01. Tais valores indicam que os itens que integram as regras de associação são independentes. Considere a regra 11 da Tabela 5.14 (Compactação = intermediária → Área Intermediária), a confiança desta regra é de 89,97%, e representa a probabilidade do conseqüente ocorrer dado que o antecedente ocorreu. O suporte isolado do conseqüente é igual a 89,29%, portanto, é possível notar que o antecedente pouco aumentou a probabilidade do conseqüente. Assim sendo, a ocorrência de Compactação = intermediária independe da ocorrência da Área = Intermediária.

As medidas de interesse subjetivas exibidas nas Tabelas 5.14 e 5.15 foram propostas pelo especialista para avaliação das regras de associação, como segue: potencialmente causal (PC), indicando conhecimento que pode ser novidade ou que reforça um conhecimento prévio; óbvia (OB), indicando mera circunstância sem valor para o domínio do especialista. Verifica-se nas Tabelas 5.14 e 5.15, que quase todas as regras de associação são potencialmente causais. A única exceção se refere à regra intranível Altura_onda_Máx = baixa → Altura_onda_Min = baixa, com suporte de 84,29% e confiança de 100%, que foi considerada óbvia. Evidentemente, se a altura máxima da onda é baixa ($\leq 1,5m$), então a altura mínima da onda só pode ser baixa, o que explica a confiança = 100%.

A robustez dos resultados acima descritos pode ser avaliada por comparação, no nível taxonômico Localização, das regras de associação correspondentes aos subníveis Cantarell e Outros (Tabela 5.16). Para Cantarell, embora as regras apresentem para o suporte valores entre 15,65% e 13,64%, a confiabilidade se mantém alta, variando de 85,37% a 97,97%. Para as regras de associação relativas a Cantarell, o *lift* varia de 0,94 a 1,01, indicando independência entre os antecedentes e conseqüentes, a exemplo do que é mostrado nas Tabelas 5.14 e 5.15. Assim, pode-se inferir que os mesmos fatores ambientais e de forma controlam as exsudações de óleo indistintamente ao Golfo do México (isto é, em Cantarell ou fora dele).

Na presente dissertação, foram também identificadas algumas regras de associação raras, com suporte menor que 10,0%, as quais apresentaram, em contrapartida, valores muito altos de confiança e/ou *lift* (Tabela 5.17). A regra 1, Compactação = alta → Perímetro = baixo, tem suporte de 1,23%, confiabilidade de 100,00% e *lift* igual a 6,36. Neste caso, os itens possuem forte dependência positiva e as exsudações resultantes apresentam forma que se aproxima de um círculo (Figura 5.14). A regra Perímetro = baixo → Área = baixa, com 9,16% de suporte, 58,26% de confiança e 5,57 de *lift*, tem interpretação intuitiva.

Por sua vez, a regra rara Altura_onda_Máx = intermediária → Ano = 2003, com 4,87% de suporte, 56,39% de confiança e 3,66 de *lift*, pode ser tentativamente justificada pela ocorrência de um evento fraco de El Niño no período 2002-2003, que pode ter afetado o clima no Golfo do México (NASA, 2008). Uma razão climática também pode ser advogada para a regra Altura_onda_Mín = intermediária → Estação = outono, com suporte de 1,75%, confiança de 60,00% e 3,38 de *lift*.

Tabela 5.14 – Lista de regras de associação intraníveis.

Nº da regra	Antecedente	Conseqüente	Grupo	Interesse Subjetivo	SUP (%)	CONF (%)	LIFT
1	TSM_Mín = intermediária	Altura_onda_Mín = baixa	CMO	PC	94,48	96,74	1,00
2	TSM_Mín = intermediária	Altura_onda_Máx = baixa	CMO	PC	82,66	84,64	1,00
3	TSM_Mín = intermediária	Clorofila_alta = não	CMO	PC	88,38	90,49	1,00
4	TSM_Máx = intermediária	TSM_Mín = intermediária	CMO	PC	93,64	99,86	1,02
5	TSM_Máx = intermediária	Altura_onda_Mín = baixa	CMO	PC	90,58	96,61	1,00
6	TSM_Máx = intermediária	Clorofila_alta = não	CMO	PC	84,55	90,17	1,00
7	TSM_Máx = intermediária	Altura_onda_Máx = baixa	CMO	PC	78,90	84,14	1,00
8	Altura_onda_Mín = baixa	Clorofila_alta = não	CMO	PC	87,34	90,21	1,00
9	Altura_onda_Máx = baixa	Altura_onda_Mín = baixa	CMO	OB	84,29	100,00	1,03
10	Altura_onda_Máx = baixa	Clorofila_alta = não	CMO	PC	76,04	90,22	1,00
11	Compactação = intermediária	Área = intermediária	F	PC	88,57	89,97	1,01
12	Compactação = intermediária	Perímetro = intermediário	F	PC	83,90	85,22	1,01
13	Perímetro = intermediário	Área = intermediária	F	PC	82,73	98,38	1,10

