PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM PROCESSO DE AQUISIÇÃO E MAPEAMENTO DE DADOS PARA AS BACIAS SEDIMENTARES MARGINAIS BRASILEIRAS

JOAQUIM VINICIUS CARVALHO ASSUNÇÃO

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Paulo Henrique Lemelle Fernandes

Porto Alegre
2012

A851p Assunção, Joaquim Vinicius Carvalho
Um processo de aquisição e mapeamento de dados para as bacias sedimentares marginais brasileiras / Joaquim Vinicius Carvalho Assunção. – Porto Alegre, 2012.
108 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. Paulo Henrique Lemelle Fernandes.

1. Informática. 2. Banco de Dados. 3. Informática na Geologia. 4. Algoritmos. I. Fernandes, Paulo Henrique Lemelle. II. Título.

CDD 005.74

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Um Processo de Aquisição e Mapeamento de Dados para as Bacias Sedimentares Marginais Brasileiras", apresentada por Joaquim Vinicius Carvalho Assunção como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Processamento Paralelo e Distribuído, aprovada em 27/03/2012 pela Comissão Examinadora:

Prof. Dr. Paulo Henrique Lemelle Fernandes - Orientador — PPGCC/PUCRS

Prof. Dr. Duncan Dubugras Alcoba Ruiz - — PPGCC/PUCRS

Dra. María Alejandra Gómez Pivel - — Pesquisadora FACIN/PUCRS

Homologada em 22/05/2012, conforme Ata No. 011 pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

Agradecimentos

Agradeço a todos aqueles que fizeram parte desta curta e intensa jornada de Mestrado. Foram inúmeras lições que fizeram com que eu crescesse muito durante esses dois anos, tanto em aspectos técnicos e científicos, como humanos e culturais. Realizei aqui, um sonho que abriu margens para uma nova realidade, e consequentemente, novos sonhos. Mas dentre tudo que foi obtido, as novas e boas amizades tem especial destaque.

Em especial agradeço aqueles com que mais convivi e aqueles que tiveram impacto direto na qualidade deste trabalho. Em ordem cronológica...
Ao Prof. Dr. Duncan Ruiz por ter me selecionado e me encaminhado nos primeiros meses de mestrado, e por ter me escolhido para o estágio de docência na disciplina Aspectos Avançados de Banco de Dados.
Ao meu orientador, Prof. Dr. Paulo Fernandes, por ter me aceitado no projeto PaleoProspec; por estar sempre disposto a avaliar novas ideias, e pelas rápidas e francas respostas quanto as mesmas; pelo reconhecimento e boa convivência que fizeram com que esse jornada tenha sido mais agradável.
A Dra. Maria Pivel que teve fundamental importância neste trabalho, agregando qualidade de maneira com que as criações computacionais não ficassem incoerentes com a realidade geológica; pela boa convivência e por sempre estar com boa disposição para tirar dúvidas e debater sobre qualquer assunto.

Aos colegas e mestres: Christian Quevedo, pelas valiosas críticas em minha prévia de defesa; por todas as dicas e modelos de PEP etc. emprestados e principalmente pela criteriosa revisão de minha dissertação.
Luciana Espíndola, por ter atuado como revisora em trechos de minha dissertação; pela presença e atuação crítica em minhas prévias; pela presença, bom humor e dicas de escrita; finalmente, e não menos importante, pela filmagem de minha defesa :)
Luciano Blomberg, pelos modelos emprestados, pelas dicas técnicas e dicas sobre a pós.

Aos demais amigos do GPIN, os quais tive maior convívio no primeiro ano de mestrado: Ana Winck, Juliano Carvalho, Luiz Vaz, Luiz Gonzalez e Peterson Colares; pela amizade, momentos de decontração e dicas sobre o mestrado.
Aos demais amigos do Paleoprospec: Afonso Sales, Claiton Correa, Eli Maruani, Gabriel Couto, Leonardo Peres, Lucas Oleksinski e Thais Webber; pela amizade, boas conversas e momentos de descontração que fizeram o peso dessa jornada se tornar mais leve.

Aos amigos e colegas de HP: Fernando Castilho e Tiago Silva, os quais compartilhei, por alguns meses, as tardes no salão de desenvolvimento da HP. Agradeço-os pela amizade e pela experiência compartilhada.

A todos os que aqui citei e aos demais amigos que indiretamente me deram forças para continuar em dias de excessivas preocupações e incertezas. Obrigado a todos!

UM PROCESSO DE AQUISIÇÃO E MAPEAMENTO DE DADOS PARA AS BACIAS SEDIMENTARES MARGINAIS BRASILEIRAS

RESUMO

Encontrar petróleo é uma tarefa difícil que requer grandes quantidades de informações e recursos. Ao longo de décadas de pesquisa, os geólogos da Petrobras acumularam grandes quantidades de dados. Além disso, na geologia do petróleo outras fontes de dados são importantes, fontes estas, que em geral estão dispersas e possuem várias formas de represenção. Esta dissertação relata a criação de um banco de dados que agrega diversos dados de origem paleoclimática e paleogeográfica provenientes do Atlântico Sul. Grande parte destes dados foram extraídos de cartas estratigráficas, convertidos e armazenados em forma de um modelo numérico. Modelo este, que é resultado de agregações de dados provenientes das bacias sedimentares brasileiras e da criação de uma solução algorítmica capaz de mapear os dados coletados ao longo da área designada. Estes dados são relativos a um período de tempo entre a idade geológica atual até 140 milhões de anos atrás. Os 140 milhões de anos correspondem ao período de deslocamento do continente Sul-Americano desde a costa da África até a posição atual. Durante esse deslocamento houveram diversas mudanças naturais nas bacias sedimentares oceânicas até chegarem ao estado atual. O grupamento destes dados potencializa a descoberta de conhecimento relativo aos fatores necessários para a deposição de matéria orgânica e geração de petróleo no fundo do mar, assim, estes novos fatores podem vir a melhorar as probabilidades de descoberta de petróleo.

Palavras-chave: Geoinformatica; Algoritmos; Banco de Dados; Bacias Sedimentares; Descoberta de Conhecimento.

A PROCESS OF DATA ACQUISITION AND MAPPING FOR BRAZILIAN SEDIMENTARY MARGINAL BASINS

ABSTRACT

Finding oil is a hard task that requires large amounts of information and resources. Along decades of research, geoscientists from Petrobras have accumulated great amout of data. Therefore, in petroleum geology other sources of important data are generally dispersed and have many forms of representation. This Master Thesis reports the creation of a database that stored various geophysical, paleogeographic and paleoclimatic data from the South Atlantic. Great part of these data were extracted from stratigraphic charts, converted, and stored as a numeric model. This model is the result of an aggregation of data from the Brazilian Sedimentary Basins and the creation of an algorithm solution able to map collected data at the designated geographical area. These data cover the past 140 millions years. The 140 millions of years correspond to a drift period of the South American Plate from the African west coast to the present location. During this drift several natural changes happened in the Ocean Sedimentary Basins until they reached the actual state. The grouping of these data enhances the potential to knowledge discovery on the factors necessary for the deposit of organic matter and oil generation, in consequence, these new factors may improve the chances of finding oil.

Keywords: Geoinformatics; Algorithm; Database; Sedimentary Basins; Knowledge Discovery.

Lista de Figuras

Lista de Abreviaturas

KDD	*Knowledge Discovery in Databases*	25
GKD	*Geographic Knowledge Discovery*	25
ESA	*European Space Agency*	29
Lat	*latitude*	29
Lon	*longitude*	29
SGBD	Sistema de Gerenciamento de Banco de Dados	33
SDM	*Spatial Data Mining - Mineração de Dados Espaciais*	38
ETL	*Extract Transform and Load*	39
DW	*Data Warehouse*	40
IMS	*Information Management System*	40
VSAM	*Virtual Storage Access Method*	40
ISAM	*Indexed Sequential Access Method*	40
ANP	Agência Nacional de Petróleo	45
PaleoGeoDB	*Paleo Geographic Database*	49
RPB	Referente a Parte de Bacia	51
PRD	Porcentagem Relativa de Depósitos	51
PPSP	Potencial Papel no Sistema Petrolífero	51
NOAA	*National Oceanic and Atmospheric Administration*	53
LC	*Linha da Costa*	54
L3K	*Linha de -3.000 metros*	54
QP	*Quebra de Plataforma*	54

Sumário

Capítulo 1

Introdução

Quando se fala em exploração de petróleo há uma grande quantidade de dados paleoclimáticos e paleogeográficos a serem explorados e analisados. A estratigrafia e sísmica geram grandes volumes de dados. Esses dados usualmente são gerados e acumulados de maneiras distintas, tanto na representação como na forma e local de armazenamento.

O trabalho relatado nesta dissertação, mostra a criação de um modelo numérico que simula a evolução de dados naturais, bem como um banco de dados que suporta esse modelo. O modelo é gerado através da coleta de diversos dados paleogeográficos e paleoclimáticos. Esses dados correspondem a margem continental brasileira e são relativos à área em que hoje se localiza a parte sul do Oceano Atlântico. Eles são correspondentes a escalas de tempo geológico de modo que é possível perceber a sucessão de sedimentos em um determinado local.

Em Ciência da Computação trabalhamos com tecnologias muito recentes. De fato, a diferença temporal entre as tecnologias atuais e as tecnologias ultrapassadas, em geral, é de apenas alguns anos. Em outras ciências, como a Geologia, dados antigos representam um conhecimento que é importante não apenas para compreendermos fatos passados, mas também (e principalmente) para ajudar a compreender melhor os dados geológicos atuais.

Quando se trata de dados da natureza, a quantidade e a diversidade dos dados tende a ser enorme. O grande volume de dados tende a dificultar analises e extrações de conhecimento. Dois dos motivos pelos quais grandes volumes de dados se tornam complexos, para análise e extração de conhecimento, é a falta de unificação de padrões e as várias formas e locais de armazenamentos.

O trabalho relatado nesta dissertação trata da unificação de locais e padrões de grandes volumes de dados geofísicos. Este volumes são expressivos não somente por abranger uma grande área física, mas também por compor dados de outras idades geológicas.

Para entendermos melhor o presente é preciso estudar o passado, deste modo foram analisados e agrupados dados relativos a um período de 140 milhões de anos atrás. Este período é relativo à separação dos continentes da América do Sul e da África. Nesse intervalo de tempo foram formadas as Bacias Sedimentares Marginais brasileiras como conhecemos hoje.

Um conjunto de dados numéricos mostrando a evolução de elementos geofísicos, bem como a

trajetória da deriva do continente Sul-Americano, abre novas oportunidades de pesquisas. Como consequência de análises nos dados paleoclimáticos e paleogeográficos é possível obter indicadores de probabilidade de estabelecimento de condições favoráveis para a deposição e preservação de sedimentos ricos em matéria orgânica no espaço e no tempo e consequentemente na predição de ocorrência de potenciais rochas geradoras de petróleo.

1.1 Motivação

Mais de 50 anos de pesquisas e explorações realizadas pela Petrobras nas bacias sedimentares da costa brasileira propiciaram o acúmulo de grandes quantidades de informações [Mil07].

Muitos são os meios utilizados para armazenamento destas informações. Os gráficos originados de sísmicas e cartas estratigráficas (ver Capítulo 2) são meios altamente utilizados e que armazenam grandes quantidades de informações.

Contudo, sabe-se que grandes volumes de dados acumulados durante anos tendem a possuir informações ocultas e potencialmente úteis (ver seção 2.4). Especialmente quando se trata de dados geológicos, o potencial para descobertas é muito grande, pois as formações geológicas dos tempos atuais são resultados de longos processos, entre outros, de transformação das próprias matérias da natureza ao decorrer no tempo.

As cartas estratigráficas (para saber mais, veja a seção 2.1.5) agregam grandes quantidades de dados de forma gráfica. Isto, *a priori*, facilita nosso entendimento e ajuda-nos a termos uma visão geral dos dados, porém nos traz uma série de limitações. Estas limitações dizem respeito a certas técnicas estatísticas e computacionais para a descoberta de informações e padrões entre os dados. Para utilização destas técnicas é necessário primeiramente adaptar os dados geofísicos em modelos numéricos.

Além dos dados provenientes das cartas estratigráficas, temos inúmeros dados geofísicos e climáticos que possuem relação indireta com a formação do petróleo (o capítulo 2 exemplifica). Unir estes dados torna mais fácil a visualização das informações e a descoberta de conhecimento.

A organização de modelos numéricos de dados geofísicos também tem como benefício organizar dados atualmente caóticos. A literatura possui vários trabalhos que se destinam a organizar estes dados. Podemos citar Cheng-fang [Fan10], que criou modelos de qualidade para garantir medidas como normalização dos dados, integridade, precisão e segurança dos dados de sísmica. Estes são especificamente alguns dos problemas que os Sistemas Gerenciadores de Banco de Dados (SGBD) foram criados para resolver [Sil97]. Sendo assim, podemos além de gerar conhecimento, organizar grandes quantidades de informações geofísicas.

Segundo Ketzer [Ket10], "Poder prever locais com maior probabilidade de sucesso pode ser muito vantajoso para a estatal brasileira em termos financeiros e de tempo.". De fato, grandes empresas petrolíferas investem em pesquisa afim de melhorar o acerto quanto aos locais a serem perfurados. Para melhorar a predição dos locais e por consequência, minimizar custos, é necessário um banco de dados

sólido que contenha grandes volumes de dados relevantes e dispostos em um modelo numérico para que as técnicas de KDD e GKD possam ser aplicadas. Para criar o modelo numérico são necessárias soluções que atendam as questões de pesquisa. Questões que começam a partir da transformação dos dados geofísicos (imagens, Cartas Estratigráficas etc.) em dados que sejam suportados pelos algoritmos de mineração de dados.

Capítulo 2

Fundamentação Teórica

Este capítulo se propõe a introduzir conceitos relativos a duas áreas do conhecimento. A Seção 2.1, no domínio das geociências, trata sobre os dados geológicos que fazem parte (direta e/ou indiretamente) deste trabalho. Nesta mesma área, há ainda uma síntese sobre a geologia do petróleo e seus principais fatores, tanto em sua composição como em sua representação (Seção 2.2). Na segunda área do conhecimento, computação, serão mostrados conceitos básicos sobre banco de dados (Seção 2.3) e descoberta de conhecimento em banco de dados (KDD) (Seção 2.4), seguido de alguns conceitos básicos sobre algoritmos e ETL.

Primeiramente serão apresentados os conceitos geológicos relativos à formação das placas tectônicas, os quais são a base para a formação e evolução das bacias. Em seguida, serão apresentados os conceitos relativos às coordenadas geográficas, dados paleogeográficos, paleoclimáticos e paleoceanográficos, bem como a função de alguns destes dados na geologia do petróleo. Após, é realizada uma síntese sobre Banco de Dados e Sistemas Gerenciadores de Banco de Dados, juntamente com alguns conceitos de KDD, GKD , ETL e algoritmos.

2.1 Dados Geológicos

Esta seção trata de diversos paleodados e dados geológicos. Estas expressões (paleodados, dados geológicos) são muito utilizadas nesta dissertação, de modo que é importante esclarecer algumas formalidades quanto as mesmas.

O termo “paleo” vem do grego “palaiós” que significa antigo. Em palavras como: paleodados ou paleoclimáticos, o termo “paleo” age como prefixo, indicando referência a algo do passado. Sendo assim, “paleodados” refere-se a dados do passado. Nesta dissertação, “paleodados” refere-se a dados em uma janela de tempo compreendida entre a idade geológica atual até 140 milhões de anos atrás.

Dados geológicos são todos aqueles referentes à geologia. Logo, a grande maioria dos dados geológicos (especialmente aqui utilizados) são paleodados, mas nem todo paleodado é um dado geológico.

Em uma visão geral, esta seção é dividida em subseções que apresentam conceitos importantes sobre os paleodados utilizados. Começando pela sub-seção "Evolução Tectônica e Deriva Continental", os conceitos são apresentados em uma visão "*Top-Down*". Assim, os conceitos da evolução tectônica servem como uma visão geral para uma melhor compreensão das outras seções, que apresentam outros conceitos geológicos que neste trabalho se relacionam com a deriva continental.

2.1.1 Evolução Tectônica e Deriva Continental

Em um período anterior há aproximadamente 200 milhões de anos existia um único continente denominado Pangéia (do Grego, "todas as terras"). Ao longo do tempo, esse supercontinente começou a se fragmentar. Formou Laurásia e Gondwana, que se fragmentaram de modo a formar os continentes como conhecemos hoje [USG09]. A África e a América do sul são resultados da fragmentação do Gondwana. Essa separação teve início há aproximadamente 150 milhões de anos atrás. A figura 2.1 exemplifica.

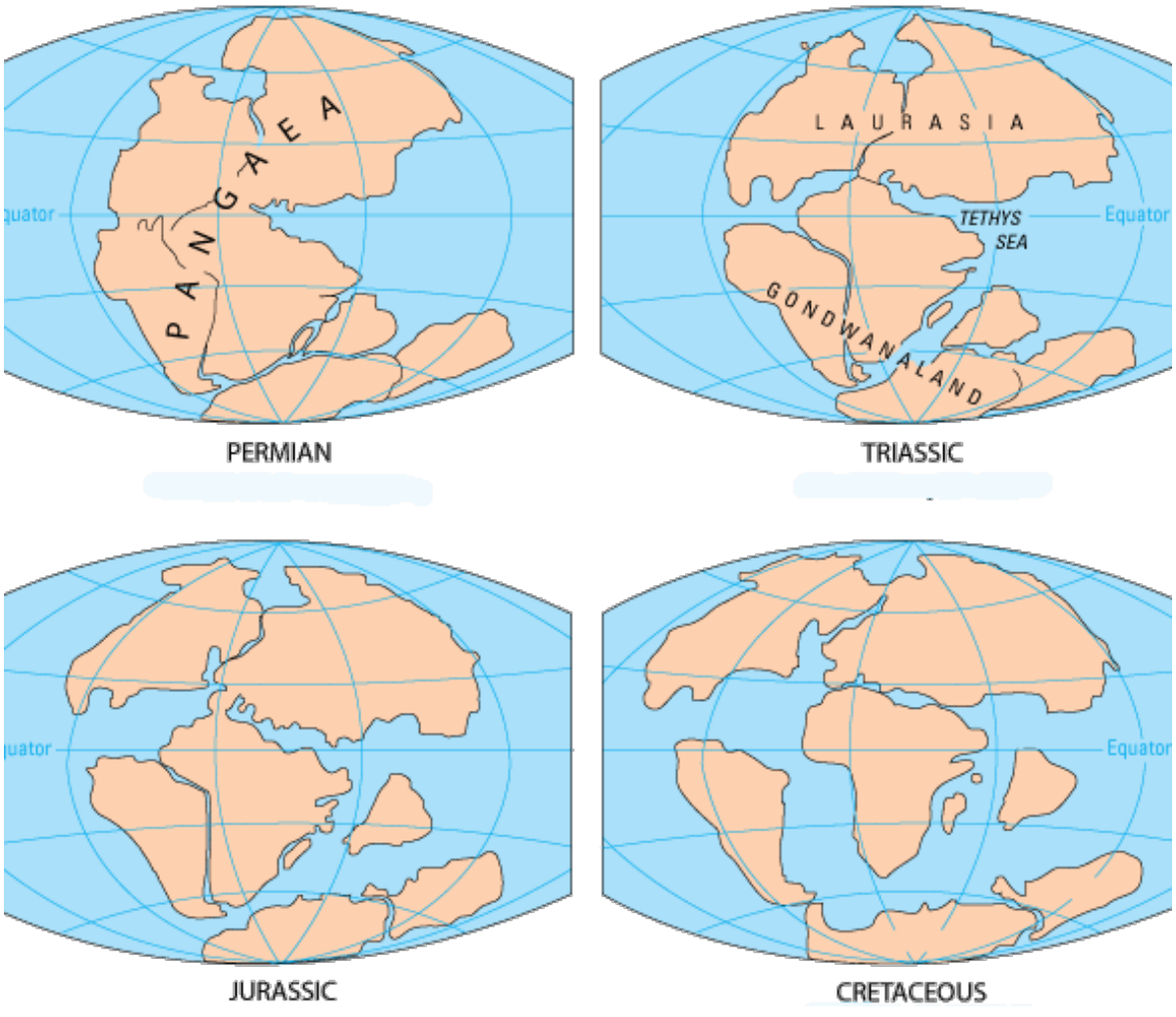


Figura 2.1: Representação esquemática da fragmentação do supercontinente Pangéia

A Teoria da Deriva Continental é a teoria que trata da separação dos continentes. Tida como teoria fundamental da geologia e da geomorfologia, ela foi publicada em 1915 por Alfred Wegener e sofreu várias atualizações até a data atual. Inicialmente refutada por físicos, a teoria de Wegener ficou anos sob debate da comunidade científica. O principal problema eram as hipóteses de Wegener quanto às forças que moviam os continentes, e a velocidade com que eles se moviam [Pre06].

Alguns anos mais tarde, geólogos vieram a provar o encaixe dos continentes devido a similaridades de alguns locais. Rochas similares com a mesma idade foram encontradas na costa leste da América do Sul e na costa oeste da África. Fósseis de alguns dinossauros da mesma espécie também foram encontrados em ambos continentes. O fato de alguns fósseis de répteis, como o Mesosaurus, existirem somente na África (costa oeste) e na América do Sul (costa leste) sugere que os continentes estavam conectados [Pre06].

Em 1960, o geólogo Harry Hess expôs a renovação constante dos assoalhos oceânicos. A ideia principal partira da existência de poucas rochas com mais de 100 milhões de anos, o que sugere que as rochas mais novas se sobrepõem às mais antigas no assoalho oceânico [Hes62]. Em 1965, J. Tuzo Wilson propôs que a ilha do Havaí e outras ilhas vulcânicas teriam se formado pela movimentação das placas sobre pontos quentes do manto da terra. Essa teoria ajudou a identificar o "Círculo de Fogo do Pacífico" juntamente com parte da placa do Pacífico, reforçando a teoria da tectônica das placas. A figura 2.2 mostra as grandes placas que compõem a superfície do planeta. Note os pontos onde há vulcanismo (em vermelho) e o limite noroeste da placa do pacífico [Pre06].