CMO: Condições Meteo-Oceanográficas

PC: Potencialmente Causal

F: Forma

OB: Óbvia

Tabela 5.15 – Lista de regras de associação interníveis.

Nº da regra	Antecedente	Conseqüente	Grupo	Interesse Subjetivo	SUP (%)	CONF (%)	LIFT
1	TSM_Mín = intermediária	Compactação = intermediária	CMO; F	PC	96,10	98,40	1,00
2	TSM_Mín = intermediária	Perímetro = intermediário	CMO; F	PC	81,82	83,78	1,00
3	TSM_Mín = intermediária	Área = intermediária	CMO; F	PC	86,95	89,03	1,00
4	TSM_Máx = intermediária	Compactação = intermediária	CMO; F	PC	92,27	98,41	1,00
5	TSM_Máx = intermediária	Área = intermediária	CMO; F	PC	83,83	89,40	1,00
6	TSM_Máx = intermediária	Perímetro = intermediário	CMO; F	PC	79,03	84,28	1,00
7	Altura_onda_Mín = baixa	Perímetro = intermediário	CMO; F	PC	81,04	83,70	1,00
8	Altura_onda_Mín = baixa	Área = intermediária	CMO; F	PC	86,17	89,00	1,00
9	Altura_onda_Mín = baixa	Compactação = intermediária	CMO; F	PC	95,26	98,39	1,00
10	Altura_onda_Máx = baixa	Compactação = intermediária	CMO; F	PC	82,92	98,38	1,00
11	Área = intermediária	Clorofila_alta = não	F; CMO	PC	80,78	90,47	1,00
12	Compactação = intermediária	Clorofila_alta = não	F; CMO	PC	89,29	90,70	1,00
13	Perímetro = intermediário	Clorofila_alta = não	F; CMO	PC	76,10	90,50	1,00
14	Localização = Outras	TSM_Mín = intermediária	L; CMO	PC	82,40	98,07	1,00
15	Localização = Outras	TSM_Máx = intermediária	L; CMO	PC	79,29	94,36	1,01
16	Localização = Outras	Altura_onda_Mín = baixa	L; CMO	PC	81,17	96,60	1,00
17	Localização = Outras	Clorofila_alta = não	L; CMO	PC	76,88	91,50	1,01
18	Localização = Outras	Compactação = intermediária	L;F	PC	82,79	98,53	1,00

CMO: Condições Meteo-Oceanográficas

PC: Potencialmente Causal

F: Forma

L: Localização

Tabela 5.16 – Lista com a comparação entre as regras de associação correspondentes aos subníveis Cantarell e Outras do nível taxonômico Localização.

Nº da regra	Antecedente	Conseqüente	Grupo	SUP (%)	CONF (%)	LIFT
1	Localização = Cantarell	Compactação = intermediária	L; F	15,65	97,97	1,00
2	Localização = Outras	Compactação = intermediária	L; F	82,79	98,53	1,00
3	Localização = Cantarell	TSM Mín = intermediária	L; CMO	15,26	95,53	0,98
4	Localização = Outras	TSM Mín = intermediária	L; CMO	82,40	98,07	1,00
5	Localização = Cantarell	TSM Máx = intermediária	L; CMO	14,48	90,65	0,97
6	Localização = Outras	TSM Máx = intermediária	L; CMO	79,29	94,36	1,01
7	Localização = Cantarell	Altura onda Mín = baixa	L; CMO	15,65	97,97	1,01
8	Localização = Outras	Altura onda Mín = baixa	L; CMO	81,17	96,60	1,00
9	Localização = Cantarell	Clorofila alta = não	L; CMO	13,64	85,37	0,94
10	Localização = Outras	Clorofila alta = não	L; CMO	76,88	91,50	1,01

CMO: Condições Meteo-Oceanográficas L: Localização
F: Forma

Tabela 5.17 – Lista de regras de associação raras.