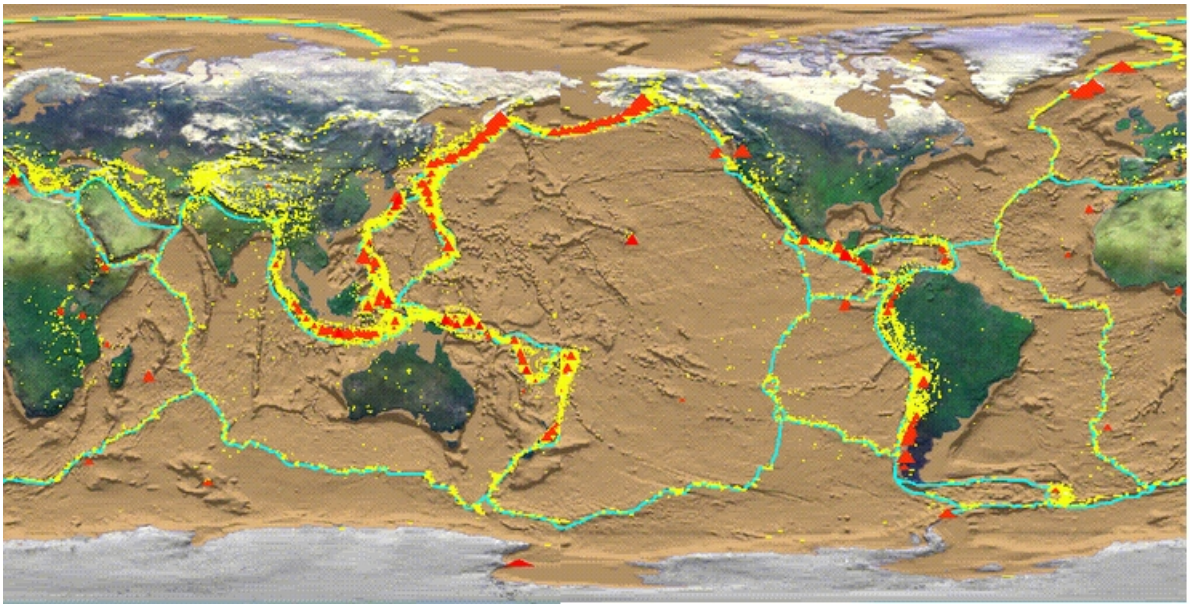


Figura 2.2: Mapa das placas e do relevo da Terra

Existem modelos de como ocorreu a deriva continental. Esses modelos respondem a questões como: Qual foi a trajetória do continente desde a fragmentação de Pangéia, quanto tempo levou para se deslocar uma distância x em um ponto y e uma idade z, etc. Este trabalho visa dar suporte

a uma reprodução numérica para o modelo de Moulin [Mou10], uma opção ao modelo de Müller [Mul08], que segundo o consultor do projeto, Dr. Daniel Aslanian, é um modelo que apresenta menos concordância com as condições de contorno definidas pelos dados de anomalias magnéticas.

As bacias sedimentares brasileiras foram originadas de eventos tectono-estratigráficos [Mil00], que segundo o modelo de Moulin [Mou10], modificaram a Placa da América do Sul de modo a deslocá-la até a posição atual. Enquanto os continentes se separavam, levavam junto consigo as bacias sedimentares, que não apenas se deslocavam, mas também sofriam transformações em sua superfície sedimentar. Nesse processo, a superfície sedimentar do assoalho oceânico relativo às Bacias Sedimentares brasileiras recebeu uma contínua deposição de novos estratos. Isto significa que novos sedimentos se sobrepõem aos mais antigos. Esses sedimentos estão diretamente ligados com a formação do petróleo, pois alguns deles formam (ou auxiliam a formar) os hidrocarbonetos, elementos químicos que compõem o petróleo [Mil00].

Esses hidrocarbonetos são gerados a grandes profundidades e trazidos à superfície através desses eventos geológicos [Jah08].

Os hidrocarbonetos são compostos químicos constituídos apenas de carbono (C) e hidrogênio (H). Essa configuração química permite com que os hidrocarbonetos agreguem átomos de oxigênio (O), nitrogênio (N) e enxofre (S), assim eles são capazes de formar diferentes compostos como o gás natural e o petróleo [Jah08].

2.1.2 Bacias Sedimentares Oceânicas

Segundo Popp [Pop84] as bacias sedimentares são divididas em 8 tipos distintos. No Brasil há apenas 5 tipos delas: Interior (tipo I), Intra continental (tipo II), Rift-Valley (tipo III), Costeira estável (tipo V) e Delta Terciário (tipo VIII). Neste trabalho utilizamos apenas os tipos III, V e VIII, pois as bacias marginais brasileiras se enquadram apenas nestes tipos.

Estas bacias se estendem ao longo da costa brasileira e parte da costa Uruguaia. Começando ao sul pela bacia de Pelotas, que tem início no Uruguai e se estende por todo o Rio Grande do Sul, até a bacia do Foz do Amazonas que vai até a cidade do Oiapoque.

Estas bacias possuem diferentes tamanhos. Sua distância em relação à costa varia muito de bacia para bacia. Grande parte das bacias está dentro do território brasileiro. Este território compreende um espaço de 200 milhas náuticas de distância da costa em direção ao Atlântico. A figura 2.3 mostra as bacias e a linha que marca o limite do território marítimo brasileiro.

2.1.3 Gravimetria e Magnetometria

É sabido que a terra possui um campo gravitacional, e que a gravidade gera a aceleração gravitacional que é a taxa de aumento de velocidade na queda de um corpo. Se a terra possuísse a forma de uma esfera perfeita esta aceleração seria igual em qualquer parte do globo. Porém nosso planeta é um geóide com distribuição distinta de massa [Net04].

Figura 2.3: Bacias e território marítimo brasileiro.

A diferente distribuição de massa, juntamente com a forma da terra (geóide ou esfera imperfeita) gera um campo gravitacional variável. Essa variação pode ser utilizada para encontrar domos de sal, determinar limites de quebra de plataforma e avaliar dados de uma bacia inexplorada [Net04]. A figura 2.4 (obtida em março de 2011 pelo satélite GOCE, da Agência Espacial Europeia (ESA)), mostra o campo gravitacional no globo, onde o vermelho representa uma gravidade maior.

Da mesma forma que a terra possui uma gravidade, também possui um campo magnético. Embora até hoje a origem do campo magnético da terra seja desconhecida, sabe-se que a capacidade magnética de algumas rochas é maior do que outras. Esse conjunto de informações pode ser utilizado para determinar a conformação de uma bacia e detectar corpos metálicos no fundo do mar [Net04].

2.1.4 Sistema de Coordenadas Geográficas

Coordenadas geográficas compreendem um sistema de localização global, onde uma latitude é uma distância angular em relação à linha do equador e longitude é uma distância angular em relação ao Meridiano Inicial (Greenwich).

Valores de Longitudes possuem duas formas de representação. A primeira são valores positivos para pontos a leste de Greenwich que se estendem até 180º e valores negativos para pontos a oeste, que se estendem até -180º. A segunda é uma angulação inteira que aumenta na direção leste, até completar a volta no globo e chegar a Greenwich com 360º [Inp11].

Comumente é utilizada uma representação com letras indicando a localização (***N**orth, **S**outh, **E**ast, **W**est*). Essa forma de representação substitui os sinais. De fato os sinais apenas representam uma

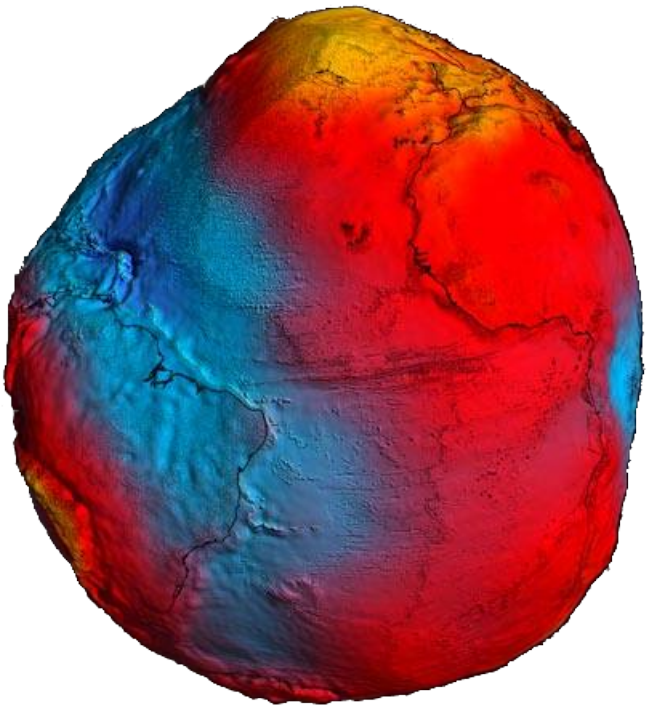


Figura 2.4: Mapa do campo gravitacional da terra.

região. Um número positivo para uma latitude indica o Norte (**N**), assim como um número positivo para a longitude representa o Leste (**E**). Da mesma forma sinais negativos indicam para Lat/Lon representam, respectivamente, Sul (**S**) e Oeste (**W**).

Neste trabalho utilizamos as coordenadas geográficas para marcação dos locais. Cada ponto é uma referência ao estado atual da Terra, logo, se uma coordenada X indica a costa brasileira, a mesma coordenada 'X' há 15 milhões de anos atrás, deve indicar algum ponto dentro do continente Sul Americano.

Em termos de longitude utilizamos a notação que vai de 0 a 360º devido à praticidade, pois a grande maioria dos dados coletados utilizam esta notação.

2.1.5 Cartas estratigráficas

As cartas estratigráficas são ferramentas muito utilizadas no estudo das bacias sedimentares. Esquematicamente ficam evidenciados muitos atributos da área estudada. São eles: A sucessão de estratos e sua representatividade na área estudada e no tempo geológico; a natureza litológica das camadas e as variações laterais de fácies sedimentares; as lacunas na história geológica daquela bacia e muitos outros [Mil07].

Com as cartas estratigráficas é possível saber, de forma visual, a sucessão dos depósitos de sedimentos nas bacias. O processo de criação de uma carta estratigráfica reflete os conceitos dos autores. Ali estão presentes tanto dados reais, como possíveis dados. Deste modo uma carta estratigráfica

possui um processo contínuo de evolução, onde uma área que hoje possui x dados atribuídos, amanhã poderá ter x+1 dados. A figura 2.5 mostra a carta estratigráfica referente à Bacia de Santos.

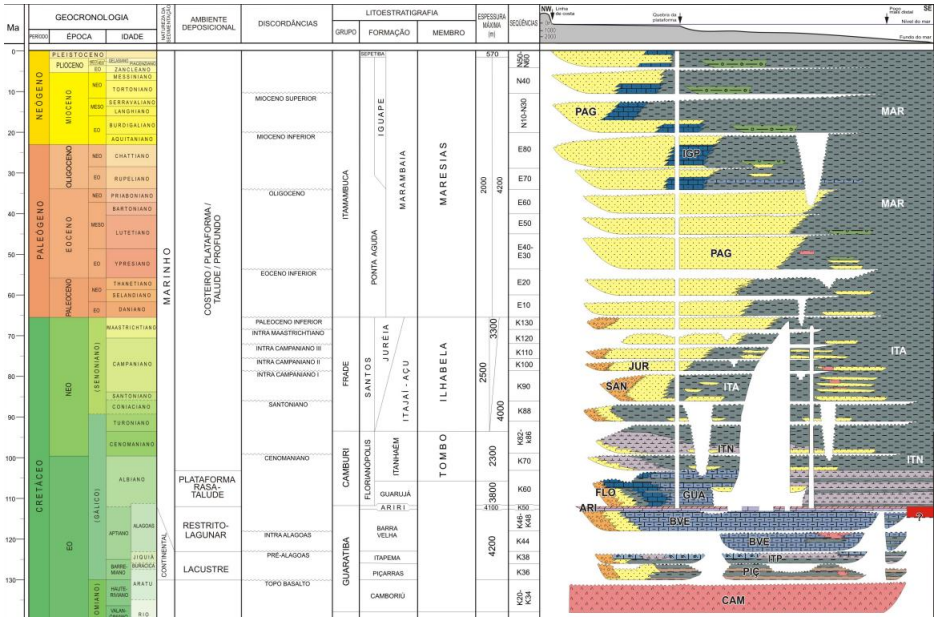


Figura 2.5: Carta Estratigráfica da Bacia de Santos.

O trabalho aqui descrito, devido meramente ao desconhecimento, iniciou com a versão de 2003 das cartas. Assim que descobrimos as versões mais recentes (2007) o trabalho foi reiniciado. Contudo, a pesquisa e a experiência ganha com as cartas antigas serviram de pilares para o trabalho final. A figura 2.6 realiza uma comparação entre as cartas de 2003 e 2007, referentes à bacia de Pelotas, para as idades geológicas mais recentes que a idade Santoniana (aproximadamente 85 milhões de anos atrás).

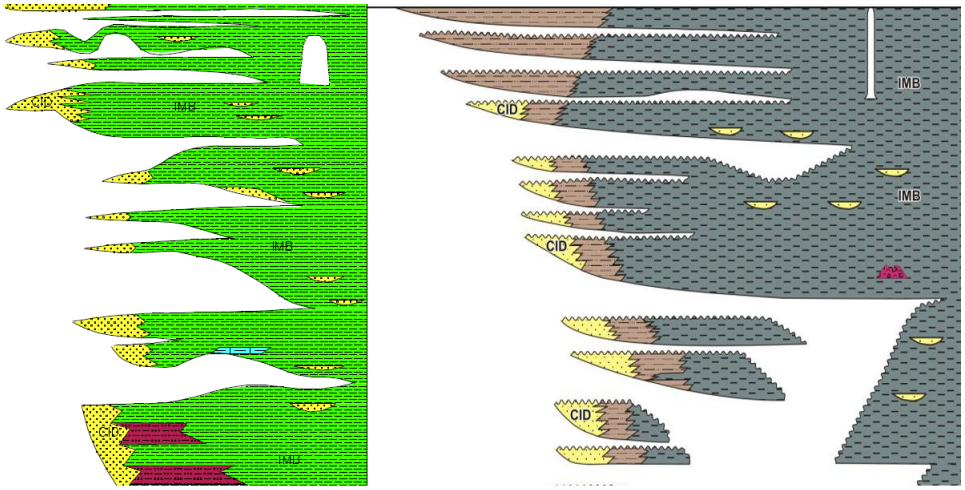


Figura 2.6: Comparação parcial entre as cartas de 2003 e 2007 da bacia de Pelotas.

Como podemos ver na figura 2.6 a diferença entre as cartasé significativa. Além da riqueza de detalhes, as novas cartas possuem algumas divergências com as cartas mais antigas. Assim é de fundamental importância que o repositório das informações seja flexível para a atualização dos dados.

2.2 Geologia do Petróleo

A geração e acúmulo de hidrocarbonetos requer uma série de elementos e processos essenciais. Sabemos que é necessária a existência de um depósito sedimentar rico em matéria orgânica como fonte dos hidrocarbonetos (rocha geradora), uma rocha porosa capaz de armazenar os hidrocarbonetos (rocha reservatório), uma rocha que impeça a fuga do fluido (rocha selo) e uma trapa (ou armadilha) que permita o acúmulo.

Em condições favoráveis, o hidrocarboneto da rocha geradora é submetido a elevada pressão e temperatura e o fluido migra até a rocha reservatório onde é finalmente trapeado. Porém, a simples presença destes elementos (os diferentes tipos de rocha necessários) não garante a existência de reservas de hidrocarbonetos, uma vez que é fundamental que a origem e desenvolvimento de cada um dos elementos e os processos sigam uma ordem temporal favorável.

Grande parte das informações são representadas nas cartas estratigráficas, porém, a ordem temporal utilizada pelos elementos leva em consideração outros processos para a formação dos hidrocarbonetos. Pressão e temperatura são dois fatores importantes que também devem ser considerados.

2.2.1 Formação das Rochas

Como vimos na seção acima (2.2), as rochas são essenciais para a formação do Petróleo. Dentre os 3 tipos de rochas, as essenciais são as sedimentares, pois fazem parte do processo, porém as ígneas e metamórficas podem eventualmente ser importantes como reservatórios para o óleo. Esta seção inicia com um breve resumo sobre a formação das rochas. A seguir, as subseções mostram algumas particularidades dos três tipos de rocha. Os três tipos de rochas possuem subclassificações que são características possivelmente úteis na aplicação das técnicas de mineração de dados.

Há milhões de anos atrás quando a Terra era poeira cósmica, em torno dos 3.000^{o}C, algumas substâncias começaram a liquefazer-se. O ferro liquefeito começou a formar o núcleo, o silício e os óxidos metálicos começaram a formar o manto. Quando esta temperatura começou a baixar, a crosta começou a se solidificar [Pop98].

A solidificação da crosta gerou as primeiras rochas. Estas rochas são classificadas como rochas ígneas ou magmáticas por originarem-se do magma, que consiste nestas mesmas rochas fundidas a temperaturas entre 800 a 1.500^{o}C [Pop98][Pre06].

Quando a crosta da Terra esfriou a uma temperatura de 374^{o}C, o vapor da atmosfera começou a se condensar em chuva, o que posteriormente formou os primeiros mares. As rochas sedimentares se originaram por consequência da ação das águas que reduziam as rochas ígneas à fragmentos e faziam

com que esses fragmentos se consolidassem, criando as primeiras rochas sedimentares.

As rochas que ficavam presas em altas temperaturas na superfície da Terra (grandes profundidades, mas ainda no manto) e sofriam efeitos de forte pressão acabavam sofrendo metamorfismo ao passar dos anos. Estas são classificadas como rochas metamórficas.

Rochas Sedimentares

Camadas de partículas que encontramos com abundância na superfície terrestre, como a areia e conchas de organismos são alguns precursores de rochas sedimentares. Essas partículas formam-se na superfície de restos de rochas que vão sendo alteradas e erodidas por meio de intemperismos [Pre06]

Intemperismos são fenômenos físicos e químicos que levam à degradação de uma rocha. Juntamente com a erosão, o intemperismo produz dois tipos de sedimentos, os clásticos e os químicos e biológicos [Pre06].

Basicamente, a diferença entre sedimentos clásticos e químicos/biológicos é como eles são gerados. Os clásticos são gerados através da fragmentação e retrabalhamento de fragmentos de rocha. Os químicos e biológicos são produzidos por meio de intemperismos e reações biológicas locais [Pre06]

Como exemplo de rocha sedimentar clástica, pode-se citar o **Folhelho**. Ele possui uma grande importância econômica já que é um potencial gerador de hidrocarbonetos [Jah08].

2.3 Banco de Dados

Um banco de dados constitui um conjunto de registros dispostos em uma estrutura que possibilita a reorganização dos mesmos. Atualmente quando falamos em banco de dados, falamos em SGBD . Além do banco de dados em si, os SGBDs provem a interface necessária para realizar as mais diversas operações possíveis em um banco de dados.

Um SGBD é constituído por um conjunto de dados associados a um conjunto de programas para acesso a esses dados [Sil97]. Segundo Silberschatz [Sil97] o principal objetivo de um SGBD é ser tanto conveniente, como eficiente para recuperação e armazenamento dos dados.

Os SGBDs apresentam grandes vantagens em relação a outras formas de armazenamento como planilhas e arquivos de texto. Com SGBDs é mais fácil de evitar alguns problemas que ocorrem quando se trabalha com grandes volumes de dados. São eles: inconsistência dos dados; redundância dos dados; dificuldade de acesso aos dados; problemas de integridade; problemas de atomicidade; anomalias de acesso concorrente e problemas de segurança [Sil97].

Um banco de dados pode ter outra classificação de acordo com suas características. O trabalho relatado nesta dissertação constitui um banco de dados temporal. Estes se diferenciam dos bancos convencionais pela presença de dados do passado e/ou futuro [Tan93].

Os bancos de dados convencionais são projetados para capturar os dados mais recentes [Tan93]. Eles possuem uma estrutura relacional formada por tuplas e atributos que podem ser visualizados em

duas dimensões. Essa relação também é conhecida como *snapshot relation*, pois captura uma imagem da realidade.

Nos bancos de dados temporais a estrutura é formada por 3 dimensões. As duas dimensões do modelo relacional mais a dimensão tempo. Essa estrutura também é chamada de *time cube* [Tan93]. Nesse trabalho possuímos muitos dados temporais representados. Resumidamente, mapeamentos de litologias em uma latitude e uma longitude, ao decorrer de uma idade geológica, ligados a valores de batimetria e anomalias gravimétricas.

2.4 Descoberta de Conhecimento em Banco de Dados (KDD)

Análises com base em informações diversas é uma prática comum. Com o avanço da tecnologia, houve um grande avanço na coleta e armazenamento de dados. Como consequência, os bancos de dados tornam-se cada vez maiores. Assim, da mesma forma que os dados são acumulados, informações ficam ocultas em meio aos grandes volumes de dados.

Analisar os dados e informações tem se tornado uma tarefa mais complexa e demorada. Devido a estes motivos, técnicas computacionais e algoritmos para análise de dados foram criados. O KDD é um processo que compreende várias etapas de análise de dados, visando descobrir informações previamente desconhecidas.

Segundo Fayyad et al. [Fay96] a Descoberta de Conhecimento em Banco de Dados (KDD) é um processo não trivial de identificar padrões interessantes de dados. Han e Kamber [Han01] definem padrões interessantes como aqueles que são facilmente entendidos por humanos, são válidos com um certo grau de certeza, são potencialmente úteis e previamente desconhecidos.

Devido ao grande crescimento no volume de dados das bases atuais, técnicas de KDD tornam-se cada vez mais necessárias para se obter conhecimento em meio a informações dispersas. As aplicações do KDD se espalham por diversas áreas do conhecimento. Astronomia, negócios de marketing, detecção de fraudes, investimentos e telecomunicações são alguns exemplos de áreas nas quais o KDD é utilizado.

Processos de KDD não estão presentes no banco de dados final (aqui descrito), porém um dos principais objetivos deste banco de dados é tornar os dados paleogeográficos mais compatíveis com os processos de KDD, e por consequência, permitir novas descobertas a partir dos dados envolvidos. Deste modo serão apresentados os conceitos e técnicas relacionados ao KDD.

Em uma visão geral Fayyad [Fay96] define KDD nas seguintes partes como mostra a figura 2.7. Há ainda uma definição de Han e Kamber [Han01], ligeiramente distinta, que define um Data warehouse logo no início do processo.

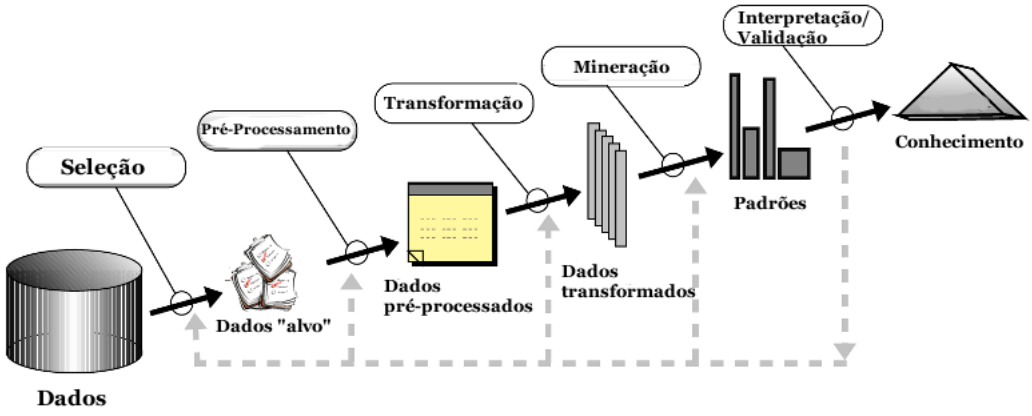


Figura 2.7: Etapas de um processo de KDD segundo Fayyad [Fay96].

2.4.1 Data Mining

A Mineração de Dados (*Data Mining*) é comumente é confundida com KDD, porém mineração de dados é uma parte de todo processo de KDD [Mil09] [Han01] [Tan05].
Segundo Hand et al. [Han01B] Data Mining é uma nova disciplina concebida da intersecção de várias disciplinas como a estatística, banco de dados, reconhecimento de padrões e inteligência artificial.
Fayyad et al. [Fay96] define Data Mining como "aplicação de algoritmos específicos para extração dos padrões de dados".
Tan et al. [Tan05] define Data Mining como uma parte integral do KDD, onde o processo como um todo visa descobrir informação útil em dados brutos.

Em meio a execução de um processo de mineração de dados, são utilizados algoritmos que trabalham os dados de acordo com as suas configurações previamente estabelecidas e o propósito do processo de KDD a ser realizado. Matheus et al. [Mat93], em 1993 tentou classificar os algoritmos de mineração em quatro classes:

- Identificação de classes: Com base em similaridade entre os registros, o algoritmo os agrupa em diferentes classes.
- Classificação: Encontra regras que identificam características de uma determinada classe.
- Análise de dependência: Encontra regras que predizem o valor de um atributo com base no valor de outro atributo.
- Detecção de desvio: Descobre desvios quanto a uma característica esperada e objetos fora de um grupo a qual deveriam pertencer (*outliers*)

Essa classificação, proposta em 1993, foi alterada pela comunidade científica. Hoje os algoritmos de detecção de desvio, por exemplo, são chamados de algoritmos de segmentação de dados. Existem

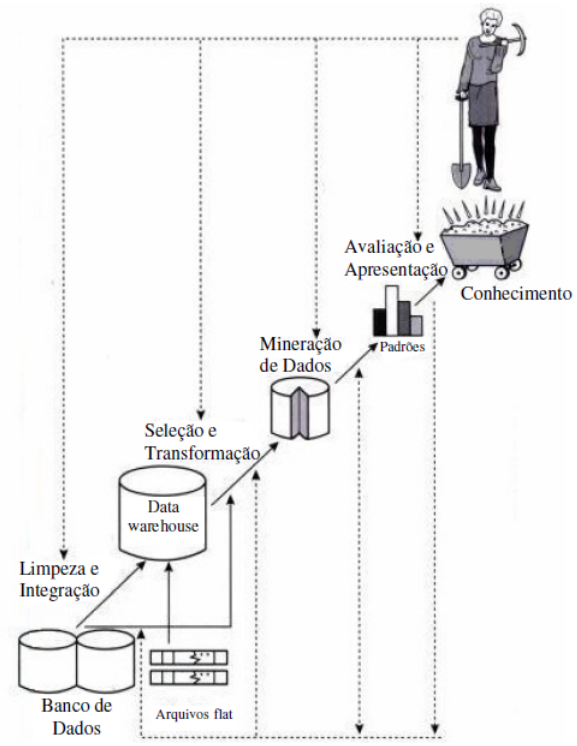


Figura 2.8: Etapas de um processo de KDD segundo Han e Kamber [Han01].

diversos tipos de algoritmos de mineração de dados, dentre os mais importantes [Han01], [Tan05] pode-se citar:

- Algoritmos de classificação: Prevêem variáveis discretas, com base em outros atributos do conjunto de dados.
- Algoritmos de regressão: Prevêem variáveis contínuas, como lucro ou perda, baseando-se nos outros atributos do *dataset*
- Algoritmos de segmentação (*clustering*): Dividem dados em grupos de itens que têm propriedades semelhantes.
- Algoritmos de associação: Encontram correlações, que podem gerar regras de associação, entre atributos diferentes em um conjunto de dados.

Na sessão 4.3.5 é apresentado um plano de mineração para os dados obtidos. Este plano baseia-se em duas técnicas de mineração de dados, são elas: Classificação e Associação. Para que o plano possa ser entendido é necessário compreender alguns conceitos sobre estas técnicas.

Classificação: Segundo Tan et al.[Tan05], classificação em mineração de dados é a tarefa de classificar objetos em uma de várias categorias pré-definidas. Um exemplo clássico é a classificação de galáxias conforme sua forma (espiral, elíptica, etc.). A figura 2.9 mostra o modelo de uma atividade

de classificação de dados, onde cada atributo x, proveniente de um conjunto de atributos, ao passar pelo modelo pré-definido de classificação, é atribuído a uma classe

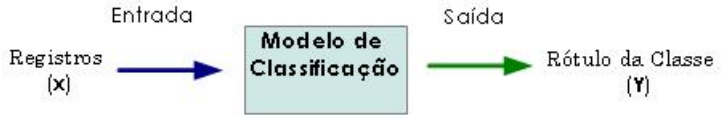


Figura 2.9: Modelo genérico de uma tarefa de classificação.

Modelos de classificação por sua vez possuem duas classificações que são definidas quanto ao seu objetivo. Estes modelos podem ser descritivos ou preditivos.

Modelos descritivos servem para explicar quais características de um registro o incluem em uma determinada classe. Um exemplo clássico é um conjunto de dados que possua maus pagadores e bons pagadores. Então, a partir deste *dataset*, podemos definir quais são as características que distinguem maus pagadores dos bons pagadores com um determinado grau de precisão. Assim, com o resultado de uma classificação por modelos descritivos podemos gerar árvores de decisão que nos ajudem a escolher indivíduos, como por exemplo, os bons pagadores.

Modelos preditivos ajudam a classificar indivíduos em uma determinada classe. Imagine que surja uma nova espécie de animal. A partir de uma árvore de decisão, concebida com um modelo descritivo de classificação, podemos inferir a classe do animal (mamífero, réptil, etc.).

Regressão - Segundo Tan et al.[Tan05], regressão é uma técnica de modelagem preditiva, onde a variável a ser estimada é contínua. Formalmente, regressão é a tarefa de aprendizagem de uma função f que mapeia cada conjunto de atributos x em uma saída contínua y. Assim, a meta da regressão é encontrar uma função que suporte os dados de entrada com um erro mínimo.

Segmentação - Segmentação (*clustering*) é uma técnica para separar os dados em grupos distintos de acordo com suas características. Algoritmos de segmentação são extremamente úteis em diversas áreas, seja para separar dados de modo a facilitar sua manipulação ou para uní-los por utilidade prática [Tan05].

Na Biologia, por exemplo, a segmentação é utilizada em grandes bancos de dados de DNA para encontrar similaridades em grupos de genes, e separá-los de acordo com suas similaridades. Na medicina, a segmentação pode ajudar a detectar padrões entre doenças, isolando seus fatores e características. Além das áreas científicas, a segmentação também pode ser utilizada em empresas para classificar clientes em grupos distintos e assim oferecer produtos mais propícios ao perfil do cliente.

Associação - Uma regra de associação é uma expressão implícita na forma $X- > Y$, onde a força da regra é determinada pelas variáveis: suporte e confiança. O suporte determina a quantidade de ocorrências que contém os itens X e Y, ou seja, representa a relevância da regra. Já a confiança determina a frequência de Y em relação a X [Tan05] [Han06]. Formalmente suporte e confiança são determinados por:

$$Suporte, s(X->Y) = \frac{\sigma(x \bigcup Y)}{N}$$
$$Confianca, c(X->Y) = \frac{\sigma(X \bigcup Y)}{\sigma(X)} \quad (2.1)$$

2.5 Descoberta de Conhecimento Geográfico (GKD)

Geographic Knowledge Discovery é um tipo especial de KDD. Segundo Miller [Mil09], uma das diferenças entre o GKD e o KDD está na relação entre os dados que estão presentes nas dimensões. No processo de KDD suas várias dimensões são relativamente independentes, enquanto que no GKD as várias dimensões de dados geradas são inter-relacionadas e possuem medidas padrões entre as dimensões [Mil09]. Outra particularidade está na existência da etapa de mineração de dados no KDD, enquanto que, no processo de GKD a etapa similar é a mineração de dados espaciais.

2.5.1 Spatial Data Mining

Spatial Data Mining (SDM) se difere do Data mining convencional por trabalhar com dados espaciais. Segundo Shekhar [She01] extrair padrões interessantes e úteis de Bases de Dados Espaciais é mais difícil do que a mineração em dados convencionais devido a complexidade do relacionamento dos dados espaciais.

Miller e Han [Mil09], classificam as regras de mineração de dados espaciais em cinco categorias, são elas:

- Associações espaço-temporais
- Generalizações espaço-temporais
- Segmentação de dados espaço-temporais
- Regras de evolução
- Meta regras

Associações espaço-temporais são similares a regras de associação utilizadas na mineração de dados convencionais, onde a ocorrência de x é seguida pela ocorrência de y em $z\%$ das vezes. A diferença na associação espaço temporal é que o foco é alterado dos dados em si, para a alteração dos dados no espaço-tempo.

Generalizações espaço-temporais é a agregação de dados segundo padrões em sua hierarquia. Por exemplo, um processo de generalização baseado em dados paleoclimáticos de Porto Alegre poderia gerar a seguinte regra: "Verões em Porto Alegre são quentes e secos.".

Segmentação de dados espaço-temporais é muito similar a segmentação de dados convencional. Ambas compartilham a mesma ideia de dividir indivíduos de acordo com sua característica, porém devido as dimensões extras, a dificuldade, bem como a possibilidade de se obter informações é muito maior.

Regras de evolução são regras que se aplicam especialmente no domínio espaço-temporal. Elas descrevem a maneira como entidades se comportam ao longo do tempo. Essas regras tendem a gerar um volume muito grande de dados, por este motivo é indicado o uso de dados e variáveis de controle antes do experimento.

Podemos supor a seguinte situação no contexto da evolução litológica do Atlântico Sul, a seguinte regra de evolução gerada para uma rocha x, potencial geradora de petróleo no local (bacia) y.

Dados: y: Bacia ou parte de uma bacia sedimentar costeira;
x: Uma área com rocha potencialmente geradora de petróleo no local y;
$p1, p2, p3, p4, p5$: Pontos quaisquer;
t: Tempo geológico;
m: Valor qualquer, negativo e em milhões de anos.

A seguinte regra de evolução poderia ser gerada: "um segmento x apresenta uma rota espaço-temporal similar dentre espaços de tempo t; após t>m, x se estende em direção aos pontos $p1$,$p2$,$p3$,$p4$,$p5$".

Meta regras são regras que geram resultados com base na análise das regras principais. Em outras palavras, elas servem para descrever dados sobre as outras regras aplicadas e seus resultados.

Um dos objetivos da base de dados descrita nessa dissertação é simplificar o processo de descoberta de conhecimento. Uma vez que a estrutura do banco permite que dados presentes no espaço geográfico sejam mapeados como dados comuns, o *Spatial Data Mining* se torna mais simples. Abraham e Roddick [**?**], já previam que bancos temporais poderiam ser abstraídos e utilizados como bancos convencionais, porém, é indicado a utilização de um processo de GKD para se obter melhores resultados.

2.6 Algoritmos e ETL

Na seção 2.3 foram apresentados, de forma condensada, conceitos sobre banco de dados e SGBDs. Nas seções 2.4 e 2.5 foram abordados conceitos necessários para o entendimento de KDD e GKD. Esta Seção apresenta conceitos básicos sobre processos de Extração, transformação e carga (ETL) de dados, juntamente com conceitos de algoritmos, que neste trabalho foram criados para o mapeamento e carga dos dados.

Quando possuímos um grande volume de dados a serem gerados, e neste caso, mapeados automaticamente; bons algoritmos se tornam cruciais. Sedgewick [Sed98] define algoritmos como métodos solucionadores de problemas e adequados para implementação em um programa de computador. Quanto maior a quantidade de dados e maior o número de *loops*, os programas tendem a possuírem um custo computacional maior. Assim, para transformar e carregar grandes quantidades de dados o

desempenho torna-se fundamental.

Se por um lado uma boa estrutura, com bons algoritmos se traduz em ganhos de performance, e por consequência, economia de tempo. Por outro lado uma estrutura e algoritmos ruins, podem inviabilizar processos. Para tratamento dos dados desde a coleta até a carga no banco da dados é comum a criação de ferramentas que automatizam processos. O processo de extração, transformação e carga de dados, geralmente é utilizado para montar um *Data Warehouse* ou um *Data Mart*.

O processo de ETL é dividido em 3 partes. A primeira parte consiste em extrair dados de fontes externas que podem estar em diversos meios e formatos. Geralmente estes dados provem de estruturas relacionais de um banco de dados, porém também é comum estarem em forma de texto puro, provenientes de relatórios, vindos de web sites, etc. Estes formatos também podem vir de estruturas não relacionais de banco de dados, como Sistemas de Gestão da Informação (IMS) , Métodos de acesso de armazenamento virtual (VSAM) ou Métodos de acesso sequencial indexado (ISAM) [Kim04].

A parte de transformação é altamente dependente da parte de extração, sendo que, quanto mais dados e mais diversificadas as fontes, maior tende a ser o processo de transformação dos dados. A transformação também é muito dependente do que se precisa no banco de dados, alguns exigem formatos mais específicos, o que demanda mais conhecimento das necessidades técnicas e de negócio [Kim04].

Por fim ,a parte de carga dos dados compreende o processo de carregar os dados extraídos e transformados para o banco de dados. Essa fase demanda iteração direta, na maioria dos casos, com o DW. Isto significa que as estruturas de banco de dados e DW tem de estar bem formadas para receber os dados, bem como as ferramentas de ETL tem de estar de acordo com as necessidades do projeto.

Um DW é um conjunto de dados orientada a assunto, não volátil, integrado e variante no tempo que provem suporte para tomada de decisão [Kim02]. Em outras palavras, um DW é um conjunto de dados selecionados de um banco de dados, organizados de forma orientada ao assunto de maneira a facilitar consultas e possibilitar a visualização de informações de forma rápida.

De certo modo, o processo de ETL para criação de um data warehouse tem o mesmo perfil do processo de ETL para a criação do banco de dados descrito nesta dissertação. Em ambos os processos, os dados são coletados objetivando organizá-los. Porém, neste trabalho, o processo de ETL visa extrair informações de diversas fontes de dados (diferente do DW, onde geralmente os dados são coletados de uma ou mais base de dados) para organiza-los em um banco de dados.

Do mesmo modo que processos de KDD passaram a utilizar algoritmos e softwares que automatizam parte do processo, os processos de ETL também evoluíram. Ferramentas para ETL são necessárias devido ao grande volume de dados. Quando os dados são de diversos tipos e formatos, a ETL torna-se ainda mais importante.

Capítulo 3

Questões de Pesquisa

O mapeamento de dados em um plano espacial altamente sinuoso, como é a margem continental brasileira, se mostra um grande desafio computacional. Vários fatores estão envolvidos, pois o mapeamento envolve diversos dados que devem ser interligados, e fatores que devem possuir concordância entre si (ver capítulo: Fundamentação Teórica 2). Assim, a coleta, a padronização e a normalização dos dados, são algumas das tarefas precedentes e necessárias para a criação de um bom modelo. O modelo, que por sua vez é necessário e indispensável para o mapeamento adequado dos dados.

Como questões de pesquisa, em geral, podemos resumir a algumas perguntas. Como representar dados paleogeográficos e paleoclimáticos extraídos de diversas fontes, de maneira com que todos estejam interligados entre si a ponto de representar a evolução tectônica continental em um período de 140 milhões de anos? Como manter representados de maneira fiel aos dados originais, e por consequência a realidade geológica, os dados transformados e interligados? Como tornar o resultado final, um banco de dados que seja prático para aplicar técnicas de KDD?

A quantidade de informações visuais, remete-nos à seguinte questão de pesquisa: "Como organizar tantas informações paleogeográficas a fim de tornar viável a utilização de técnicas de descoberta de conhecimento em banco de dados?".

3.1 Cenário de Pesquisa

As questões de pesquisa descritas no inicio do capítulo, foram planejadas e trabalhadas ao longo do desenvolvimento do trabalho. Poucas referências foram encontradas e com subáreas ligeiramente parecidas, alguns trabalhos serviram como fundamentação teórica. Porém, possivelmente devido à natureza deste trabalho, não foram encontrados trabalhos relacionados com grande relevância a ponto de ajudar na metodologia, na criação do modelo do banco de dados ou no mapeamento dos dados.

Inicialmente, um processo de transcrição dos dados foi realizado com as cartas estratigráficas das quatro bacias sedimentares mais ao sul da costa brasileira (detalhes na seção 4.1).

Essas cartas foram trabalhadas para expandir seus dados por toda a região sul da costa de forma

a completar os dados faltantes. Foi realizado ainda um processo de KDD nos dados obtidos e desse processo foram gerados alguns resultados, como por exemplo, coordenadas geográficas com possíveis reservas de petróleo.

Este primeiro projeto serviu de *startup* (e para testes). Dentre os problemas de pesquisa encontrados, os problemas com a atualidade dos dados litoestratigráficos e como eles devem ser representados, se destacaram dentre os demais.

As incertezas quanto às informações geofísicas que foram obtidas se mostraram agravantes. Ao longo que as primeiras questões de pesquisa eram respondidas, novas questões eram geradas. Qual é a precisão dos dados originais? Quais os limites de precisão do mapeamento para que os mesmos sejam considerados realísticos? Existem meios presentes na literatura para mapear dados desta natureza? Qual seria uma boa maneira de mapear milhões de dados de maneira fiel e automática?

O capítulo 4, constitui-se no relato do desenvolvimento que produziu os resultados deste trabalho, da pesquisa e de todo processo realizado com os resultados da mesma. Nele será possível compreender como o banco de dados foi populado, quais as metodologias utilizadas e qual o resultado final.

Capítulo 4

Desenvolvimento

Este capítulo se propõe a introduzir as etapas de desenvolvimento deste trabalho desde a criação do primeiro modelo do banco de dados até o banco de dados final, bem como sua integração com o modelo estrela (formato para DW e aplicação de processos de KDD).

O banco de dados aqui descrito é uma evolução de outros modelos primitivos que sofreram uma série de alterações e foram divididos, basicamente, em três grandes etapas. As duas primeiras etapas são relatadas com o intuito de mostrar a evolução do trabalho e facilitar a compreensão do motivo de algumas escolhas relativas ao modelo final, que está na terceira etapa. Este capítulo também aborda a solução desenvolvida para mapeamento dos dados estratigráficos, bem como a ferramenta criada para realizar o mapeamento desses dados em áreas sinuosas como a costa brasileira.

4.1 Introdução ao problema

Como descrito na seção 2.1.5, as cartas estratigráficas são de fundamental importância para entender a formação dos hidrocarbonetos, e por consequência as atividades petrolíferas. Assim, as cartas estratigráficas foram o ponto de partida deste trabalho.

Segundo Wang [Wan08] as atuais bases de dados geográficas são modeladas sem levar em consideração possíveis processos de KDD. Isso gera problemas quando a base de dados começa a ter um grande volume, pois torna-se difícil encontrar conhecimento em meio a dados dispersos [Mil09]. Assim, para chegar em um modelo robusto e passível de processos de KDD seguimos as etapas de KDD sugeridas por Fayyad et al.[Fay96].

Contudo o modelo se mostrou incompleto para o preenchimento das informações. Então, mais uma etapa foi criada, a etapa de estimativa de dados, que ficou entre as etapas de *transformação* e *mineração*. A figura 4.1 ilustra as etapas do processo e a seguir são descritas as atividades e questões de pesquisa de cada etapa.

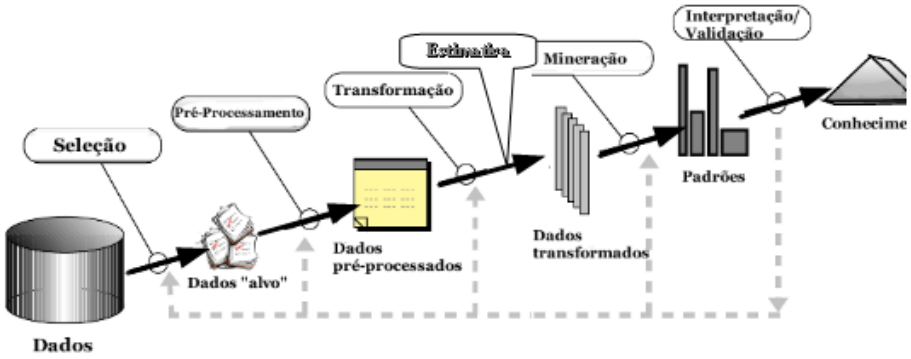


Figura 4.1: Modelo das etapas de KDD aplicadas aos processos.

4.1.1 Seleção dos dados

Com o foco especificamente nas cartas estratigráficas, em um procedimento inicial e incerto, devido aos nossos conhecimentos em geociências o trabalho foi limitado a quatro bacias.

Foram coletados dados de quatro das bacias sedimentares mais ao sul da costa brasileira. Estes dados foram coletados das cartas estratigráficas de 2003, devido ao nosso desconhecimento das cartas mais recentes. Essa seleção de dados serviu para iniciar um trabalho prévio para o banco de dados litoestratigráfico (descrito na seção 4.2).

A vantagem de um trabalho prévio é que ele gera experiências, assim, hipóteses e procedimentos podem ser validados. Toda a experiência ganha, serve como base para planejar o próximo modelo. Um modelo mais robusto e construído em menos tempo devido a redução na taxa de retrabalho.

Como forma de mapeamento para os dados selecionados foram utilizadas coordenadas geográficas (Lat-Lon). Os dados de batimetria da respectiva área também foram coletados. Esses dados foram obtidos do site TOPEX [Smi10], o qual informa coordenadas com precisão de um minuto.