Nº da regra	Antecedente	Conseqüente	Grupo	SUP (%)	CONF (%)	LIFT
1	Compactação = alta	Perímetro = baixo	F; F	1,23	100,00	6,36
2	Perímetro = baixo	Área = baixa	F; F	9,16	58,26	5,57
3	Altura_onda Mín = intermediária	Estação = outono	CMO; CT	1,75	60,00	3,38
4	Altura_onda Máx = intermediária	Ano = 2003	CMO; CT	4,87	56,39	3,66

CMO: Condições Meteo-Oceanográficas CT: Contexto Temporal
F: Forma

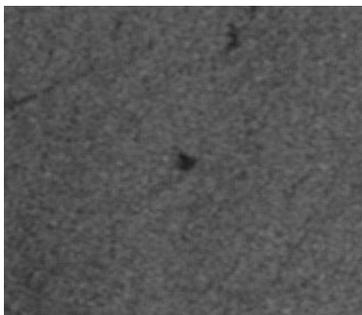


Imagem do satélite RADARSAT-1 (SNB - órbita ascendente), adquirida sobre a porção *offshore* do Golfo do México, Baía de Campeche em 8 de agosto de 2004.

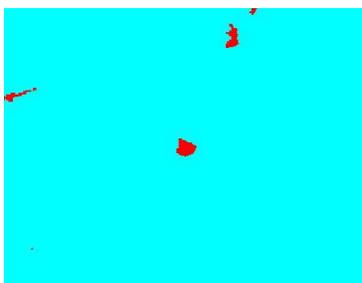
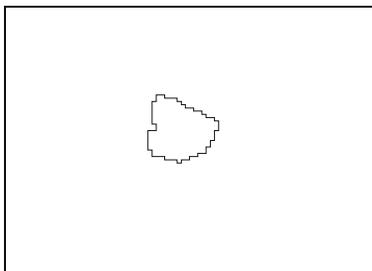


Imagem classificada pelo método USTC, a classe azul representa a superfície rugosa do mar, enquanto que a vermelha representa as superfícies lisas.



Polígono identificado como exsudação de óleo, com área de 0,60 km² (classificada como intermediária), perímetro de 3,90 km (classificado como baixo) e compactação de 0,50 (classificada como alta).

Figura 5.14 – Exemplo de um polígono, de um total de dezenove, que representa a regra rara “Compactação= alta → Perímetro= baixo”.

5.4 – Análise do potencial de uso do novo conhecimento

O conhecimento extraído, depois de ser avaliado e validado na etapa de pós-processamento, é consolidado na fase de utilização do conhecimento, podendo ser incorporado a um sistema inteligente, utilizado diretamente pelo usuário final para apoio a algum processo de tomada de decisão ou, simplesmente, relatado às pessoas interessadas (Sinoara, 2006).

A utilização efetiva do conhecimento descoberto em um processo de apoio a decisão depende de fatores como a possibilidade do usuário compreender o conhecimento e conseguir identificar, por exemplo, o conjunto de regras mais interessantes sob seu ponto de vista (Melanda, 2005). Para dar suporte ao especialista na utilização do conhecimento descoberto é de grande relevância a avaliação das regras quanto a sua qualidade, compreensibilidade, utilidade e grau de interesse. Neste sentido, o emprego das medidas objetivas para a avaliação da força de associação entre os lados da regra, colabora na identificação das regras interessantes.

Contudo, somente o especialista do domínio e/ou usuário final da aplicação estão aptos a mensurar o quão realmente interessante são as regras obtidas pelo processo de KDD, se elas vão de acordo ao seu conhecimento pré-existente e se as mesmas poderão apoiá-lo no processo de tomada de decisão.

Verifica-se que as regras de associação intraníveis referentes às Condições Meteo-oceanográficas, quase todas potencialmente causais, os antecedentes são TSM_Mín = intermediária, TSM_Máx = intermediária, Altura_onda_Mín = baixa e Altura_onda_Máx = baixa. Os conseqüentes dessas regras são Altura_onda_Mín = baixa, Altura_onda_Máx = baixa, Clorofila_alta = não e TSM_Mín = intermediária. Tal resultado sugere a possibilidade de um relacionamento inesperado entre a temperatura da superfície do mar (TSM) e a altura significativa de ondas, com possível influencia nos fenômenos associados à produção de clorofila-a no ambiente marinho. Esse conhecimento apresenta potencial promissor para novas investigações científicas pelo especialista do domínio. As regras de associação intraníveis correspondentes à forma, todas potencialmente causais, indicam que área, perímetro e compactação possuem majoritariamente valores intermediários.

As regras de associação interníveis, todas potencialmente causais, possuem como antecedente três níveis taxonômicos (Condições Meteo-oceanográficas, Forma e Localização; Tabela 5.15). Os antecedentes relacionados às Condições Meteo-oceanográficas são TSM_Mín = intermediária, TSM_Máx = intermediária, Altura_onda_Mín = baixa e Altura_onda_Máx = baixa, as quais apresentam como conseqüentes Compactação = intermediária, Perímetro = intermediário e Área = intermediária. Tal resultado sugere a existência de controles ambientais na forma das exsudações de óleo. Por outro lado, os antecedentes relacionados à Forma (Área = intermediária, Compactação = intermediária e Perímetro = intermediário) estão associados ao conseqüente Clorofila_alta = não. Finalmente, o antecedente Localização = outros possui como conseqüentes TSM_Mín = intermediária, TSM_Máx = intermediária, Altura_onda_Mín = baixa, Clorofila_alta = não e Compactação = intermediária.

CAPÍTULO 6

CONCLUSÕES E RECOMENDAÇÕES

A pesquisa aqui apresentada tem por objetivo a identificação de regras de associação interessantes em uma base de dados de execuções de óleo, obtida a partir da interpretação de imagens RADARSAT-1 do Golfo do México, Baía de Campeche. Tal interpretação foi realizada pelo Laboratório de Sensoriamento Remoto por Radar (LABSAR), situado na COPPE/UFRJ, em parceria com a empresa canadense RADARSAT *Internacional* Inc. (RSI) e com a companhia estatal mexicana PEMEX Exploração e Produção (PEP).

As regras de associação descrevem padrões de relacionamento entre itens de uma base de dados. Uma de suas aplicações mais características refere-se ao exame de compras de consumidores para determinar produtos que costumam ser adquiridos em conjunto. A presente dissertação faz uso dessa abordagem de maneira inovadora, considerando uma exsudação de óleo com uma transação, à qual são associados atributos de natureza diversa (localização, contexto, temporal, forma, batimetria do centróide do polígono, características do imageamento RADARSAT-1, tendência ao agrupamento e condições meteo-oceanográficas).

Por questões de simplicidade, foram consideradas apenas regras de associação envolvendo dois itens, ou seja, um item no antecedente e um no conseqüente, produzindo conjuntos freqüentes de comprimento dois (*2-itemsets*). O aplicativo utilizado foi o CBA (*Classification Based Association*).

O processo aqui utilizado pode ser dividido em cinco etapas: (1) identificação do problema, que envolve a compreensão do domínio da aplicação e a seleção do conjunto de dados; (2) pré-processamento, que consta da integração, limpeza e redução dos dados; (3) extração de padrões, que abrange a geração de regras de associação; (4) pós-processamento, que inclui a simplificação e a avaliação dos novos padrões extraídos, considerando o atendimento a requisitos mínimos de suporte e confiança; o grau de dependência da regra de associação foi avaliado pela medida de

interesse objetiva denominada *lift*; (5) utilização do conhecimento, na qual o usuário final consolida o conhecimento extraído.