4.1.2 Pré-Processamento

Levando em consideração o limite territorial brasileiro, que é de 200 milhas náuticas, optamos por dividir o *eixo x* da carta em quadrantes de 10,01Km. Assim obtivemos 37 quadrantes entre a costa e o limite territorial brasileiro.

Como nosso foco é a criação de um banco de dados para utilização de técnicas de GKD e KDD, surgiu a primeira necessidade. Como representar as litologias de forma numérica? A primeira solução, a representação por potência de dois, se mostrou prática, pois com ela várias litologias podem ser representadas com um valor. Isto também é prático para decompor os valores e por consequência obter as litologias que compõem o território. A figura 4.2 mostra um exemplo.

Figura 4.2: Exemplo de representação com potências de dois.

4.1.3 Transformação

A transformação dos dados se mostrou uma etapa relativamente curta. Basicamente os dados foram selecionados para criar um arquivo .arff, formato utilizado pelo WEKA [Hal09].

Como atributo classe, foi criado um atributo formado por outros dois atributos. O primeiro atributo se refere as áreas licenciadas pela Agência Nacional de Petróleo (ANP) , isto significa, *1* para áreas licenciadas e *0* para áreas não licenciadas, o outro atributo é relativo as atividades de extração de petróleo no local, *1* se o ponto referido possui atividade de extração, *0* caso não possua. Assim definimos o atributo classe como um **ou** lógico entre os outros dois atributos.

4.1.4 Estimativa de dados

Com os dados reais representados numericamente, é necessário encontrar uma maneira de mapear estes dados de forma a representar uma grande área. Como as unidades numéricas representantes dos dados reais são apenas algumas centenas, achamos que a maneira mais realista seria centralizar os dados fontes na área de sua respectiva bacia.

Com os dados fontes centralizados em cada bacia, processos de interpolação foram realizados com o intuito de estimar novos dados para cobrir a área restante de cada bacia, levando em consideração que a área das bacias é uniforme e portanto, deveria ter mesmo comportamento sedimentar. Com isso, para cada duas bacias vizinhas, os dados da bacia mais ao sul eram interpolados com a bacia mais ao norte. Este processo gerou os dados faltantes para o preenchimento das bacias. Assim, esses dados completam a área norte da bacia ao sul, e o sul da bacia ao norte.

4.1.5 Mineração

Após definido o atributo classe partiu-se para as atividades de mineração. O algoritmo de Classificação J48 foi utilizado para criar um mapa de probabilidade dos possíveis locais com ocorrência de petróleo. Como critérios de classificação foram utilizadas as similaridades litográficas decorrentes das idades geológicas e da profundidade do terreno.

Foi realizado um breve processo de mineração com o algoritmo APRIORI. Consideramos apenas os registros com valor booleano *verdadeiro* no atributo classe (ver:4.1.3). Como resultado dessa

operação foram obtidas algumas informações sobre as áreas em questão. Essas informações foram valiosas a ponto de validação do modelo. A seção a seguir descreve os resultados como um todo.

4.1.6 Interpretação e Validação

Dentre as quatro bacias estudadas nesta primeira etapa, as áreas em que o atributo classe foi *'verdadeiro'* apresentaram Folhelho, que é uma rocha geradora e selo (liga rochas, impedindo que vase o óleo, ver seção 2.2), ou na idade Campaniana ou na Coniaciana, que datam de períodos de 70,6 a 83,5 e 85,8 a 88,6 milhões de anos atrás (respectivamente). Encontrar essa associação foi importante, pois essas idades geológicas estão em um período de tempo que é justamente o período necessário para maturação dos hidrocarbonetos. Esta maturação significa uma potencial transformação dos hidrocarbonetos em petróleo ou gás. Parte do resultado deste etapa de mineração se encontra no apêndice B.

O tempo de maturação dos hidrocarbonetos possui uma variação maior que a mostrada acima, pois depende de vários fatores como pressão, temperatura, etc. Porém a maioria data desse período de tempo.

Podemos associar a descoberta como uma validação do trabalho, embora haja alguns problemas quanto à metodologia, a descoberta mostra que o mapeamento se mostrou relativamente confiável.

O fato de termos realizado o trabalho com cartas antigas nos limita em precisão e acurácia dos dados. A forma de obter dados por interpolação matemática teve a consequência de perdermos a ligação com os dados originais. Estes fatores levaram à criação de uma nova base de dados. Assim, dentre a atualização das fontes (dados geofísicos), foram aproveitadas as técnicas e metodologias bem sucedidas, ao mesmo tempo em que aquelas que apresentaram problemas ou se mostraram insatisfatórias foram remodeladas ou substituídas. A estrutura do banco de dados foi continuada e expandida, de modo que o banco de dados litoestratigráfico pode ser considerado uma nova versão destes dados produzidos e descritos nesta seção.

4.2 Banco de dados Litoestratigráficos

Tendo em vista as oportunidades de melhoria do trabalho anterior, foi iniciado o processo de criação de um novo banco de dados. A metodologia beneficiou-se da experiência adquirida nos testes descritos acima. As metodologias do processo que se mostraram eficazes, como descrito no final da seção anterior, continuaram a ser utilizadas. Bem como, aquelas que apresentaram problemas foram alteradas e/ou substituídas.

Na etapa de **seleção**, seção 4.1.1, foram realizadas as extrações dos dados mais recentes e atualizada a forma de organizar as litologias de modo a facilitar o processo de mineração. Também foi introduzido o valor '1' para marcar um quadrante como parcialmente sem depósitos, ou seja, em uma área de, por exemplo, 10km pode haver informação somente para parte do terreno.

A etapa de **pré-processamento** (seção 4.1.2) sofreu uma significativa mudança. Para esta etapa foi desenvolvida, em Delphi, uma ferramenta para Extração Transformação e Carga (ETL) dos dados coletados. A ferramenta realiza alterações de formatos, importa planilhas Excel, realiza cálculos, aplica os algoritmos criados para estimativas e carrega os dados para o banco de dados. A figura 4.3 mostra a interface da ferramenta, juntamente com um breve resumo de suas funções.

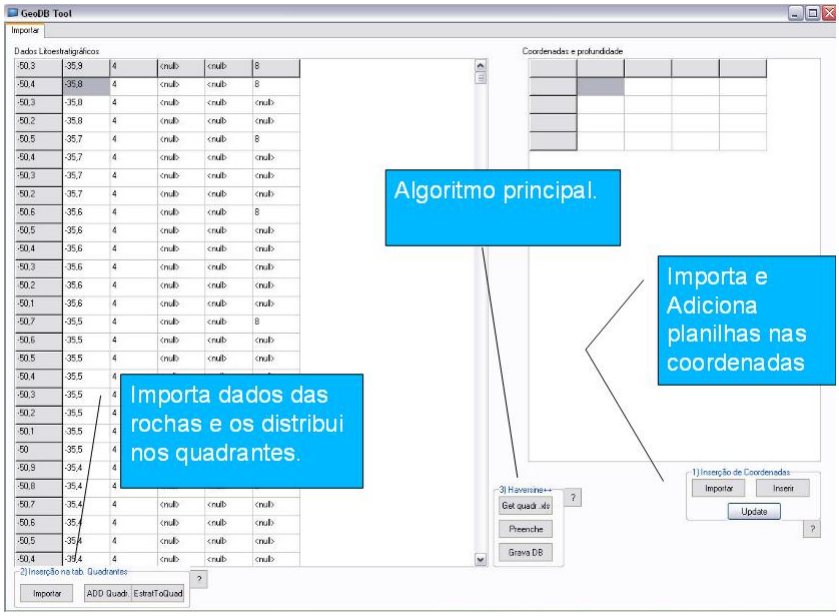


Figura 4.3: Ferramenta de ETL para o banco de dados litoestratigráfico

A principal função da ferramenta se encontra na parte de carga dos dados, pois a mesma comporta os algoritmos criados para mapear os dados nas coordenadas estimadas. A função de mapeamento é basicamente dividida em três partes: o *upload* da planilha pré formatada, o preenchimento dos dados na costa e a gravação dos dados estimados no banco de dados.

O mapeamento dos dados seguiu a mesma metodologia da base anterior. Contudo, a técnica de estimativa e mapeamento dos dados foi refeita. Após os dados originais serem mapeados nas bissetrizes, foi utilizada a fórmula de Haversine [Gel89] para calcular a distância entre pontos. Assim, como parâmetros para o ponto médio foram usadas a distância entre o ponto em questão e os pontos mais próximos em ambas as bissetrizes.

A formula de Haversine (4.1) foi escolhida para se obter uma distância mais precisa, já que trabalhamos em um plano esférico. Essa fórmula leva em consideração a curvatura de uma esfera (nesse caso a Terra) e calcula a distância entre dois pontos.

$$
\begin{aligned}
R &= earth'sradius(meanradius = 6.371km) \\
\triangle lat &= lat2 - lat1 \\
\triangle long &= long2 - long1 \\
a &= sin^2(\triangle lat/2) + cos(lat1).cos(lat2).sin^2(\triangle long/2) \\
c &= 2.atan2(\sqrt{a}, \sqrt{(1-a)}) \\
d &= R.c
\end{aligned}
\tag{4.1}
$$

Para estimar o valor de um ponto x no espaço foi criado um algoritmo que se baseia na distância entre o ponto em questão e os pontos mais próximos de cada bissetriz. Assim, o algoritmo cria uma matriz de valores que são mapeados em pontos no espaço. Então, para cada idade geológica, faz-se os seguintes passos:

1. Encontra-se o valor do ponto na primeira bissetriz;
2. Varre-se a matriz até encontrar o valor correspondente na segunda bissetriz;
3. Aplica-se a fórmula de Haversine para obter a distância total;
4. Varre-se a matriz, checando os índices correspondentes;

Para cada índice correspondente, dentro de uma idade geológica, segue-se os seguintes passos:

1. Verifica-se o valor para checar se já foi preenchido; caso sim, segue para o próximo índice; caso não continua-se com os próximos passos;
2. Aplica-se Haversine entre o ponto a ser preenchido e os pontos ao extremo das bissetrizes para obter a distância entre os mesmos;
3. Aplica-se regra de três entre as distâncias para descobrir a porcentagem de distância relativa as bissetrizes;
4. Aplica-se uma fórmula de randomização dando chances inversamente proporcionais a distância do ponto com a bissetriz.
5. Atribui-se os valores presentes no ponto da bissetriz selecionada.

Após todos os valores litoestratigráficos serem obtidos e devidamente mapeados, foram realizados alguns testes com o auxílio de dados pré-conhecidos e ferramentas de modelagem espacial. Ao final do processo, dados foram extraídos diretamente da base para verificar o modelo de batimetria. A figura 4.4, criada via SURFER [Gol06], mostra um modelo 3D criado com os dados (correspondentes a idade geológica mais atual) extraídos diretamente do banco de dados.

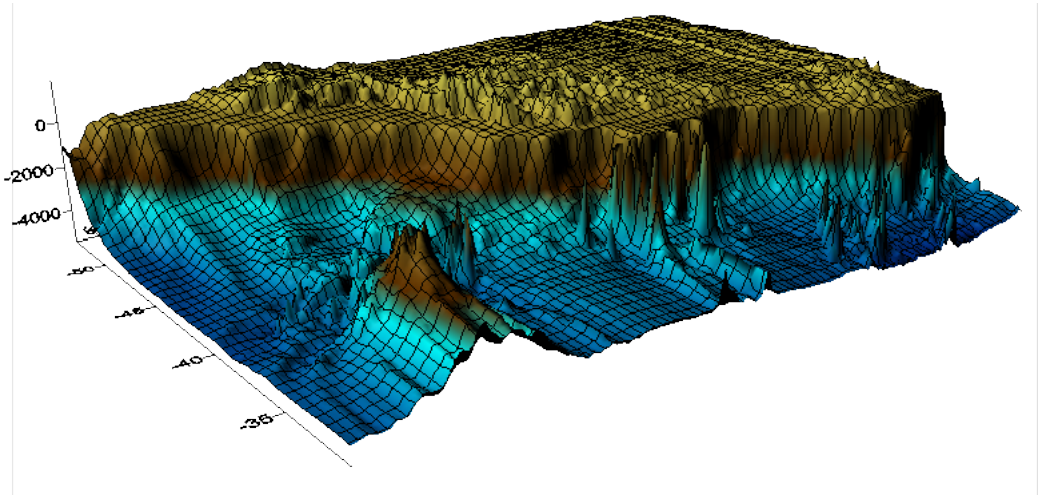


Figura 4.4: Modelo extraído do banco de dados

4.3 PaleoGeoDB

Paleo Geographic Database é o banco de dados criado para comportar o modelo numérico final. Este modelo numérico, constitui todos os dados coletados, transformados e integrados.

Após a conclusão da Base de dados litoestratigráfica, foi expandido o modelo do banco de dados de modo a comportar outros dados geofísicos. O banco de dados batizado como PaleoGeoDB, possui uma estrutura aprimorada em relação ao banco de dados anterior. A estrutura do mapeamento das cartas foi alterada para haver mais coesão com a realidade.

No banco de dados litoestratigrafico (etapa 2, ver Seção 4.2) os dados eram mapeados nas bissetrizes e os demais dados eram estimados através de algoritmos. Porém, foi constatado que mapear os dados de uma carta para uma bacia, sem realizar estimativas de dados, torna o mapeamento mais real. As estimativas dos dados não são propícias, devido ao fato de que os dados presentes nas cartas estratigráficas já possuem uma relação de ordem e distância.

Outra técnica descartada foi a interpolação de valores entre as bacias, pois geologicamente, não faz sentido interpolar valores entre bacias, uma vez que estas possuem divisões naturais em que os sedimentos foram depositados durante as eras. Estas divisões existem justamente por haver características distintas entre as bacias.

Outra mudança, em relação ao banco de dados da etapa 2, foi a relação dos dados com a área relativa aos mesmos. Esta mudança foi adotada devido ao fato de que as cartas correspondem as formações geológicas das bacias, e não as 200 milhas náuticas como fora feito no trabalho prévio.

Devido a estes fatores, alteramos a nossa abordagem em relação ao mapeamento das litologias. Nesta nova abordagem cada carta corresponde a uma única bacia, assim como a distância entre os pontos se tornou variável devido ao tamanho das bacias em relação à costa.

4.3.1 Modelo para o Banco de Dados

Dentre as questões que cercam a criação de um modelo, talvez as mais evidentes sejam: "Quais dados possivelmente estarão presentes?"e "Como garantir que eles possam ser adicionados de forma harmoniosa com os demais dados?". Devido a estas questões foram levantados os dados que possivelmente seriam assimilados no banco de dados. Com base no modelo anteriormente criado, foram realizadas as melhorias e atualizações para criar o modelo atual.

Primeiramente foram selecionados os dados relevantes para o banco de dados. Além dos dados já utilizados no banco anterior, foram adicionados a arquitetura do banco suporte a dados como: gravimetria, Gás Carbônico, isotopos de Oxigênio, formação específica em uma bacias etc. A figura 4.5 ilustra o modelo proposto para o banco de dados.

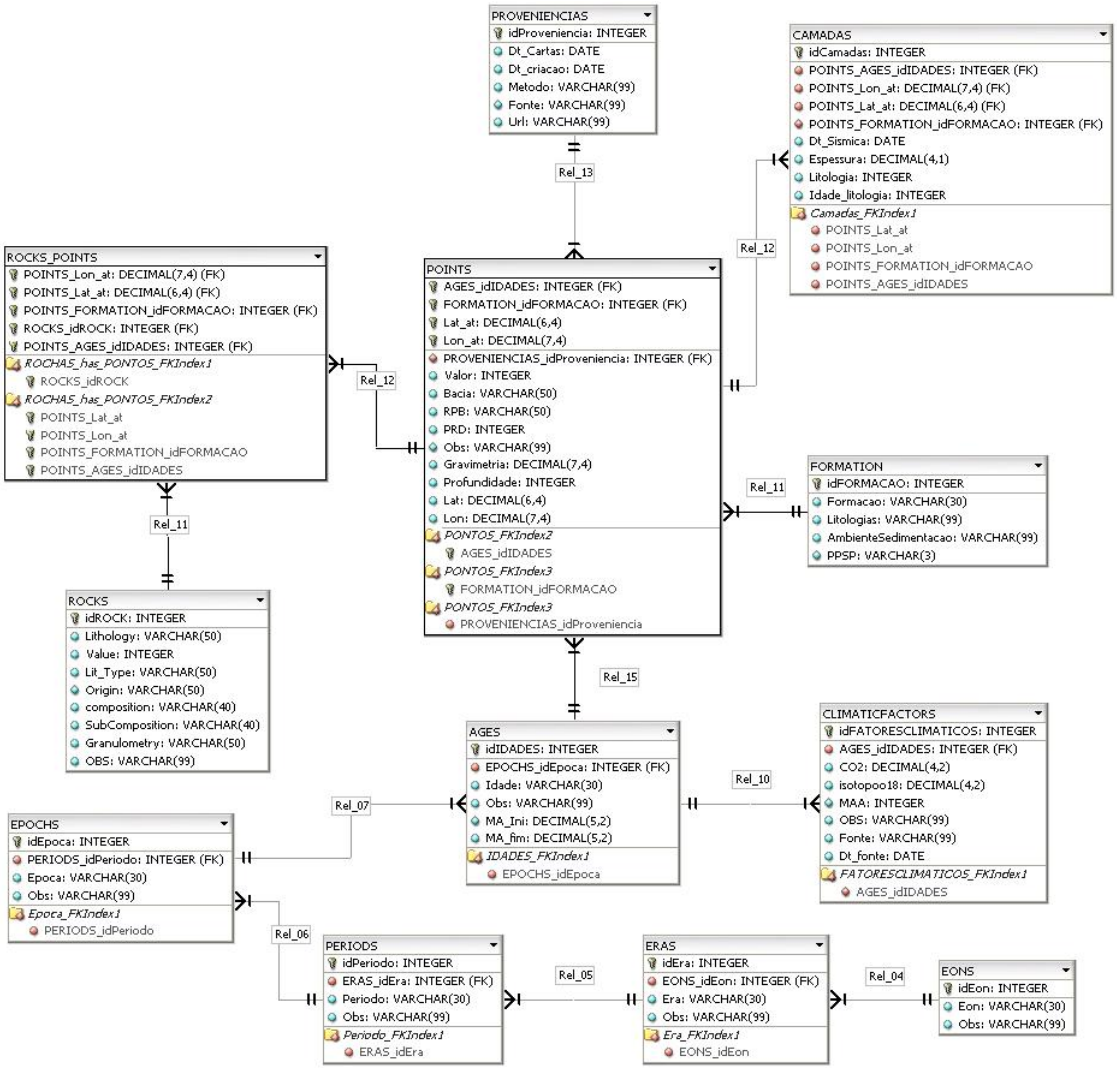


Figura 4.5: Modelo proposto para o Banco de Dados PaleoGeoDB.

Como mostrado na figura 4.5 o modelo permite armazenar dados paleogeográficos e paleoclimáticos. A seguir, são descritas, detalhadamente, as tabelas e os campos do banco de dados.

No banco de dados temos cinco tabelas estáticas que descrevem o tempo geológico. São elas: **Idades**, **Epocas**, **Periodos**, **Eras**, **Eons**. Basicamente estas tabelas guardam os nomes dos tempos geológicos, juntamente com uma campo para observação. No caso da tabela **Idades**, que é a divisão mínima do tempo geológico aqui utilizado, são utilizados dois campos para marcar em quantos milhões de anos atrás a idade teve início e fim (MA_ini, MA_fim).

A tabela **FatoresClimaticos** armazena dados climáticos em geral. Esses dados geralmente são obtidos em janelas de tempo menores que as das idades geológicas. Inicialmente, foram criadas tabelas para dados de Gás Carbônico (CO_2) e Isótopos de Oxigênio. Ambos possuem um registro para cada milhão de ano.

A tabela **Pontos** é a principal tabela no banco de dados, pois esta representa as informações da bacia no espaço e tempo. O campo *RPB* (Referente a Parte de Bacia) serve para armazenar a sub parte da bacia que o ponto pertence e pode ser completado com as seguintes opções: Linha de Costa, Plataforma, Talude e Sopé. O campo *PRD* serve para armazenar a Porcentagem Relativa de Depósitos em uma área.

Essa tabela utiliza coordenadas geográficas para marcar um local. Ao final duas 'malhas' são mostradas. Uma com informações de batimetria e gravimetria, dispostas em linhas e colunas com precisão de 1 minuto de grau. Outra com dados dos sedimentos, disposta sinuosamente no globo e com distância irregular entre os pontos.

A malha com os dados atuais de gravimetria e batimetria é disposta apenas para o tempo presente, já que esses dados não existem para outras idades geológicas. Porém, a malha com os dados dos sedimentos é expandida para as idades geológicas anteriores. As coordenadas atuais (Campos: lat_at, lon_at) são mantidas fins de referência, contudo, cada ponto em distintas idades geológicas possuem distintas coordenadas.

As coordenadas em idades geológicas passadas são *flags* que marcam onde cada ponto esteve no passado. Assim estas coordenadas constituem a reconstituição da deriva continental. Esta reconstituição é estimada de maneira abstrata e deve sofrer ajustes para se adaptar pelo modelo de deriva continental proposto por Moulin et al. [Mou10].

Como um ponto representa uma determinada área na bacia, tem-se mais de uma litologia para o mesmo ponto. Porém também há partes da área que não possuem depósitos. A porcentagem da área que não possui depósitos deve ser subtraída, assim deve ser armazenada a porcentagem da área total que sofreu depósitos, para isso temos o campo *PRD*.

Para controle de validade e qualidade dos dados temos a tabela **proveniencias**. Esta tabela serve para armazenar a data de criação dos dados, a data das fontes provenientes das cartas estratigráficas, a descrição de outras fontes, o método utilizado e o endereço da página utilizada como fonte (Campo: URL).

Como uma bacia é constituída de várias formações, foi adicionada a tabela **Formacao**. Esta tabela armazena, além do nome da formação, o ambiente de sedimentação e o Potencial Papel no Sistema Petrolífero .

A tabela **camadas** serve para armazenar camadas provenientes de perfis de sísmica marítima. Nela é possível atribuir dados mais precisos que servem para validação e aprimoramento dos dados no banco de dados. Para cada camada é possível definir dados como: litologia que constitui a camada, idade e espessura da camada. A figura 4.6 mostra um perfil de Sísmica Marítima (interpretado) e suas camadas para a bacia de Santos.