Os resultados assim obtidos indicam a presença de regras fortes que exibem valores altos de suporte e confiança. Os valores de *lift* para essas regras mostram quase sempre independência entre o antecedente e o conseqüente. Foi verificada a possibilidade de um relacionamento inesperado entre a temperatura da superfície do mar (TSM) e a altura significativa de onda, com possível influência nos fenômenos associados à produção de clorofila-a no ambiente marinho. Esse conhecimento apresenta potencial promissor para novas investigações pelo especialista de domínio. Constatou-se também que a área, o perímetro e a compactação dos polígonos representativos de exsudações de óleo apresentam valores majoritariamente intermediários de acordo com a discretização proposta pelo especialista do domínio. Além disso, outras regras fortes sugerem a existência de controles ambientais na forma de polígonos das citadas exsudações. Foi também possível inferir que os mesmos fatores ambientais e de forma controlam indistintamente as exsudações de óleo no Golfo do México (isto é, em Cantarell ou fora dele).

Finalmente, foram também identificadas algumas regras de associação raras, com suporte menor que 10,0%, as quais apresentam valores muito altos de confiança e/ou *lift*. A mais interessante delas relaciona a altura significativa de onda com o ano de 2003, o que pode ser tentativamente justificado pela ocorrência de um evento fraco de El Niño no período de 2002-2003, que pode ter afetado o clima no Golfo do México.

Para a execução de trabalhos futuros de pesquisa nessa linha, recomenda-se: (1) a comparação dos resultados das regras de associação com aqueles obtidos por técnicas clássicas de estatística multi-variada (coeficientes de correlação, análise de principais componentes, etc.); (2) o emprego de estatísticas zonais, associadas ao valor médio zonal do polígono; (3) emprego de dados individualizados por polígono, e não por imagens RADARSAT-1; (4) utilizar lógica nebulosa (*fuzzy logic*) na etapa de discretização dos atributos da base de dados, durante o pré-processamento, esta abordagem permite fronteiras graduais (partição “suave”) ao invés de rígidas.

REFERÊNCIAS BIBLIOGRÁFICAS

- Agrawal, R., Imicliniski, T., Swami, A., 1993, "Mining association rules between sets of items in large databases". In: Buneman, P. & Jajodia, S. (eds), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., Mayn, pp. 207-216.
- Agrawal, R., Srikant, R., 1994, "Fast Algorithms for Mining Association Rules". In: Bocca, J. B, Jarke, M. & Zanolio C. (eds), *Proceedings of the 20th International Conference on Very Large Data Bases, VLBD' 94*, Santiago, Chile, pp. 487-499, Morgan Kaufmann.
- Anuário Estatístico – PEMEX, 2007 [on line]. Disponível em <<http://www.pemex.com/files/content/Anuario2007.pdf>>. Acesso em Abril de 2008.
- Bentz, C. M., 2006, Reconhecimento automático de eventos ambientais costeiros e oceânicos em imagens de radares orbitais. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, Brasil.
- Bentz, C. M., Miranda, F. P., 2001, "Application of remote sensing data for oil spill monitoring in the Guanabara Bay, Rio de Janeiro, Brazil". In: *International Geoscience and Remote Sensing Symposium (IGARSS 2001). Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS 2001)*.
- Brachman, R. J., Anand, T., 1996, "The Process of Knowledge Discovery in Databases": In: *Advances in Knowledge Discovery and Data Mining*. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds). Menlo Park (CA). AAAI. Pp. 37-57
- Brin, S., Motwani, R., Ullaman J.D, *et al.*, 1997, Dynamic itemset counting and implication rules for market basket data. In: Peckham (eds). *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pp. 255-264.

- CSA – Canadian Space Agency [on line]. Disponível em <<http://radarsat.space.gc.ca/>>. Acesso em Junho de 2008.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., 1996a, "From data mining to knowledge discovery: an overview". In *Knowledge Discovery & Data Mining*, Chapter 1, pp. 1-34, AAAI Press.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., 1996b, "Knowledge discovery and data mining: Towards a unifying framework". *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 82-88.
- Félix, L. C. M., 1998, Data mining no processo de extração de conhecimento de base de dados. Dissertação de Mestrado, Instituto de Ciências Matemáticas e da Computação, ICMC/USP, São Carlos, Brasil.
- Glymour, C., Madigan, D., Pregibon, D., Smyth, P., 1997, "Statistical themes and lessons for data mining". *Data Mining and Knowledge Discovery* 1, Kluwer Academic Publishers, pp. 11-28.
- Gonçalves, E. D., 2005, "Regras de associação e suas medidas de interesse objetivas e subjetivas". In: *Journal of Computer Science*, v. 4, pp 26-35.
- Guillaume, S., 2001, "Designing fuzzy inference systems from data: an interpretability-oriented review", In: *IEEE Transactions on Fuzzy Systems*, v. 9, n. 3, pp.426-443.
- Han, J., Kamber, M., 2001, *Data Mining: Concepts and Techniques*. 2^a ed. San Francisco. Morgan Kaufmann Editores.
- Han, J., Pei, J., Yin, Y., 2000, "Mining frequent patterns without candidate generation". *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 1-22. ACM Press.
- Hilderman, R. J., Hamilton, H. J., 1999, "Knowledge discovery and interestingness measures: A survey". Technical Report, Department of Computer Science, University of Regina, Canada.