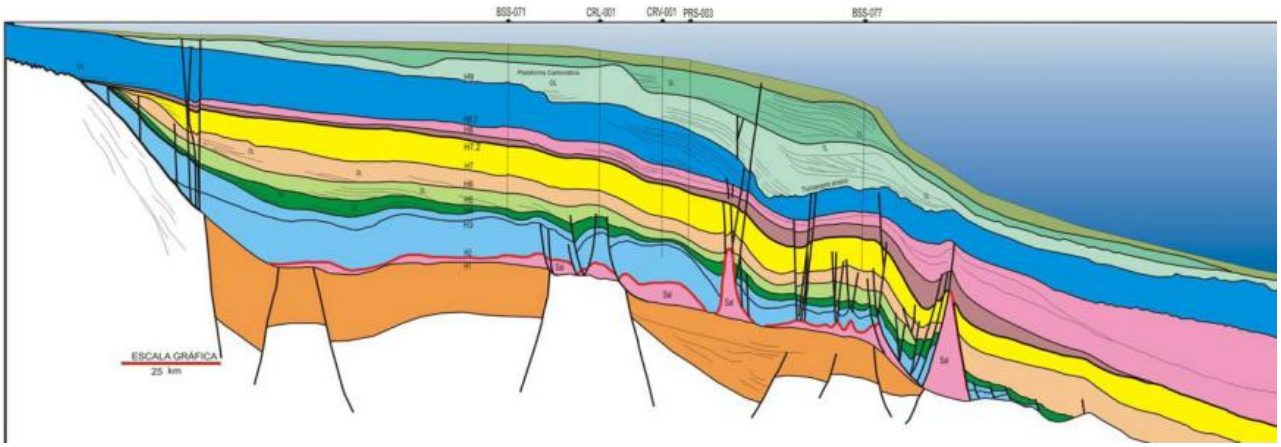


Figura 4.6: Perfil de Sísmica referente a Bacia de Santos.

A tabela **rochas** armazena basicamente dados das litologias, tipo da litologia e outras divisões. Estas divisões além de armazenar a informação tem como objetivo auxiliar no processo de KDD, pois isso torna possível executar algoritmos baseando-se em grupos de litologias.

Por fim, a tabela **rocha_pontos** serve para armazenar valores correspondentes a uma única litologia, já que a tabela *pontos* armazena os valores somados.

Em resumo, o modelo do banco permite armazenar:

- Informações sobre as litologias presentes nas bacias sedimentares;
- As litologias locais, variantes conforme o tempo geológico;
- O tempo geológico, separado em Idades, Épocas, Períodos, Eras e Eons;
- Informações explicitas sobre qual parte da bacia um ponto de dados (os dados são mapeados em pontos no mapa) pertence;
- Estimativas referente à quantidade de depósitos litológicos presentes no local e no tempo geológico;
- Dados de batimetria e gravimetria para cada ponto no tempo e espaço;
- Dados de formação de uma bacia;
- Dados de sísmica marítima, separados por camadas;

- Metadados para a proveniência dos dados.

Além do modelo completo do banco, existe um modelo estrela (Pronto para DW), com o objetivo de facilitar a obtenção de informação em tempo real. Foram criados scripts de inserção de dados, para realizar a carga neste modelo, de modo que os dados presentes no PaleoGeoDB estejam presentes no modelo estrela do banco. A figura 4.7 ilustra o formato do modelo estrela.

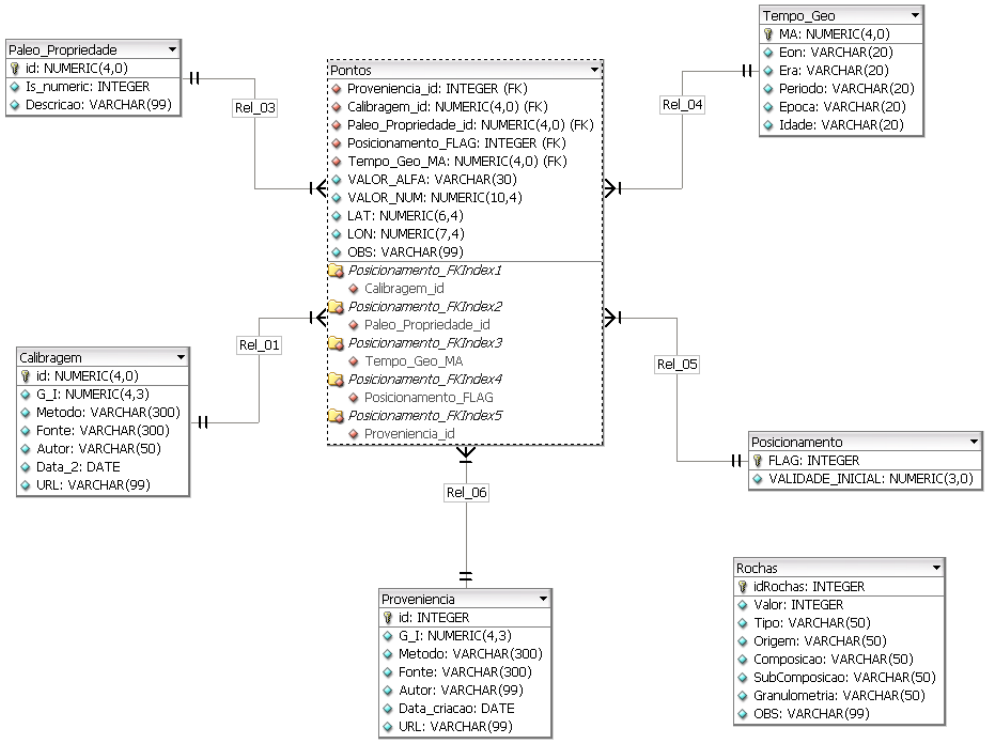


Figura 4.7: Banco de Dados modelo Estrela.

4.3.2 Obtenção e tratamento das coordenadas

Quanto ao limite territorial, a grande questão é: "Como manter os dados dentro dos limites sinuosos das bacias, de maneira a representar os dados reais de forma fiel?". A estratégia adotada foi primeiro obter as linhas que limitam o território. Assim, basicamente temos que obter as coordenadas que marcam a linha da costa e a linha do final da bacia.

Para obter as coordenadas foram utilizadas 2 fontes. Do *National Oceanic and Atmospheric Administration* (NOAA) [Sol90] foi obtida a linha da costa. O NOAA foi escolhido por três razões; a precisão dos dados e o fato de termos mais de uma fonte (assim é possível comparar os dados automaticamente, realizando mais uma etapa de validação).

Para obter as demais coordenadas, foi utilizado o TOPEX, que gera coordenadas com precisão de quatro casas decimais, juntamente com a altitude/profundidade do local. A terceira razão pela qual o NOAA foi utilizado é devido ao fato de que o TOPEX não gera linhas de contorno continental, assim

para gerarmos uma linha de costa (LC) seria necessário verificações de coordenadas com outras fontes.

Foi obtido um total de 4.142.875 pares de coordenadas, considerando as coordenadas relativas à terrenos acima do nível do mar. Para validar as linhas foram utilizadas duas ferramentas. Primeiramente foi montado um Mapa 3D com o auxílio do software Surfer [Gol06]. Depois criado um arquivo .KML para o Google Earth.

No Google Earth os dados de costa se mostraram deslocados algumas dezenas de metros em direção Oeste. Porém todos uniformes, ou seja, mostravam o contorno idêntico ao do Google Earth, apenas com uma certa diferença em relação à longitude.

Para resolver o problema os dados foram mesclados com os do TOPEX. Foram filtrados, dentre aqueles dados que possuem 0 (zero) no valor de profundidade, os que mais se aproximam da linha da costa. Assim foram obtidos os dados do TOPEX com base nos dados do NOAA.

4.3.3 As três linhas divisórias

Diretamente do NOAA foi obtida a linha da costa. Para obter a linha limite, foram utilizados os dados do TOPEX que já estavam no banco de dados. Para isso foi usado um filtro simples via SQL. Aquelas regiões que estão próximas a -3.000 metros formam a linha (L3K). Este filtro foi aplicado até obter uma boa relação entre precisão e quantidade de dados que formam a linha. No final o filtro ficou com uma variação em torno de 15 metros (-2.985 a -3.015).

Para aproveitar melhor as informações presentes nas cartas estratigráficas uma terceira linha foi estabelecida. Como as cartas tem um indicador de quebra de plataforma (QP), faz sentido que os dados sejam distribuídos de acordo com o local.

O exemplo a seguir ajuda-nos a visualizar melhor a situação. Digamos que a bacia x em uma latitude y possua 300 metros de distância da costa. Imagine que a quebra de plataforma inicie a 50 metros da costa. Com apenas duas linhas teríamos um dado de litologia a cada 8,1 km, isto daria apenas 6 dados antes da quebra de plataforma independente de quantos dados houvessem na carta. Se naturalmente houvessem 9 dados na carta antes dos 50m, então teríamos 3 dados fora do local. Ou seja, 3 dados estariam no Talude ao invés de estarem na Plataforma continental.

Não foram encontradas fontes que possuíssem coordenadas ou mesmo informações que indicassem pontos de quebra de plataforma. Assim, os pontos foram estabelecidos e exportados via Surfer [Gol06] com base no modelo 3D criado com os dados do banco.

Após a criação da linha de QP, obtivemos três linhas no banco de dados, LC, QP e L3K. Para a execução do algoritmo de preenchimento (Seção 4.3.4) pequenos espaços foram preservados dentre as bacias, visando prevenir a sobreposição de dados. Além disso, do ponto de vista geológico, é melhor que haja pequenos espaços em branco (o que representa a divisão das bacias) do que presença de dados de outra bacia.

O software desenvolvido exporta as coordenadas geradas. Assim, com as mesmas foi criado um

arquivo .kml que é utilizado pelo Google Earth para visualizar pontos, linhas e outras referências. A figura 4.8 mostra a pré-visualização das linhas.

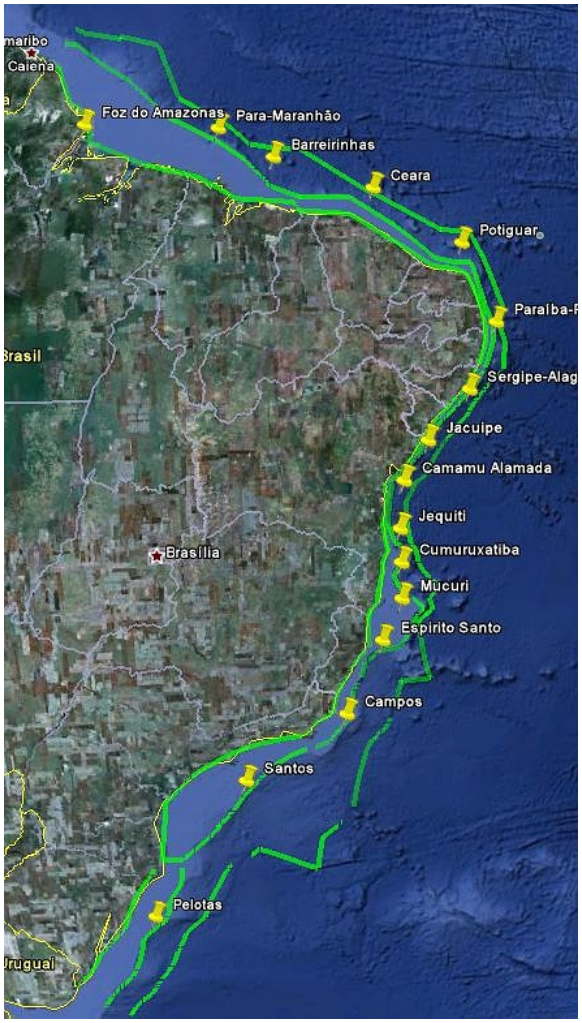


Figura 4.8: Linhas: LC, QP e L3K geradas.

4.3.4 Algoritmos de preenchimento

Devido à enorme quantidade de dados, e a enorme área para atribuir esses dados, a necessidade de um bom sistema para automatização do processo é fundamental. Possivelmente devido ao fato da originalidade e especificidade do trabalho aqui descrito, não foram encontradas técnicas na literatura ou trabalhos correlatos que fossem úteis para a distribuição e mapeamento automático dos dados.

Em uma visão alto nível, foram elaboradas 3 soluções. Uma utilizando distorção de imagens, outra utilizando ângulos em relação à costa, e outra com base em fórmulas e distribuição dos dados.

Solução A: Distorção de imagens.

A técnica de fusão de imagens consiste, em uma interpolação de pixels entre duas figuras de modo a criar uma única figura preservando suas principais características [Hon10]. Em outras palavras, esta

técnica visa fundir duas ou mais imagens para sintetizar a informação significante de cada imagem em uma única imagem. Esta técnica é amplamente utilizada para sensoriamento remoto, geração de imagens médicas e aplicações militares [Wan04].

Como não faz sentido (do ponto de vista geológico) unir dados de duas bacias, a técnica seria adaptada para uma distorção da imagem, e aplicada em subdivisões dos gráficos das cartas estratigráficas. Cada carta deveria ser fatiada em 34 partes, que correspondem às 34 idades geológicas pertinentes a este trabalho, deste modo cada fatia corresponderia a uma bacia em uma idade geológica. Estas fatias seriam distorcidas, através de algoritmos presentes na literatura, de modo a completar a área de uma bacia. Com uma imagem cobrindo uma bacia, a atividade a seguir seria transcrever as legendas presentes nas cartas para os respectivos valores.

A vantagem desta técnica é que o mapeamento dentro dos limites teria uma excelente precisão, já que a imagem da carta (fonte original) seria ajustada de modo gráfico nos limites da bacia. Porém essa técnica possui alguns inconvenientes. Como definir a distância entre um dado e outro de forma a não perder informações é uma das questões, já que a transcrição gráfica automatizada não é flexível como a análise humana quanto às litologias dos fontes. Outro grande problema é a transcrição para os valores em si, já que as litologias presentes nas cartas são compostas de cores e traços.

A figura 4.9 representa o recorte da bacia de Jequitinhonha relativo à idade Ypresiana. Verticalmente esta imagem seria deformada de um limite a outro da bacia e horizontalmente da costa até a profundidade de 3.000 metros no Atlântico.

Figura 4.9: Recorte da carta de Jequitinhonha, relativo a idade Ypresiana.

Solução B: Linhas de distribuição com base em ângulos de 90º partindo da costa. Partindo da ideia de que os dados devem ficar perpendiculares à linha de costa relativa à bacia, observou-se que a criação de um limite angular centrado nos 90º poderia ser uma boa opção. Assim, para cada linha de dados a ser mapeada um cálculo deve ser realizado para determinar a direção da linha.

O algoritmo consiste em formar uma matriz de pontos onde: O valor na Matriz [i,j] são pontos para determinar a angulação. i é um ponto na linha da costa (i_0 para o início da bacia de Pelotas, i_n para o limite Oeste da bacia de Foz do Amazonas) e j o registro limite previamente escolhido indo em direção ao Oceano.

O valor de *j* é uma coordenada utilizada para calcular a melhor variação do ângulo (em relação aos 90º) de modo a ajustar as linhas e impedir espaços em branco ou sobreposições. A figura 4.10 mostra linhas em forma de 'T', onde a linha que segue em direção ao Oceano é a linha de preenchimento dos dados. A linha verde mostra um caso de sucesso, ao contrário das linhas vermelha e preta, onde a linha vermelha cruza com a preta havendo sobreposição de dados.

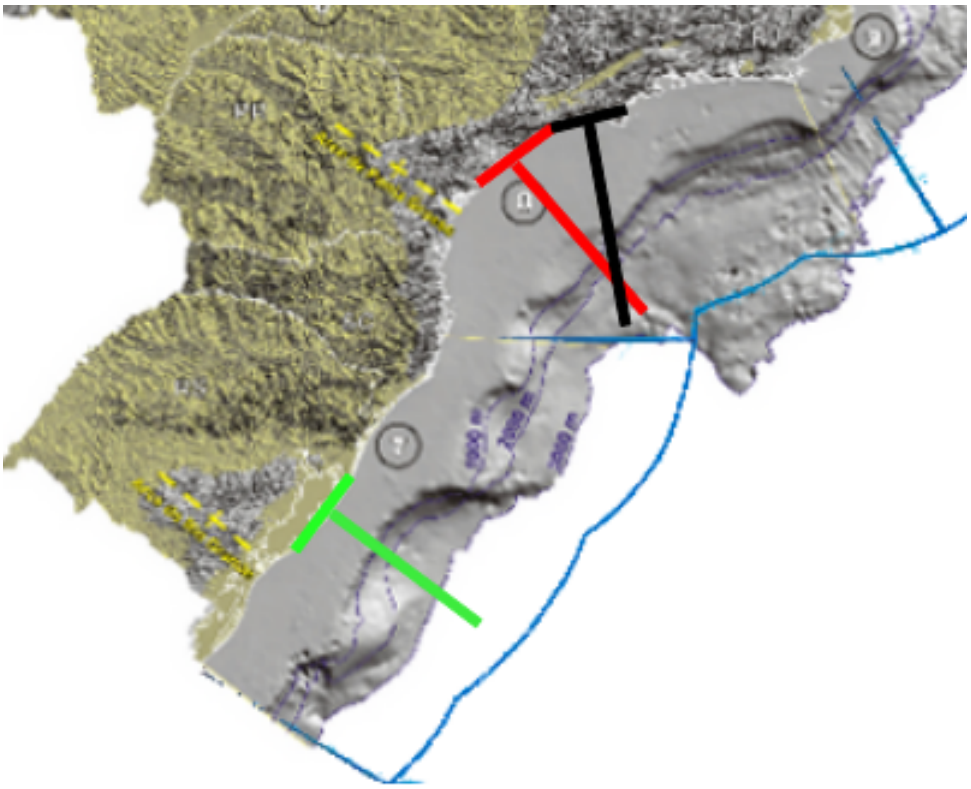


Figura 4.10: Exemplo de problema com a utilização de ângulos como parâmetro.

Como a quantidade de dados é imensa e a costa brasileira é muito sinuosa, controlar a angulação das linhas de preenchimento automaticamente de forma a evitar espaços em branco e colisões se mostra uma tarefa complexa. Por esse motivo esta técnica foi descartada.

Solução C: Fórmulas para distribuição de dados. A ideia desta solução é que, a partir de cada ponto o próximo ponto seja localizado via um conjunto de fórmulas. Para construção deste algoritmo duas fórmulas são essenciais: Haversine (ver Fórmula 4.1) para encontrar a distância entre dois pontos e uma fórmula para se obter a curvatura entre dois pontos.

Para se obter a curvatura, simbolizada por Θ (Teta), entre duas coordenadas geográficas utiliza-se a seguinte fórmula:

$$\Theta = atan2(sin(\Delta long).cos(lat2), \\ cos(lat1).sin(lat2) - sin(lat1).cos(lat2).cos(\Delta long)) \tag{4.2}$$

Estas fórmulas são utilizadas em meio ao algoritmo, de modo que para cada ponto são chamadas funções que utilizam ambas as fórmulas. A seguir o algoritmo é descrito, em duas etapas e em alto nível.

Para cada carta primeiramente divide-se a mesma na quebra de talude, então para cada uma das partes:

1. Obtém-se a distância entre os pontos nas extremidades (Distância Total);
2. Verifica-se a quantidade de dados (das cartas) neste intervalo;
3. Divide-se o resultado do passo 1 pelo resultado do passo 2 para obter o tamanho dos segmentos;
4. Havendo o ponto inicial, o tamanho dos segmentos e a curvatura é possível obter a coordenada do próximo ponto.

A ordem de etapas descritas acima faz parte de um laço lógico que se repete n vezes para o preenchimento de uma bacia, mais 34 vezes para preenchê-la nas idades geológicas. O resultado dos passos lógicos, descritos acima, é o preenchimento completo de uma bacia. É importante lembrar que os dados de idades mais antigas não são atribuídos diretamente as coordenadas, estes apenas são dispostos na extensão da bacia para posteriormente serem mapeados com *flags* de posição e deslocamento.

O resultado dos passos lógicos, descritos acima, é o preenchimento completo de uma bacia. Porém algumas variáveis de entrada são necessárias para iniciar o algoritmo. Essas variáveis são os pontos iniciais (pontos na LC) e finais (pontos da QP) que determinam a distância total (passo 1). Duas questões são importantes para a definição dessas variáveis: "Quais os pontos iniciais e finais a serem utilizados? E como obtê-los?". Para um correto mapeamento as linhas ao serem preenchidas devem ficar com um ângulo inicial de 90 graus em relação a bacia. Os itens listados a seguir mostram os passos lógicos criados para definir os pontos a serem usados nas retas perpendiculares a costa.

1. Obtém-se a distância entre os pontos iniciais de cada bacia (Distância Total);
2. Verifica-se a quantidade mínima de pontos encontrados nesse intervalo, por LC, QP e L3K;
3. Verifica-se o tamanho do segmento entre esses pontos;
4. Para as outras duas linhas faltantes usa-se a mesma quantidade de pontos com base no segmento.

Após obter a quantidade de pontos, com a mesma distância de seguimento, traça-se as linhas de LC, QP, e L3K. Cada uma dessas linhas é armazenada em um vetor, então, entre dois vetores traça-se as retas como descrito no primeiro algoritmo.

Como dados de entrada para o algoritmo, são necessárias as coordenadas limites de cada bacia. Porém, devido à sinuosidade da costa, e a extensão do terreno das bacias, são necessárias coordenadas extras para ajustar a curvatura e a direção das linhas. Para isso foram criadas linhas, de modo visual, no Google Earth; suas coordenadas foram capturadas e exportadas. Essas coordenadas utilizadas podem ser conferidas no apêndice A.

Esta solução se mostrou eficaz, tanto em termos de precisão como em performance. A figura 4.11 mostra um exemplar do resultado obtido pelo programa desenvolvido que utiliza-se deste algoritmo para mapear e preencher os dados no banco de dados. O resultado de todas as bacias pode ser conferido no apêndice F.