- Hipp, J., Güntzer, U., Nakhaeizadeh, G., 2000, "Algorithms for Association Rule Mining – A General Survey and Comparison". *Proceedings of the 6th SIGKDD – International Conference on Knowledge Discovery and Data Mining*, pp. 58-64.
- Houtsma, M., Swami, A., 1995, "Set-oriented mining of association rules in relational databases", In: Yu, P. S. & Chen, A. L. P., *Proceedings of the 11th International Conference on Data Engineering*, pp. 25-33.
- Kampel, M., Geata, S. A., Lorenzetti, J. A., Pompeu, M. "Estimativa por satélite da concentração de clorofila a superficial na costa sudeste brasileira, região oeste do Atlântico Sul: Comparação dos algoritmos SeaWiFS", In: XII Simpósio Brasileiro de Sensoriamento Remoto, 2005, Goiânia, Brasil, pp. 3633-3641.
- Levy, E., 1999, The Lowdown on Data Mining [on line]. Disponível em <<http://www.infosystems.eku.edu/loy/cis300/datamining.pdf>>. Acesso em Dezembro de 2007.
- Liu, B., Hus, W., Chen, S., Ma, Y., 1998, "Integrating classification and association rule mining". In: *Proceedings of the 4th Int. Conf. on Knowledge Discovery and Data Mining*. Nova Iorque, pp. 80-86.
- Mannila, H., Toivonen, H., Verkamo, A. I., 1994, "Efficient Algorithms for Discovering Association Rules", *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, pp. 181-192.
- Mata, M. M., Garcia, C. A. E., 1996, "Variabilidade da topografia oceânica superficial no Atlântico Sul Ocidental observada pela altimetria TOPEX/POSEIDON". *Anais VIII Simpósio Brasileiro de Sensoriamento Remoto*, Salvador, Brasil. INPE, pp. 781-786.
- Melanda, E. A., 2005, Pós- processamento de regras de associação. Tese de Doutorado, Instituto de Ciências Matemáticas e da Computação, ICMC/USP, São Carlos, Brasil.
- Mendoza, A., Miranda, F. P., Pedroso, E. C., *et al.*, 2003, "Operational application of

- RADARSAT-1 for the monitoring of natural oil seeps in the Southern Gulf of Mexico”. In International Geoscience and Remote Sensing Symposium, Toulouse, IEEE.
- Miranda, F. P., Marmol, A. M. Q., Pedroso, E. C., *et al.*, 2004, “Analysis of RADARSAT-1 data for offshore monitoring activities in the Cantarell Complex, Gulf of Mexico, using the Unsupervised Semivariogram Textural Classifier (USTC)”. *Canadian Journal of Remote Sensing*, v.30, n.3, pp. 424-436.
- NASA (National Aeronautics and Space Administration), Locus Tutorial Research Project One: Influence of El Niño on the Gulf of Panama Seasonal Productivity Cycle [online]. Disponível em <http://daac.gsfc.nasa.gov/oceancolor/locus/tutorial_1.shtml>. Acesso em Setembro de 2008.
- Park, J. S., Chen, M.-S., Yu, P. S., 1997, “Using a hash-based method with transaction trimming for mining association rules”. In: *IEEE Transactions on Knowledge and Data Engineering*. 813-825.
- Peake, W.H. and Oliver, T. L., 1971, *The response of terrestrial surfaces at microwave frequencies*. Ohio State University Electroscience Laboratory, 2440-7, Technical Report AFAL-TR-70-301, Columbus, Ohio.
- Pei, J., Han, J., Mao, R., 2000. “CLOSET: An efficient algorithm for mining frequent closed itemsets”. In: Gunopulos, D. & Rastogi R. (eds). *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21-30.
- Piatetsky-Shapiro, G., 1991, “Discovery, analysis and presentation of strong rules”, In Gregory Piatetsky-Shapiro and William Frawley, editors, *Knowledge Discovery in Databases*, AAAI/MIT Press, Cambridge, MA, pp. 229-248.
- RADARSAT International (RSI), 1996, RADARSAT Illuminated. Your Guide to Products & Services, RADARSAT user guide produced by RSI.
- Rezende, S. O., Pugliesi, J. B., Melanda, E. A., *et al.*, 2003, “Mineração de dados”. In: S. O. Rezende (eds). *Sistemas Inteligentes: Fundamentos e Aplicações* (1ª ed.), Capítulo 12, Editora Manole, pp. 307-335.