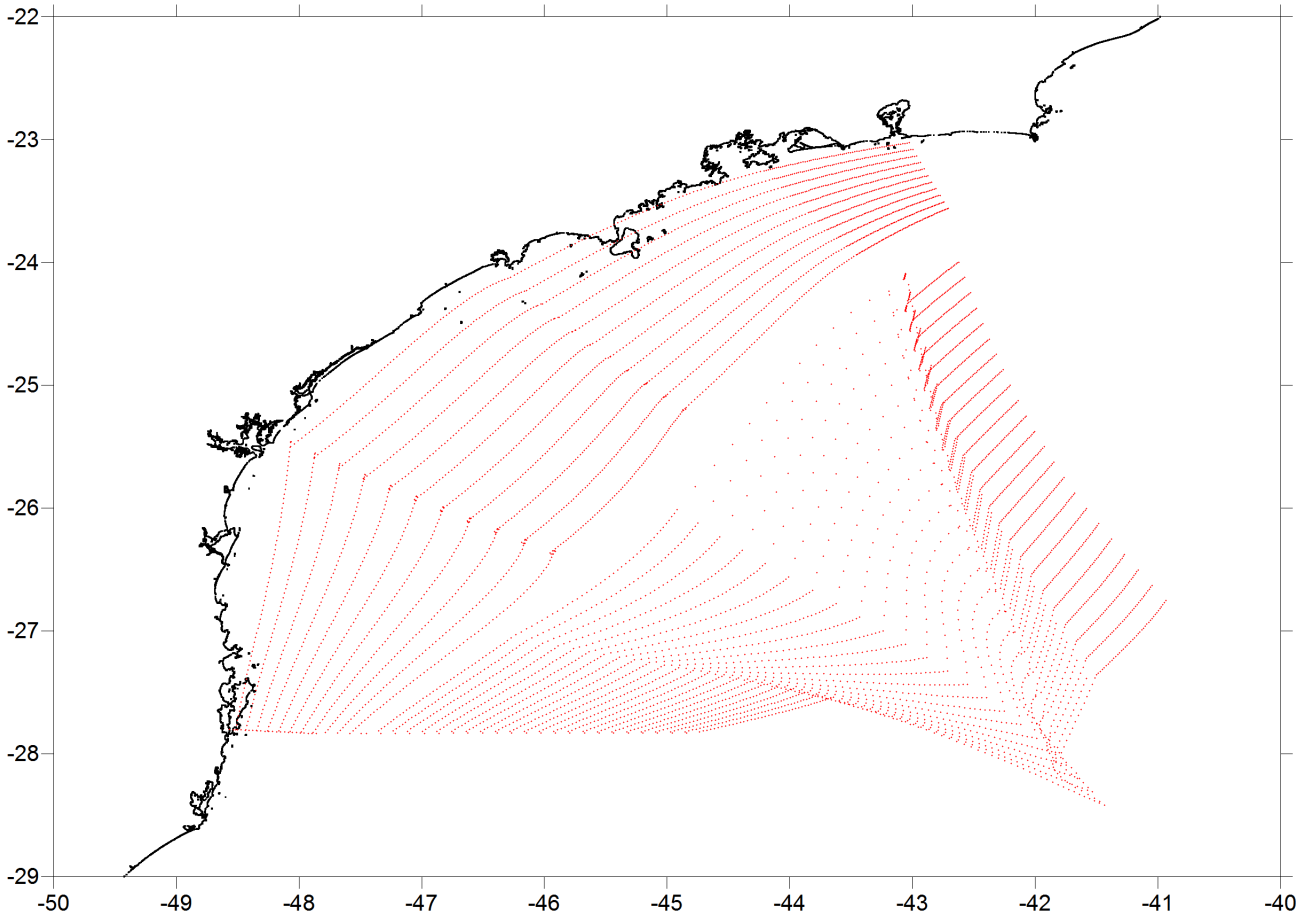


Figura 4.11: Visão gráfica da localização dos dados na Bacia de Santos.

Este algoritmo pode ser utilizado para preenchimento de quaisquer dados numéricos ao longo de qualquer região em uma superfície esférica. Para isso, devem apenas ser estabelecidos as coordenadas de limite e controle (coordenadas internas a uma determinada área que ajudam na precisão das linhas).

Para implementar os algoritmos e carregar os dados no banco a ferramenta de ETL, descrita na seção 4.2, foi atualizada. A ferramenta mais do que realizar a ETL e executar os algoritmos, demonstra bom funcionamento dos algoritmos criados, validando a teoria desenvolvida. Seu desempenho prova-se satisfatório ao realizar milhões de cálculos na inserção dos registros (mais detalhes da ferramenta, podem ser conferidas no apêndice E).

Problemas encontrados.

Foram encontrados alguns problemas com a solução. Estes problemas referem-se a precisão no mapeamento dos dados e em geral não chegam a ser significantes já que as áreas trabalhadas são bem

extensas e a falha na precisão das fórmulas é pequena.

O primeiro fator que leva à imperfeição da solução se deve ao fato do algoritmo estar construído com base em duas fórmulas que não são 100% precisas. A figura 4.11 é uma amostra da primeira geração de dados produzidos. Podemos notar que ela possui alguns defeitos de sobreposição dos dados.

A sobreposição, ocorre por três fatores. O primeiro refere-se à precisão da formula de Haversine. Isso implica que quanto maior a precisão dos dados, maior é o grau de falha da fórmula. Este fato foi constatado com análise entre as bacias que possuíam muitos dados para áreas muito pequenas (e consequentemente exigiam mais precisão). Estas tiveram uma sobreposição mais acentuada dos que as demais bacias.

O segundo fator é a fórmula de curvatura que pode vir a tomar uma direção oposta a esperada, e assim fazer com que dados sejam mapeados muito próximos à outros (em termos de visualização, sobrepostos). Caso sobrem dados ao final de um segmento, a curvatura tende a fazer com que o próximo dado após o final do segmento (primeiro que sobrou), entre as retas que cortam as bacias perpendicularmente, seja mapeado antes do último dado. A figura 4.12 exemplifica.

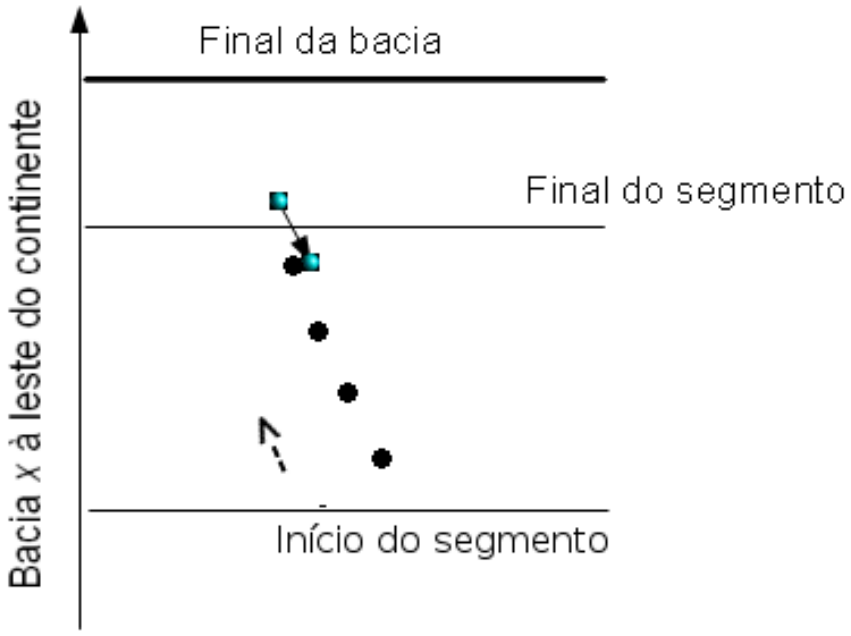


Figura 4.12: Problema com a quantidade excessiva de dados e a fórmula de curvatura.

Como os dados são mapeados em linhas ao longo da costa, onde cada linha representa o mesmo grupo de litologias, este fator não chega a ser um problema. Em outras palavras, os dados sobrepostos provenientes de uma mesma linha paralela à costa são os mesmos.

O terceiro fator é o ajuste da quantidade de dados e seus segmentos. Este fator pode ser corrigido alterando alguns valores em variáveis dentro do algoritmo. Eliminando este fator de erro, diminui-se

os problemas com o segundo fator, pois quando o segmento dos dados é muito grande, sobram dados no final. Assim, o cálculo da distância, junto a curvatura, ajuda a criar anomalias no mapeamento.

4.3.5 Plano de mineração

Mais que um plano de mineração, esta seção se dispõe a criar um plano de KDD e GKD, onde a mineração de dados deve ser executada de diversas maneiras e com diversas configurações para os algoritmos. Este plano se propõe não apenas para descoberta de conhecimento, mas também para validar e melhorar a precisão no mapeamento dos dados.

Para executar este plano é importante seguir sua ordem, pois em alguns casos um processo depende dos resultados de outros. A generalização espaço-temporal por exemplo, define uma regra geral baseada em uma hierarquia dos dados e um conjunto de padrões. Para isso é necessário que regras de Associação e Associação espaço-temporal sejam previamente conhecidas.

O plano aqui descrito, trata sobre técnicas de mineração, citando itens como objetivos dentro de um tópico. Algumas técnicas em GKD são possivelmente mais difíceis de se executar que outras. Primeiramente pela própria natureza do GKD e a mineração espacial. Segundo, porque as possibilidades de descoberta em meio aos dados espaço-temporais são tantas, que possivelmente sejam necessários novos algoritmos para se obter toda informação previamente desconhecida.

Associação

Visa encontrar similaridades entre as rochas e fatores climáticos. Dentre algumas possíveis similaridades podemos citar:

- Similaridade entre as rochas encontradas nos locais onde há extração atualmente (para validação da base).
- Similaridades entre rochas geradoras presentes em uma formação e rochas de outras formações com potencial previamente desconhecido.
- Correlações entre gravimetria e demais dados geológicos.

Classificação

O plano de mineração para Classificação, Visa criar modelos de indução com base nos geodados, possivelmente um mapa de ajuda a tomada de decisão para perfuração de poços (locais com petróleo). Para isso devem ser usados modelos de classificação preditiva onde os atributos sugeridos são: Lat, Lon, Batimetria, Gravimetria, Isótopos de Oxigênio, e CO_2.

Para atributo classe sugere-se, com base na(s) rocha(s) geradora(s) da bacia: marcar '1' caso haja presença de rocha com potencial para ser geradora (ver seção 2.2) e 0 caso não haja a rocha na coordenada.

Regras de evolução

Com base em mineração espacial é possível determinar padrões de evolução dos geodados. Dentre outros possíveis resultados, provenientes de um processo de GKD, pode-se citar:

- Criação de um gráfico que mostre o caminho percorrido pelos sedimentos. Isso pode ser importante para detectar domos de sal, que naturalmente tendem a se achatar horizontalmente, alterando a área ocupada pelos sedimentos.

- Possível criação de uma malha que ligue valores de sedimentos de modo a detectar anomalias e, por consequência, detectar falhas ou domos de sal.

A figura 4.13 mostra uma possível malha entre os sedimentos representados por 'x' e 'Losangos'. Ao criar uma evolução desta malha as anomalias ficarão visíveis, o que facilitaria o processo para descoberta de domos de sal que podem esconder grandes reservas de óleo no pré-sal.

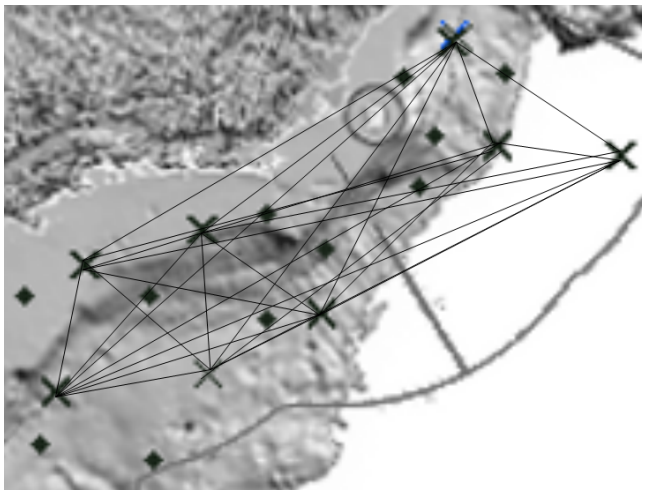


Figura 4.13: Exemplo de uma possível malha (entre duas litologias) criada a partir de mineração com regras de evolução.

Associação espaço temporal

Similar a associação convencional, porém com o foco em encontrar similaridades entre as alterações dos dados no espaço-tempo. Dentre algumas possíveis similaridades podemos citar:

- Similaridade entre a evolução das rochas geradoras de hidrocarbonetos.

- Padrões temporais entre CO_2, Isotopo de Oxigênio, gravimetria e as rochas com papel importante nos locais produtores de petróleo.

Generalização espaço-temporal

Com esta etapa da descoberta de conhecimento, podemos agregar algumas informações previamente conhecidas a fim de induzir regras gerais em torno de algo. Possivelmente um conjunto de fatores que levem a uma indicação do tipo: Uma bacia que possuiu sedimentos x em uma idade y possui reservas em uma área ocupada pelo sedimento x. Esta etapa é importante para difundir o conhecimento gerado, porém demanda conhecimentos com alto grau de confiança como entrada.

Segmentação de dados espaço-temporais

Podemos segmentar os dados com base em seu comportamento espaço-temporal. Essa etapa requer como entrada dados de associação, regras de evolução e associação espaço-temporal. A ideia é separar os dados geológicos em grupos. De modo geral, dentre estas separações, pode-se prever:

- Áreas com um determinado percentual de chance de haver petróleo, sal etc.
- Áreas com características paleogeográficas ou paleoclimáticas em comum.

Capítulo 5

Conclusões

Neste Capítulo são relatados os resultados obtidos, e após, são feitas as considerações sobre as contribuições científicas. De modo simples os resultados da pesquisa e do desenvolvimento do trabalho serão ligados com as possibilidades de benefícios tanto para a academia quanto para a indústria petrolífera. Finalmente serão apresentados alguns dos possíveis trabalhos futuros.

5.1 Resultados Obtidos

Um banco de dados paleogeográficos e paleoclimaticos foi desenvolvido. Seus dados foram coletados de diversas fontes e grande parte deles foram adaptados para serem representados de forma numérica. Esta forma numérica constitui-se num modelo representativo para os dados e paleodados das bacias sedimentares brasileiras.

Milhões de dados compõem a estrutura proposta; em sua maioria dados estratigráficos que representam camadas de depósitos sedimentares referentes tanto ao período atual, como à períodos passados desde a separação dos continentes da África e da América do Sul. Acompanhando os dados estratigráficos, estão diversos metadados que ajudam no entendimento da composição sedimentar do terreno. Além de outros dados que são importantes indicadores climáticos, como os níveis de Gás Carbônico.

Como relatado no capítulo 4, o trabalho evoluiu de um sistema simples e pouco coerente com os aspectos geológicos, a um sistema realístico que representa os dados naturais e abre margem para representá-los de forma evolutiva ao longo dos 140 milhões de anos.

Foram obtidas, e mapeadas ao longo da costa brasileira, duas malhas de dados. A primeira possui coordenadas, batimetria e dados gravimétricos. A segunda representa os dados, e metadados, litoestratigráficos, que foram mapeados automaticamente conforme descrito na seção 4.3, mais especificamente com a solução criada e descrita na seção 4.3.4.

A primeira malha pode ser visualizada na figura 4.4. Esta malha foi obtida com auxílio da ferramenta de ETL criada (descrita na seção 4.2). Esta ferramenta sofreu alterações de modo a comportar

os algoritmos desenvolvidos para a geração da segunda malha de dados. A versão final da ferramenta pode ser visualizada no apêndice E, juntamente com comentários descrevendo suas funções.

Os resultados referentes à segunda malha de dados podem ser conferidos no apêndice F. Os gráficos mostrados neste apêndice representam pontos em determinadas coordenadas, onde cada ponto possui um conjunto de informações que por sua vez estão armazenadas na tabela *pontos* do banco de dados (descrito no Capítulo 4.3.1).

O banco de dados final, somente para a idade geológica atual, armazena 83.273 pontos de dados ao longo das bacias. Conforme descrito na seção 4.3.2, foram obtidas 4.142.875 coordenadas para a primeira malha, contendo a gravimetria e a batimetria do terreno.

5.2 Contribuição Científica

É sabido que a extração de óleo é uma atividade em constante evolução e com descobertas recentes. Uma das contribuições científicas deste trabalho é ajudar na construção de conhecimento relativo aos elementos necessários para a formação do óleo.

É importante salientar que o banco de dados como um todo é resultado de pesquisas aplicadas de forma a agrupar dados e representar a evolução das bacias sedimentares brasileiras. Assim, a contribuição não se baseia em um simples modelo de banco de dados. Em uma visão geral, o banco de dados é resultado de todo processo de pesquisa e desenvolvimento descrito nesta dissertação.

O modelo criado é uma alternativa ao modelo estrela, que apesar de ser prático para a criação de um DW, demanda um grande espaço para ser armazenado, pois a tabela central sofre grande aumento de dados a cada nova *paleo_propriedade* adicionada (ver figura 4.7).

O modelo estrela possui diversas vantagens em relação ao modelo criado; este é mais claro, eficaz para criação de DW e possivelmente mais prático para KDD. De fato, o modelo criado não se propõe a ser o melhor modelo, este se propõe a ser uma alternativa que sacrifica flexibilidade, entre outros aspectos, para obter ganhos em relação ao armazenamento, uma vez que diversos tipos de dados distintos (*paleo_propriedade*) tendem a ser armazenados.

O modelo criado possui uma forma específica para comportar os dados requeridos pelo cenário da pesquisa, este foi testado com diversas cargas de dados com auxílio da ferramenta de ETL criada. Além disso, o formato do modelo, com uma tabela central baseadas em coordenadas, permite facilmente a adição de novos dados paleoclimáticos. Caso os novos dados tenham ligações com uma região, basta ligá-los a tabela principal (*pontos*), caso sejam dados referentes a toda Terra (ou toda a área em questão), estes podem ser ligados a tabela *Idades* (como exemplo: Fatores Climáticos).

O problema mais comum de se trabalhar com dados geográficos é o fato de que o valor de distância referente a longitude é alterada conforme a distância da linha do equador. Porém o algoritmo desenvolvido utiliza a fórmula de Haversine para mapear estes dados, assim a distância é calculada considerando a variação da longitude. Deste modo, o algoritmo detém uma, relativa, boa precisão em relação a distância dos pontos.

O algoritmo de mapeamento de dados mostra-se uma importante contribuição, uma vez que pode ser utilizado para propagar quaisquer dados, com representatividade numérica, ao longo de uma superfície na Terra.

Os dados presentes no banco são úteis para futuras pesquisas em geoinformática. Estes servem como pilar para a construção de conhecimento que poderá ser aplicado diretamente nas geociências e por consequência na indústria petrolífera.

Por fim, este trabalho gera potencial para significativas descobertas científicas, principalmente em questão dos trabalhos futuros, possíveis com base no que foi obtido. A seção abaixo descreve algumas possibilidades em relação ao que pode ser feito para extensão do trabalho descrito nesta dissertação.

5.3 Trabalhos Futuros

O trabalho relatado nesta dissertação abre caminho para muitos trabalhos futuros. Por se tratar de um trabalho multidisciplinar, que envolve diretamente duas ciências (computação e geologia), diversos ramos são possíveis para continuação da pesquisa. Novos trabalhos em ambas as ciências agregarão valor ao que foi criado. Novos conhecimentos podem ser obtidos, novas soluções poderão ser criadas para melhorar as mais distintas tarefas, principalmente aquelas ligadas à extração de óleo.

A quantidade de dados possíveis de serem agregados é praticamente infinita, além disso, a proveniência dos dados, a quantidade de um determinado elemento em um local e a atualidade dos dados; não apenas influenciam na precisão das informações extraídas mas também ajudam a extrair novos conhecimentos por meio de técnicas de KDD.

A absoluta maioria dos dados foram mapeados pela solução algorítmica criada para determinar sua localização. Devido a este fator, apesar dos testes e verificações visuais, é possível que muitos dados estejam com um mapeamento incoerente com a realidade. Desta forma, um trabalho futuro proposto é um sistema de verificação dos dados, que atualize coordenadas de pontos de acordo com dados empíricos.

Cada bacia possui várias sequências que são pacotes de rochas sedimentares relativos a um grupo no tempo geológico. Embora hajam subsequências, em geral, as duas grandes sequências presentes são a sequência Rift e a sequência Drift. Os dados presentes nestas divisões possuem diferentes fontes. Assim, um possível trabalho futuro, seria a classificação dos dados (presentes no banco) quanto as subsequências.

A ANP disponibiliza anualmente informações sobre os locais com extração ativa de petróleo. Estes dados podem ser úteis para processos de KDD, de modo a descobrir padrões entre eles e determinar chances de haver petróleo em outras bacias baseando-se nos dados extraídos destas áreas. A figura 5.1 é um mapa com as áreas de extração (2011). As áreas em amarelo ao longo da costa são as áreas das bacias sedimentares oceânicas (áreas de interesse).

Com as coordenadas do mapa é possível realizar cruzamentos com os dados presentes no banco de dados. Para isso, é necessário extrair do mapa os dados de coordenadas e, possivelmente, o volume

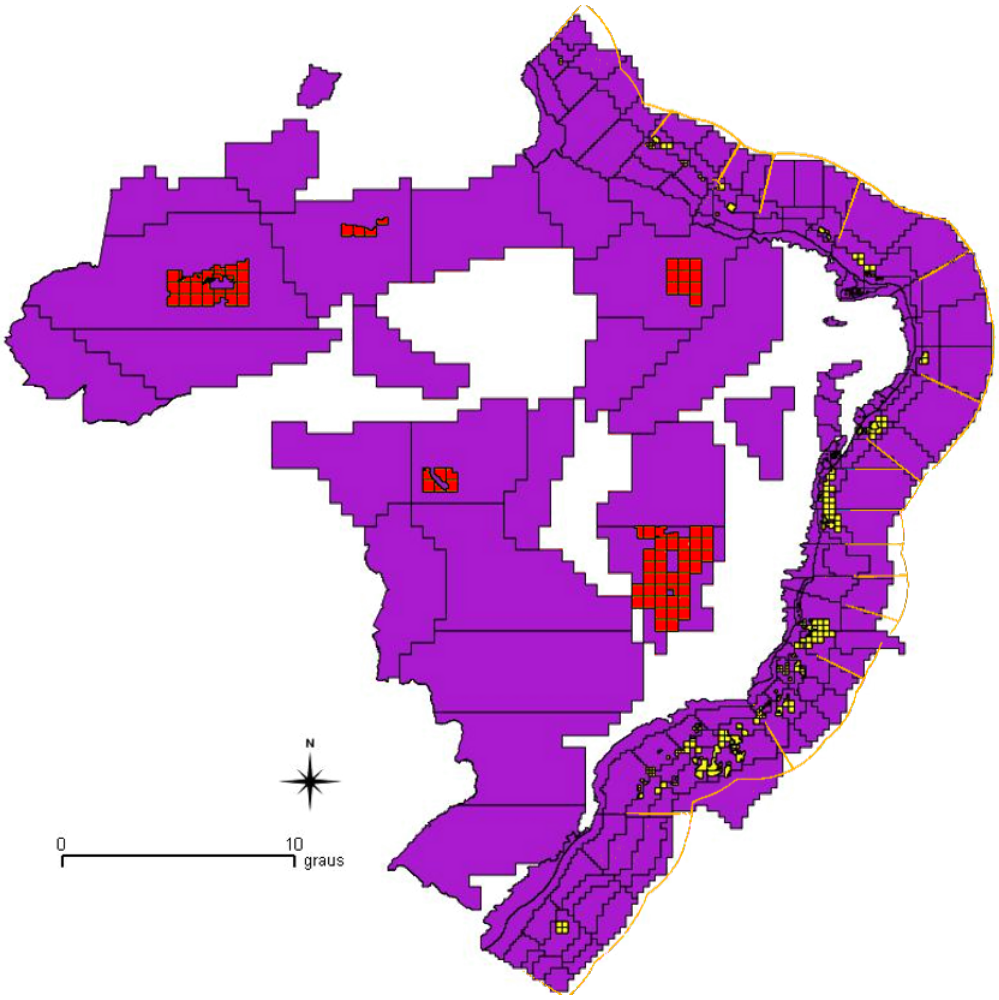


Figura 5.1: Atuais áreas de extração de petróleo e gás, limites e bacias.

de petróleo retirado e estimado para a região. Todos esses dados são disponibilizados pela ANP em forma gráfica e em documentos de texto. Extrair e popular o banco de dados com estas informações consiste em um trabalho futuro importante, pois o cruzamento dos dados empíricos mais atuais torna mais provável a descoberta de conhecimento. Assim, uma arquitetura (ou sistema) para a extração e carga destas informações se torna necessária.

Dados de Sísmica Marítima apresentam as camadas de litologias no tempo presente com uma maior precisão que os provenientes das cartas, pois estes são coletados diretamente do referido local, logo, estes dados são fundamentais para manter o banco de dados atualizado. Além disso, os dados provenientes da Sísmica, podem ser coletados e usados para agregar precisão, verificar e estabelecer uma ordem nos dados do banco.

A ampliação do modelo do banco de dados também constitui um importante trabalho futuro. A medida que novos tipos de dados forem sendo obtidos, o modelo deve ser expandido. Para isso a ferramenta de ETL criada deve ser melhorada e atualizada para servir ao novo cenário de dados e banco de dados.

A medida que muitos dados são produzidos são necessários procedimentos para assegurar a sua qualidades. Propõe-se a criação de um modelo de verificação para os dados, juntamente com um sistema especialista que valide os dados presentes no banco de dados. Este sistema deve ser flexível para aceitar não somente a entrada de dados reais, mas também a entrada de conhecimento geológico. Esse conhecimento deve ser interpretado pelo software e transformado em regras, que por sua vez, devem ser analisadas para localizar possíveis dados discrepantes, errôneos ou mapeados em local errado.

Todo o resultado da pesquisa, descrito aqui, pode ser replicado para a costa Africana. A única diferença seria, basicamente, a obtenção dos dados fontes (cartas estratigráficas, coordenadas etc.). O desenvolvimento deste trabalho ajudaria a criar um modelo de evolução mais realístico e completo, já que as bacias da costa Brasileira e Africana um dia foram uma só.

Por fim, como dito no plano de mineração 4.3.5, novos algoritmos de mineração, especializados em dados paleogeográficos e paleoclimáticos, podem ser decisivos para a descoberta de conhecimento neste banco, pois as possibilidades de descoberta em meio aos dados paleogeográficos são tantas que possivelmente sejam necessários novos algoritmos para se obter toda informação previamente desconhecida.

Referências Bibliográficas

[Cor00] Cordani, G.; Milani, J.; Thomaz Filho, A.; Campos A. "Tectonic Evolution of South America". Geological Society, 2000, 856p.

[Fan10] Cheng-fang, L.; Xue-jun, w.; Tao, M.; Tie-cheng, W.; Yong, L.; Wei, Y. "Research on petroleum seismic data resource quality management methods". *International Conference on Environmental Science and Information Application Technology (ESIAT), 2010*, Julho 2010, pp. 245 - 248.

[Fay96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth,P. "Knowledge Discovery and Data Mining: Towards a Unifying Framework". *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996, pp. 82 - 88.

[Gel89] Gellert, W.; Gottwald, M.; Hellwich, M.; Kästner, H.; Küstner H. "Global seafloor topography from satellite altimetry and ship depth soundings". Van Nostrand Reinhold, 1989, 760p.

[Gol06] Golden Software. "SURFER". Disponivel em: "http://www.goldensoftware.com/products/surfer/surfer.shtml". Acessado em: Julho de 2011.

[Hal09] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. "The WEKA Data Mining Software: An Update". University of Waikato. Disponivel em: "http://www.cs.waikato.ac.nz/ml/weka/". Acessado em: Junho de 2011.

[Han01] Han, J.; and Kamber, M. "Data mining: Concepts and techniques". San Mateo: Morgan Kaufmann, 2001, 550p.

[Han01B] Hand, D.; Mannila H.; Smyth, P. "Principles of Data Mining". MIT Press, 2001, 425p.

[Han06] Han, J; and Kamber, M. "Data mining: Concepts and techniques, Second Edition". San Mateo: Morgan Kaufmann, 2006, 800p.

[Hes62] Hess, H. "History of Ocean Basins". *Petrologic studies: A volume of honor*, vol. 1-1, Agosto 1962, pp. 599 - 620.

[Hon10] Hongbo Wu and Yanqiu Xing. "Pixel-based image fusion using wavelet transform for SPOT and ETM+ image", *IEEE International Conference on Progress in Informatics and Computing (PIC)*, vol. 1-1, Dezembro 2010, pp. 936–940.

[Inp11] INPE; Terraview. "Conceitos Cartográficos". Terraview. Disponivel em: "http://www.dpi.inpe.br/terraview/docs/pdf/ProjecaoCartografica.pdf". Acessado em: Agosto de 2011.

[Jah08] Jahn, F.; Cook, M.; Graham, M. "Hydrocarbon Exploration & Production, Volume 55, Second Edition (Developments in Petroleum Science)". Elsevier Science., 2008, 456p.

[Ket10] Ketzer, J. "Projeto mapeia parte do atlântico para exploração de petróleo". Disponivel em: "http://www.planetauniversitario.com". Acessado em: Julho de 2011.

[Kim02] Kimbal, R. and Ross, M. "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)". Wiley, 2002, 464p.

[Kim04] Kimbal, R. and Caserta, J. "The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data". Wiley, 2004, 528p.

[Mat93] Matheus, C.; Chan, P.; Piatetsky-Shapiro, G. "Systems for Knowledge Discovery in Databases". *IEEE Transactions on Knowledge and Data Engineering*, 1993, pp. 903 - 313.

[Mil00] Milani, E.J.; Thomaz Filho, A.; Campos, D.A.; Cordani, U.G. "Tectonic Evolution of South America". Geological Society., 2000, 854p.

[Mil07] Milani, E.J.; Rangel, D.H.; Bueno, G.V.; Stica, J.M.; Winter, W.R.; Caixata,J.M.; Neto, O. C. P. "Boletim de Geociências da Petrobras". Centro de Pesquisas Leopoldo A. Miguez de Mello., Vol.15 nº2, 2007, 573p.

[Mil09] Miller, H,J.; Han, J. "Geographic Data Mining and Knowledge Discovery". CRC Press, 2009, 484p.

[Mou10] Moulin, M.; Aslanian, D.; Unternehr, P. "Earth-Science Reviews". *Geological Society*, vol. 98, Agosto 2010, pp. 1 - 37.

[Mul08] Müller, D.; Sdrolias, M.; Gaina, C.; Roest W "Age, spreading rates, and spreading asymmetry of the world's ocean crust". *Geochemistry, Geophysics, Geosystems, 9*, vol. Q04006, Agosto 2008.

[Net04] Neto, A.; Ponzi, R.; Sichel, S. "Introdução a Geologia Marinha". Interciência, 2004, 279p.

[Pop84] POPP, J. H. “Geologia Geral, 3ª edição”. LTC - Livros Técnicos e Científicos S.A., Inc., 1984, 283p.

[Pop98] POPP, J. “Geologia Geral, 5ª edição”. LTC - Livros Técnicos e Científicos S.A., 1998, 400p.

[Pre06] Press, F.; Siever, R.; Grotzinger, J.; Jordan, T. “Para entender a terra, 4ª edição”. Bookman, 2006, 656p.

[Sil97] Silberchartz, A.; Korth, H.; Sudarshan, S. “Sistema de Banco de Dados”. MAKRON Books, 1997, 808p.

[Sed98] Sedgewick, R. “Algorithms in C”. Addison-Wesley, 1998, 702p.

[She01] Shekhar, S.; Huang, Y.; Han, J.; Chawla, S.; Gopal, S. “Categorization of Spatial Data Mining Techniques”. *Scientific Data Mining*, 2001, pp. 3642 - 3646.

[Smi10] Smith, W.; and Sandwell, D. “Global seafloor topography from satellite altimetry and ship depth soundings”. *Science*, vol. 45, Setembro 2010, pp. 1957 - 1962.

[Sol90] Soluri, E.A. and Woodson, V.A. “World Vector Shoreline”, *Science*, Disponivel em: “http://www.ngdc.noaa.gov/mgg/coast/wvs.html”. Acessado em: Julho de 2011.

[Tan93] Tansel, A.; Clifford, J.; Gadia, S.; Jajodia, S.; Segev, A.; Snograss, R. “Temporal Databases”. The Benjamin/Cummings, 1993, 656p.

[Tan05] Tan, P.; Steinbach, M.; Kumar, V. “Introduction to Data Mining”. Addison-Wesley Longman Publishing, 2005, 769p.

[USG09] U.S. Geological Survey. “U.S. Geological Survey”. Disponivel em: “http://pubs.usgs.gov/gip/dynamic/historical.html”. Acessado em: Agosto de 2011.

[Wan04] Wang Qiang and Shen Yi. “The effects of fusion structures on image fusion performances”, *Proceedings of the 21st IEEE Instrumentation and Measurement Technology Conference. IMTC 04.* , vol. 1-1, Maio 2004, pp. 468–471.

[Wan08] Wang, T.; Chen, X.; Bao, A.; Wang, W. “A new Geospatial Data Model to Facilitate Geographic Data Mining and Knowleadge Discovery”. *IEEE, International Conference on Systems, Man and Cybernetcs (SMC 2008)*, 2008, pp. 3642 - 3646.

Apêndice A

Documentos auxiliares

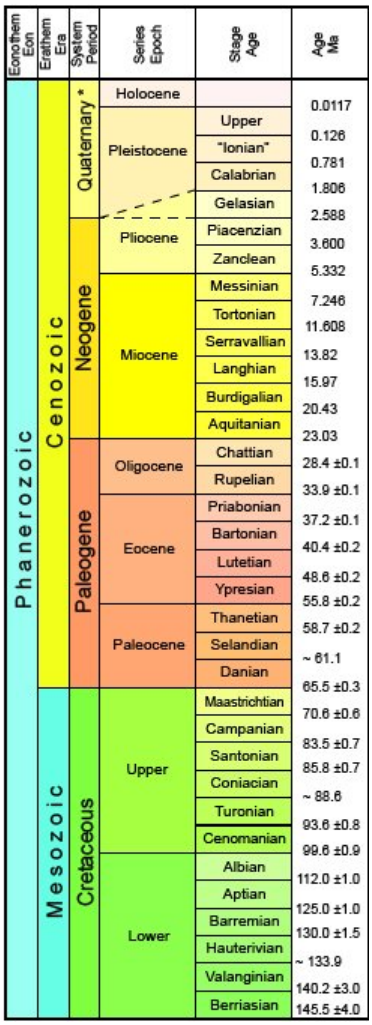


Figura A.1: Parte do *International Stratigraphic Chart* com as idades geológicas usadas.

Símbolo	Rocha	Valor
	Sandstone	2
	Siltito	4
	Folhelho	8
	Calcarenito	16
	Calcilutito	32
	Calcissiltito	64
	Diamictito	128
	Argilito	256
	Calcirrudito	512
	Halita	1024
	Conglomerado	2048
	marga	4096
	Anidrita/Gipsita	8192
	Silexito	16384
	Basalto	32768
	Dolomito	65536
	Coquina	131072
	Diabásio	262144
	metamórfica NI	524288
	Ígnea Não Especificada	1048576
	Ígnea Ácida	2097152
	Ígnea Alcalina	4194304
	Composta com dados branco	1

Figura A.2: Litologias utilizadas e seus respectivos valores.

Bacia	LC-Lat	LC-Lon	QP-Lat	QP-Lon	L3k-Lat	L3k-lon
Pelotas Sul	-33.7	306.6	-34.8277	307.8279	-35.31	309.0083
Pel-H00	-32.24	307.9	-33.15	309.48	-34.0	310.725
Pel-H1	-30.8	309.6083	-31.3985	310.2165	-32.0	311.725
Pel-H2	-29.9006	310.0250	-30.4990	311.8083	-31.0	312.391
Pel-H2-2	-29.3	310.4	-30.1	311.9	-30.2	313.9750
Pel-H3	-28.6065	311.0750	-29.0987	311.9309	-29.6	314.1750
Pel-H4	-28.0	311.6250	-28.3987	313.1000	-28.7	314.4583
Pel-San	-27.8	311.4	-27.8348	312.3855	-27,8134	314,8583
San-H0	-25.38	311.02	-26.43	314	-27.33	316
San-H1	-23.99	313.74	-25.30	315.08	-28.47	318.4
San-H2	-23.09	316.42	-23.88	316.87	-27.13	318.4750
San-Cam	-22.9	317.8	-23.4100	318.1083	-25.8354	319.4083
Cam-H0	-22.34	318.28	-23.11	319.12	-24.08	320.11
Cam-H1	-21.99	319.13	-22.42	319.61	-23.15	320.58
Cam-H2	-21.38	319.03	-21.99	320.08	-22.78	321.07
Cam-H3	-20.89	319.26	-21.18	319.7	-22.273	321.38
Cam-ES	-20.4	319.6	-20.7055	320.0446	-21.67	321.53
ES-H0	-19.93	319.9	-20.09	320.2	-20.88	321.7
ES-H1	-19.49	320.37	-19.68	320.73	-20.78	322.93
ES-H2	-18.89	320.32	-19.51	321.65	-20.00	322.66
ES-Muc	-18.4	320.3	-18.9116	322.0399	-19.01	322.45
Muc-H0	-18.01	320.5	-18.45	322.25	-18.53	322.42
Muc-H1	-17.80	320.79	-18.07	322.63	-18.13	323.00
Muc-Cum	-17.6	320.7	-17.6199	321.9583	-17.70	322.80
Cum-H0	-17.17	320.8	-17.20	321.37	-17.26	323.37
Cum-H1	-16.73	320.89	-16.74	321.3	-16.76	322.58
Cum-Jeq	-16.3	320.9	-16.3058	321.7583	-16.33	322.33
Jeq-H0	-15.85	321.15	-15.85	322	-15.85	322.25
Jeq-H1	-15.56	321.05	-15.56	321.38	-15.56	322.02
Jeq-Camamu	-14.8	320.9	-14.8	321.0980	-14.81	321.88

Figura A.3: Coordenadas utilizadas - parte 1.

Jeq-Camamu	-14.8	320.9	-14.8	321.0980	-14.81	321.88
Cama-H0	-14.05	321.07	-14.05	321.18	-14.05	321.83
Cama-H1	-13.34	321.05	-13.34	321.30	-13.34	322.22
Camamu-Jacuipe	-13.0	321.5	-13.0081	321.6083	-13.00	322.68
Jacuipe-H0	-12.22	322.25	-12.24	322.46	-12.31	323.02
Jac-Sergipe-Al	-11.5	322.6	-11.5220	322.7773	-11.64	323.55
Sergipe-H0	-10.5	323.6	-10.58	323.86	-10.76	324.4
Sergipe-H1	-9.65	324.43	-9.72	324.72	-9.77	325.1
Sergipe_Al-Pa-Pe	-8.8	324.9	-8.9101	325.1506	-8.98	325.88
Paraiba-H0	-7.9	325.19	-7.9	325.5	-7.9	326.00
Paraiba-H1	-7.18	325.2	-7.16	325.55	-7.15	325.8
Paraiba_Pe-Potigiar	-6.5	325.1	-6.47	325.25	-6.44	325.96
Potiguar-H0	-5.18	324.54	-4.76	324.77	-4.43	325.22
Potiguar-H1	-5.03	324.03	-4.79	324.13	-3.50	324.75
Potiguar-H2	-4.95	323.11	-4.62	323.22	-3.47	323.59
Pot-Cea	-3.8	321.4	-3.2930	321.7083	-2.42	321.7
Ceara-H0	-2.86	320	-2.17	320.04	-1.52	320.1
Cea-Barreirinhas	-2.75	318.1	-2.1899	318.15	-0.16	318.24
Barreir.-H0	-2.39	317.92	-1.72	317.03	-0.13	317.25
Bar-Para_Ma	-2.2	315.7	-0.5916	316.1083	0.09	316.3
Para_Ma-H0	-1.24	314.28	0.27	315.04	1.02	315.41
Pa_Ma-Foz	-0.6	312.2	1.4	313.18	2.21	313.63
Foz-H0	0.6	309.95	3.13	311.73	4.75	312.86
Foz-H1	1.7	310	3.47	311.28	4.94	312.35
Foz-H2	2.46	309.22	4.13	310.48	5.3	311.37
Foz-H3	4.11	308.79	5.15	309.48	5.65	309.76
Foz-Oeste	4.3	308.4	5.1966	309.1044	6.0	309.33

Figura A.4: Coordenadas utilizadas - parte 2.

Apêndice B

Resultados da mineração com fonte das cartas de 2003

```
 4. Campanian=5 860 ==> Classe=S 860    conf:(1)
 5. Campanian=5 Valanginian=513 816 ==> Classe=S 816    conf:(1)
 6. Campanian=5 Hauterivian=4097 814 ==> Classe=S 814    conf:(1)
 7. Campanian=5 Hauterivian=4097 Valanginian=513 799 ==> Classe=S 799    conf:(1)
 8. Coniacian=5 691 ==> Hauterivian=4097 691    conf:(1)
 9. Coniacian=5 691 ==> Classe=S 691    conf:(1)
10. Coniacian=5 Classe=S 691 ==> Hauterivian=4097 691    conf:(1)
11. Coniacian=5 Hauterivian=4097 691 ==> Classe=S 691    conf:(1)
12. Coniacian=5 691 ==> Hauterivian=4097 Classe=S 691    conf:(1)
13. Priabonian=4 671 ==> Classe=S 671    conf:(1)
14. Turonian=5 659 ==> Classe=S 659    conf:(1)
15. Coniacian=5 Valanginian=513 653 ==> Hauterivian=4097 653    conf:(1)
16. Coniacian=5 Valanginian=513 653 ==> Classe=S 653    conf:(1)
17. Coniacian=5 Valanginian=513 Classe=S 653 ==> Hauterivian=4097 653    conf:(1)
18. Coniacian=5 Hauterivian=4097 Valanginian=513 653 ==> Classe=S 653    conf:(1)
19. Coniacian=5 Valanginian=513 653 ==> Hauterivian=4097 Classe=S 653    conf:(1)
20. Serravallian=5 641 ==> Classe=S 641    conf:(1)
21. Serravallian=5 Hauterivian=4097 636 ==> Classe=S 636    conf:(1)
22. Turonian=5 Hauterivian=4097 634 ==> Classe=S 634    conf:(1)
23. Bartonian=4 629 ==> Classe=S 629    conf:(1)
24. Messinian=4 620 ==> Classe=S 620    conf:(1)
25. Serravallian=5 Turonian=5 615 ==> Classe=S 615    conf:(1)
26. Thanetian=5 614 ==> Classe=S 614    conf:(1)
27. Serravallian=5 Turonian=5 Hauterivian=4097 610 ==> Classe=S 610    conf:(1)
28. Campanian=5 Coniacian=5 609 ==> Hauterivian=4097 609    conf:(1)
29. Campanian=5 Coniacian=5 609 ==> Valanginian=513 609    conf:(1)
30. Campanian=5 Coniacian=5 609 ==> Classe=S 609    conf:(1)
31. Campanian=5 Coniacian=5 Valanginian=513 609 ==> Hauterivian=4097 609    conf:(1)
32. Campanian=5 Coniacian=5 Hauterivian=4097 609 ==> Valanginian=513 609    conf:(1)
33. Campanian=5 Coniacian=5 609 ==> Hauterivian=4097 Valanginian=513 609    conf:(1)
34. Campanian=5 Coniacian=5 Classe=S 609 ==> Hauterivian=4097 609    conf:(1)
35. Campanian=5 Coniacian=5 Hauterivian=4097 609 ==> Classe=S 609    conf:(1)
36. Campanian=5 Coniacian=5 609 ==> Hauterivian=4097 Classe=S 609    conf:(1)
37. Campanian=5 Coniacian=5 Classe=S 609 ==> Valanginian=513 609    conf:(1)
38. Campanian=5 Coniacian=5 Valanginian=513 609 ==> Classe=S 609    conf:(1)
39. Campanian=5 Coniacian=5 609 ==> Valanginian=513 Classe=S 609    conf:(1)
40. Campanian=5 Coniacian=5 Valanginian=513 Classe=S 609 ==> Hauterivian=4097 609    conf:(1)
41. Campanian=5 Coniacian=5 Hauterivian=4097 Classe=S 609 ==> Valanginian=513 609    conf:(1)
42. Campanian=5 Coniacian=5 Hauterivian=4097 Valanginian=513 609 ==> Classe=S 609    conf:(1)
43. Campanian=5 Coniacian=5 Classe=S 609 ==> Hauterivian=4097 Valanginian=513 609    conf:(1)
44. Campanian=5 Coniacian=5 Valangini an=513 609 ==> Hauterivian=4097 Classe=S 609    conf:(1)
45. Campanian=5 Coniacian=5 Hauterivian=4097 609 ==> Valanginian=513 Classe=S 609    conf:(1)
46. Campanian=5 Coniacian=5 609 ==> Hauterivian=4097 Valanginian=513 Classe=S 609    conf:(1)
```