- Roriz, C.E.D., 2006, Detecção de exsudações de óleo utilizando imagens do satélite RADARSAT-1 na porção *offshore* do delta do Niger. Dissertação de M.Sc., COPPE/UFRJ, Rio de Janeiro, Brasil.
- Sabins, F. F., 1997, Remote Sensing: Principles and Interpretation. New York, W. H. Freeman and Company.
- Savasare, A., Omiencink, E., Navathe, S., 1995, "An Efficient Algorithm Mining Association Rules in Large Databases", *Proceedings of the 21th VLDB – Very Large Data Base Conference*, pp. 432-443.
- Silberschatz, A., and Tuzhilin, A., 1996, "What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engeneering* 8, pp. 970-974.
- Silva Junior, C. L., Mano, M. F., Hargreaves, F. M., *et al.*, 2003, "Utilização de dados orbitais multisensor na caracterização de exsudações naturais de óleo no Golfo do México". In: Anais XI Simpósio Brasileiro de Sensoriamento Remoto, SBSR, Belo Horizonte, Brasil, INPE, pp. 929-936.
- Sinoara, A. R., 2006, Identificação de regras de associação interessantes por meio de análises com medidas objetivas e subjetivas. Dissertação de Mestrado, Instituto de Ciências Matemáticas e da Computação, ICMC/USP, São Carlos, Brasil.
- Soler, L. S., 2002, Detecção de manchas de óleo na superfície do mar por meio de técnicas de classificação textural de imagens de radar de abertura sintética (RADARSAT-1). Dissertação de Mestrado, INPE, São José dos Campos, Brasil.
- Srikant, R., Agrawal, R., 1996, "Mining quantitative association rules in large relational tables". *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pp. 1-12.
- Tan, P., Kumar, V., Srivastana, 2002, "Selecting the right interestingness measure for

association patterns". *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canadá, pp.32-41.

Vesecky, J.F., 1995, "Surface film effects on the radar cross section of the ocean surface", In: *Proceedings of the 1995 International Geoscience and Remote Sensing Symposium, IGARSS 1995. Quantitative Remote Sensing for Science and Applications*, Vol. 2, pp. 1375-1377.

Wang, K., Tang, L., Han, J., *et al.*, 2002,"Top down fp-growth for association rule mining". *Proceedings of the 6th Pacific – Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 334-340. Ed.Springer – Verlag.

Webb, G. I., 1995, "OPUS: An efficient admissible algorithm for unordered search". *Journal of Artificial Intelligence Research* 3, pp. 431-465.

Zaki, M. J., Parthasarathy, S., Ogihara M., *et al.*, 1997, "New Algorithm for Fast Discovery of Association Rules". *Proceedings of the Third SIGKDD – International Conference on Knowledge Discovery and Data Mining*, pp. 283-286.

Zaki, M., Hsiao ,C., 2002, "Charm: An efficient algorithm for closed itemset mining". In: Grossman, R., Han ,J., Kumar., V., Mannila, H., & Motwani, R., (Eds.), 2nd *SIAM International Conference on Data Mining*.

Zeng, Z., Kohavi, R., Manson, L., 2001, "Real world performance of association rules algorithms". *Proceedings of the Seventh ACM SINGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, pp. 401-406.