Figura B.1: Resultados da mineração na etapa 1

Apêndice C

Script de criação do banco

```
CREATE TABLE FORMACAO (
  idFORMACAO INTEGER   NOT NULL ,
  Bacia VARCHAR(50)     ,
  Formacao VARCHAR(30)     ,
  Litologias VARCHAR(99)     ,
  AmbienteSedimentacao VARCHAR(99)     ,
  PPSP VARCHAR(3)       ,
PRIMARY KEY(idFORMACAO));

CREATE TABLE PROVENIENCIAS (
  idProveniencia INTEGER   NOT NULL ,
  Dt_Cartas DATE     ,
  Dt_criacao DATE     ,
  Metodo VARCHAR(99)     ,
  Fonte VARCHAR(99)     ,
  Url VARCHAR(99)       ,
PRIMARY KEY(idProveniencia));

CREATE TABLE ROCHAS (
  idROCHAS INTEGER   NOT NULL ,
  Litologia VARCHAR(50)     ,
  Valor INTEGER     ,
  Tipo VARCHAR(50)     ,
  Origem VARCHAR(50)     ,
  Composicao VARCHAR(40)     ,
  SubComposicao VARCHAR(40)     ,
  Granulometria VARCHAR(50)     ,
  OBS VARCHAR(99)       ,
PRIMARY KEY(idROCHAS));

CREATE TABLE EONS (
  idEon INTEGER   NOT NULL ,
  Eon VARCHAR(30)     ,
  Obs VARCHAR(99)       ,
PRIMARY KEY(idEon));

CREATE TABLE ERAS (
  idEra INTEGER   NOT NULL ,
  EONS_idEon INTEGER   NOT NULL ,
  Era VARCHAR(30)     ,
  Obs VARCHAR(99)       ,
PRIMARY KEY(idEra)   ,
  FOREIGN KEY(EONS_idEon)
    REFERENCES EONS(idEon));

CREATE INDEX Era_FKIndex1 ON ERAS (EONS_idEon);
CREATE INDEX IFK_Rel_04 ON ERAS (EONS_idEon);
```

```
CREATE TABLE PERIODOS (
  idPeriodo INTEGER   NOT NULL ,
  ERAS_idEra INTEGER   NOT NULL ,
  Periodo VARCHAR(30)    ,
  Obs VARCHAR(99)      ,
PRIMARY KEY(idPeriodo)  ,
  FOREIGN KEY(ERAS_idEra)
    REFERENCES ERAS(idEra));


CREATE INDEX Periodo_FKIndex1 ON PERIODOS (ERAS_idEra);
CREATE INDEX IFK_Rel_05 ON PERIODOS (ERAS_idEra);

CREATE TABLE EPOCAS (
  idEpoca INTEGER   NOT NULL ,
  PERIODOS_idPeriodo INTEGER   NOT NULL ,
  Epoca VARCHAR(30)    ,
  Obs VARCHAR(99)      ,
PRIMARY KEY(idEpoca)  ,
  FOREIGN KEY(PERIODOS_idPeriodo)
    REFERENCES PERIODOS(idPeriodo));

CREATE INDEX Epoca_FKIndex1 ON EPOCAS (PERIODOS_idPeriodo);
CREATE INDEX IFK_Rel_06 ON EPOCAS (PERIODOS_idPeriodo);

CREATE TABLE IDADES (
  idIDADES INTEGER   NOT NULL ,
  EPOCAS_idEpoca INTEGER   NOT NULL ,
  Idade VARCHAR(30)    ,
  Obs VARCHAR(99)    ,
  MA_Ini DECIMAL(5,2)    ,
  MA_fim DECIMAL(5,2)      ,
PRIMARY KEY(idIDADES)  ,
  FOREIGN KEY(EPOCAS_idEpoca)
    REFERENCES EPOCAS(idEpoca));

CREATE INDEX IDADES_FKIndex1 ON IDADES (EPOCAS_idEpoca);
CREATE INDEX IFK_Rel_07 ON IDADES (EPOCAS_idEpoca);

CREATE TABLE FATORESCLIMATICOS (
  idFATORESCLIMATICOS INTEGER   NOT NULL ,
  IDADES_idIDADES INTEGER   NOT NULL ,
  CO2 DECIMAL(4,2)    ,
  isotopoo18 DECIMAL(4,2)   NOT NULL ,
  MAA INTEGER    ,
  OBS VARCHAR(99)    ,
  Fonte VARCHAR(99)    ,
  Dt_fonte DATE      ,
PRIMARY KEY(idFATORESCLIMATICOS)  ,
  FOREIGN KEY(IDADES_idIDADES)
    REFERENCES IDADES(idIDADES));
```

```
CREATE INDEX FATORESCLIMATICOS_FKIndex1 ON FATORESCLIMATICOS (IDADES_idIDADES);
CREATE INDEX IFK_Rel_10 ON FATORESCLIMATICOS (IDADES_idIDADES);

CREATE TABLE PONTOS (
  IDADES_idIDADES INTEGER   NOT NULL ,
  FORMACAO_idFORMACAO INTEGER   NOT NULL ,
  Lat_at DECIMAL(6,4)   NOT NULL ,
  Lon_at DECIMAL(7,4)   NOT NULL ,
  PROVENIENCIAS_idProveniencia INTEGER   NOT NULL ,
  Valor INTEGER    ,
  Bacia VARCHAR(50)    ,
  RPB VARCHAR(50)    ,
  PRD INTEGER    ,
  Obs VARCHAR(99)    ,
  Gravimetria DECIMAL(7,4)    ,
  Profundidade INTEGER    ,
  Lat DECIMAL(6,4)    ,
  Lon DECIMAL(7,4)      ,
PRIMARY KEY(IDADES_idIDADES, FORMACAO_idFORMACAO, Lat_at, Lon_at)      ,
  FOREIGN KEY(IDADES_idIDADES)
    REFERENCES IDADES(idIDADES),
  FOREIGN KEY(FORMACAO_idFORMACAO)
    REFERENCES FORMACAO(idFORMACAO),
  FOREIGN KEY(PROVENIENCIAS_idProveniencia)
    REFERENCES PROVENIENCIAS(idProveniencia));

CREATE INDEX PONTOS_FKIndex1 ON PONTOS (IDADES_idIDADES);
CREATE INDEX PONTOS_FKIndex2 ON PONTOS (FORMACAO_idFORMACAO);
CREATE INDEX PONTOS_FKIndex3 ON PONTOS (PROVENIENCIAS_idProveniencia);

CREATE INDEX IFK_Rel_15 ON PONTOS (IDADES_idIDADES);
CREATE INDEX IFK_Rel_11 ON PONTOS (FORMACAO_idFORMACAO);
CREATE INDEX IFK_Rel_13 ON PONTOS (PROVENIENCIAS_idProveniencia);

CREATE TABLE CAMADAS (
  idCamadas INTEGER   NOT NULL ,
  PONTOS_Lon_at DECIMAL(7,4)   NOT NULL ,
  PONTOS_Lat_at DECIMAL(6,4)   NOT NULL ,
  PONTOS_FORMACAO_idFORMACAO INTEGER   NOT NULL ,
  PONTOS_IDADES_idIDADES INTEGER   NOT NULL ,
  Dt_Sismica DATE    ,
  Especura DECIMAL(4,1)    ,
  Litologia INTEGER    ,
  Idade_litologia INTEGER    ,
  Lat_ini DECIMAL(6,4)    ,
  Lon_ini DECIMAL(7,4)    ,
  Lat_fim INTEGER(6,4)    ,
  Lon_fim DECIMAL(7,4)      ,
PRIMARY KEY(idCamadas)  ,
  FOREIGN KEY(PONTOS_IDADES_idIDADES, PONTOS_FORMACAO_idFORMACAO, PONTOS_Lat_at, PONTOS_Lon_at)
    REFERENCES PONTOS(IDADES_idIDADES, FORMACAO_idFORMACAO, Lat_at, Lon_at));
```

```
CREATE INDEX Camadas_FKIndex1 ON CAMADAS (PONTOS_IDADES_idIDADES,
    PONTOS_FORMACAO_idFORMACAO, PONTOS_Lat_at, PONTOS_Lon_at);
CREATE INDEX IFK_Rel_12 ON CAMADAS (PONTOS_IDADES_idIDADES,
    PONTOS_FORMACAO_idFORMACAO, PONTOS_Lat_at, PONTOS_Lon_at);

CREATE TABLE ROCHAS_PONTOS (
  ROCHAS_idROCHAS INTEGER   NOT NULL ,
  PONTOS_FORMACAO_idFORMACAO INTEGER   NOT NULL ,
  PONTOS_IDADES_idIDADES INTEGER   NOT NULL ,
  PONTOS_Lon_at DECIMAL(7,4)   NOT NULL ,
  PONTOS_Lat_at DECIMAL(6,4)   NOT NULL   ,
PRIMARY KEY(ROCHAS_idROCHAS, PONTOS_FORMACAO_idFORMACAO, PONTOS_IDADES_idIDADES,
       PONTOS_Lon_at, PONTOS_Lat_at)    ,
  FOREIGN KEY(ROCHAS_idROCHAS)
    REFERENCES ROCHAS(idROCHAS),
  FOREIGN KEY(PONTOS_IDADES_idIDADES, PONTOS_FORMACAO_idFORMACAO, PONTOS_Lat_at, PONTOS_Lon_at)
    REFERENCES PONTOS(IDADES_idIDADES, FORMACAO_idFORMACAO, Lat_at, Lon_at));

CREATE INDEX ROCHAS_has_PONTOS_FKIndex1 ON ROCHAS_PONTOS (ROCHAS_idROCHAS);
CREATE INDEX ROCHAS_has_PONTOS_FKIndex2 ON ROCHAS_PONTOS (PONTOS_IDADES_idIDADES,
           PONTOS_FORMACAO_idFORMACAO, PONTOS_Lat_at, PONTOS_Lon_at);

CREATE INDEX IFK_Rel_11 ON ROCHAS_PONTOS (ROCHAS_idROCHAS);
CREATE INDEX IFK_Rel_12 ON ROCHAS_PONTOS (PONTOS_IDADES_idIDADES,
              PONTOS_FORMACAO_idFORMACAO, PONTOS_Lat_at, PONTOS_Lon_at);
```

Apêndice D

Scripts de inserção de dados

```
insert into formacao (idFormacao,Bacia,Formacao,Litologias,ambienteSedimentacao, PPSP)
values (1, 'Pelotas', 'Fm Imbé IMB', 'Folhelho/Turbiditos','Marinho Profundo','GR');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,ambienteSedimentacao, PPSP)
values (2, 'Pelotas', 'Fm Cidreira (CID)', 'Clásticos grossos e finos','Leques costeiros progradantes. Ambiente nerítico.','');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,ambienteSedimentacao, PPSP)
values (3, 'Pelotas', 'Fm Atlântida (ATL)', 'Clásticos e carbonatos Finos','Plataforma externa','G');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,ambienteSedimentacao, PPSP)
values (4, 'Pelotas', 'Fm Porto Belo (PBL)', 'Calcarenito/Arenito','Plataforma carbonática','');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,ambienteSedimentacao, PPSP)
values (5, 'Pelotas', 'Fm Tramandaí (TRA)', 'Arenitos/Folhelhos/Siltitos','Marinho Raso (plataforma)','');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,ambienteSedimentacao, PPSP)
values (6, 'Pelotas', 'Fm Ariri (ARI)', 'Evaporitos','Transicional','C');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,PPSP)
values (7, 'Santos', 'Itajaí Açú', 'Folhelhos marinhos/Arenitos/Turbiditos ','GR');
insert into formacao (idFormacao,Bacia,Formacao,Litologias, OBS)
values (8, 'Santos', 'Barra velha', 'Folhelhos marinhos/Arenitos/Turbiditos ','Há reservatório de petróleo no pré-sal');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,PPSP)
values (9, 'Santos', 'Itapema', 'Folhelhos marinhos/Arenitos/Turbiditos ','R');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,PPSP)
values (10, 'Santos', 'Piçarras', 'Folhelhos marinhos/Arenitos/Turbiditos ','G');
insert into formacao (idFormacao,Bacia,Formacao,Litologias)
values (11, 'Santos', 'Camboriú', 'Folhelhos marinhos/Arenitos/Turbiditos ');
insert into formacao (idFormacao,Bacia,Formacao,Litologias)
values (12, 'Santos', 'Marambaia', 'Folhelhos/Diamictitos/Calcarenitos ');
insert into formacao (idFormacao,Bacia,Formacao,Litologias)
values (13, 'Santos', 'Iguapé', 'Folhelhos/Calcarenitos/Arenitos ');
insert into formacao (idFormacao,Bacia,Formacao,Litologias)
values (14, 'Santos', 'Ponta Aguda', 'Folhelhos/Calcarenitos/Arenitos');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,PPSP)
values (15, 'Santos', 'Santos', 'Arenitos','R');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,PPSP)
values (16, 'Santos', 'Guaruja', 'Carbonatos','R');
insert into formacao (idFormacao,Bacia,Formacao,Litologias,PPSP)
values (17, 'Santos', 'Jureia', 'Arenitos','R');
insert into formacao (idFormacao,Bacia,Formacao,Litologias)
values (18, 'Santos', 'Itanhaém', 'Folhelhos/marga/Arenito');
insert into formacao (idFormacao,Bacia,Formacao,Litologias)
values (19, 'Santos', 'Florianópolis', '');
insert into formacao (idFormacao,Bacia,Formacao,Litologias, PPSP)
values (20, 'Santos', 'Ariri', 'Evaporitos','C');
```

```
insert into EONS ("IDEON","EON","OBS") values(1,'Phanerozoic','The Only after Precambrians');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(1,1,'Cenozoic','');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(2,1,'Mesozoic','');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(3,1,'Paleozoic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS")
    values(1,1,'Quaternary','Under discution. International Stratigraphic Chart 2008');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(2,1,'Neogene','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(3,1,'Paleogene','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(4,2,'Cretaceous','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(5,2,'Jurassic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(6,2,'Triassic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(7,3,'Permian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(8,3,'Carboniferous','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(9,3,'Devonian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(10,3,'Silurian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(11,3,'Ordovician','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(12,3,'Cambrian','');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(1,1,'Holocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA","OBS") values(2,1,'Pleistocene','Used by us like on Age');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(3,2,'Pilocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(4,2,'Miocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(5,3,'Oligocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(6,3,'Eocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(7,3,'Paleocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(8,4,'Upper');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(9,4,'Lower');
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(1,2,'Upper','',0.126,0.0117);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(2,2,'Ionian','',0.781,0.126);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(3,2,'Calabrian','',1.806,0.781);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(4,2,'Gelasian','Begin of Epoch 2:Plaistocene',2.588,1.806);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(5,3,'Piacenzian','By us, second age, first is full pleistocene',3.600,2.588);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(6,3,'Zancleanan','Begin of Epoch 3:Pilocene',5.332,3.600);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(7,4,'Messinian','',7.246,5.332);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(8,4,'Tortonian','',11.608,7.246);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(9,4,'Serravallian','',13.82,11.608);
```

```
insert into EONS ("IDEON","EON","OBS") values(1,'Phanerozoic','The Only after Precambrians');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(1,1,'Cenozoic','');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(2,1,'Mesozoic','');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(3,1,'Paleozoic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS")
    values(1,1,'Quaternary','Under discution. International Stratigraphic Chart 2008');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(2,1,'Neogene','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(3,1,'Paleogene','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(4,2,'Cretaceous','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(5,2,'Jurassic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(6,2,'Triassic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(7,3,'Permian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(8,3,'Carboniferous','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(9,3,'Devonian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(10,3,'Silurian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(11,3,'Ordovician','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(12,3,'Cambrian','');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(1,1,'Holocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA","OBS") values(2,1,'Pleistocene','Used by us like on Age');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(3,2,'Pilocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(4,2,'Miocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(5,3,'Oligocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(6,3,'Eocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(7,3,'Paleocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(8,4,'Upper');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(9,4,'Lower');
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(1,2,'Upper','',0.126,0.0117);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(2,2,'Ionian','',0.781,0.126);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(3,2,'Calabrian','',1.806,0.781);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(4,2,'Gelasian','Begin of Epoch 2:Plaistocene',2.588,1.806);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(5,3,'Piacenzian','By us, second age, first is full pleistocene',3.600,2.588);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(6,3,'Zancleanan','Begin of Epoch 3:Pilocene',5.332,3.600);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(7,4,'Messinian','',7.246,5.332);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(8,4,'Tortonian','',11.608,7.246);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(9,4,'Serravallian','',13.82,11.608);
```

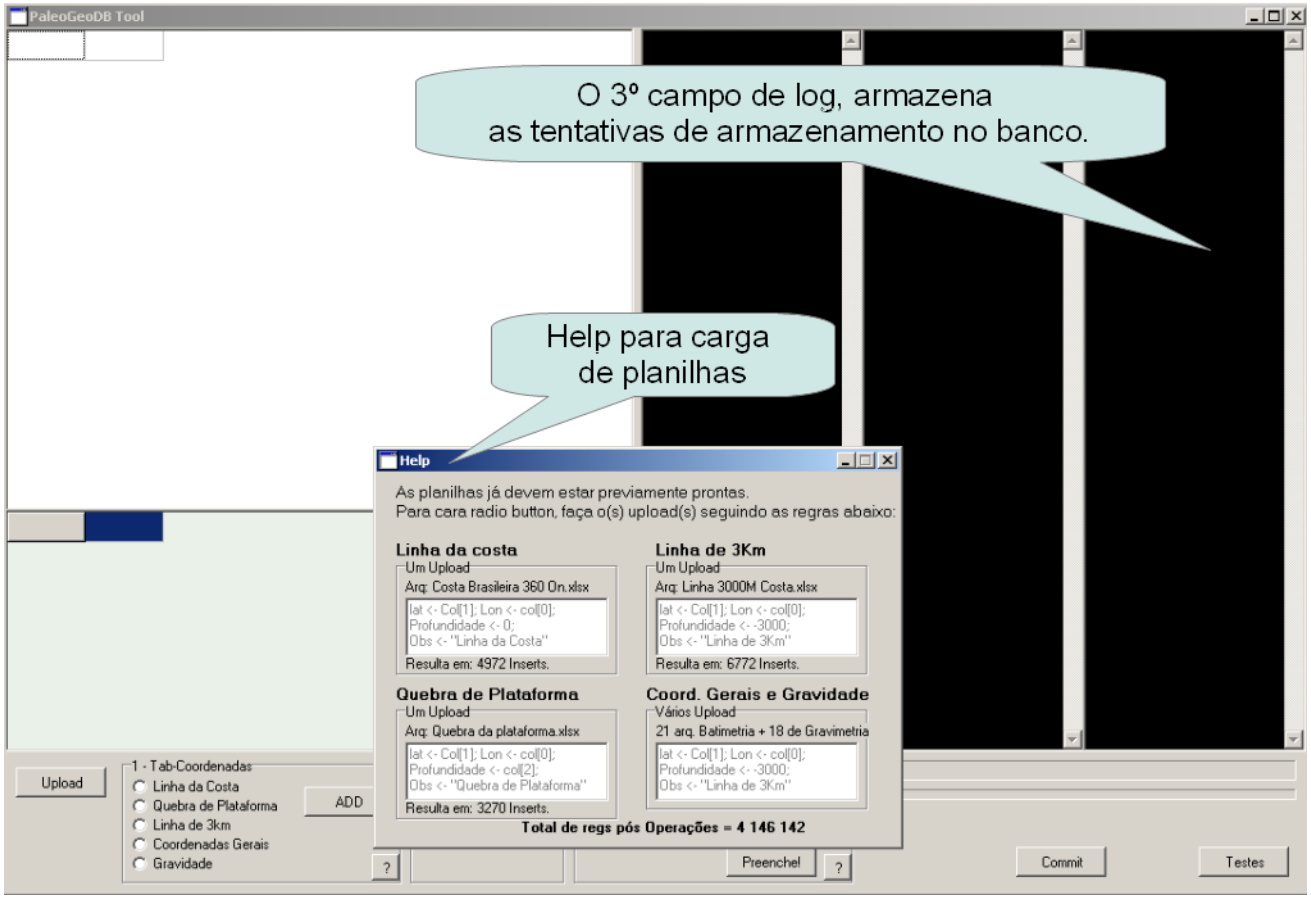
```
insert into EONS ("IDEON","EON","OBS") values(1,'Phanerozoic','The Only after Precambrians');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(1,1,'Cenozoic','');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(2,1,'Mesozoic','');
insert into ERAS ("IDERA","EONS_IDEON","ERA","OBS") values(3,1,'Paleozoic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS")
    values(1,1,'Quaternary','Under discution. International Stratigraphic Chart 2008');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(2,1,'Neogene','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(3,1,'Paleogene','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(4,2,'Cretaceous','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(5,2,'Jurassic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(6,2,'Triassic','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(7,3,'Permian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(8,3,'Carboniferous','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(9,3,'Devonian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(10,3,'Silurian','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(11,3,'Ordovician','');
insert into PERIODOS ("IDPERIODO","ERAS_IDERA","PERIODO","OBS") values(12,3,'Cambrian','');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(1,1,'Holocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA","OBS") values(2,1,'Pleistocene','Used by us like on Age');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(3,2,'Pilocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(4,2,'Miocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(5,3,'Oligocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(6,3,'Eocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(7,3,'Paleocene');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(8,4,'Upper');
insert into epocas ("IDEPOCA","PERIODOS_IDPERIODO","EPOCA") values(9,4,'Lower');
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(1,2,'Upper','',0.126,0.0117);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(2,2,'Ionian','',0.781,0.126);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(3,2,'Calabrian','',1.806,0.781);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(4,2,'Gelasian','Begin of Epoch 2:Plaistocene',2.588,1.806);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(5,3,'Piacenzian','By us, second age, first is full pleistocene',3.600,2.588);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(6,3,'Zancleanan','Begin of Epoch 3:Pilocene',5.332,3.600);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(7,4,'Messinian','',7.246,5.332);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(8,4,'Tortonian','',11.608,7.246);
insert into IDADES ("IDIDADES","EPOCAS_IDEPOCA","IDADE","OBS","MA_INI","MA_FIM")
    values(9,4,'Serravallian','',13.82,11.608);
```

```
insert into ROCHAS ("IDROCHAS","VALOR","OBS")
    values(0,0,'Sem depósitos para o local na idade geológica');
insert into ROCHAS ("IDROCHAS","VALOR","OBS")
    values(1,1,'Depósito parcial');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(2,256,'Argilito','Sedmentar', 'Clástica', 'Terrígena','Siliciclástica','Argila');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(3,8,'Folhelho','Sedmentar', 'Clástica', 'Terrígena','Siliciclástica','Argila+Silte');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(4,4,'Siltito','Sedmentar', 'Clástica', 'Terrígena','Siliciclástica','Silte');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(5,2,'Arenito','Sedmentar', 'Clástica', 'Terrígena','Siliciclástica','Areia');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(6,128,'Diamictito','Sedmentar', 'Clástica', 'Terrígena','Siliciclástica','Cascalho com + de 15% de matriz fina');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(7,2048,'Conglomerado','Sedmentar', 'Clástica', 'Terrígena','Siliciclástica','Cascalho com - de 15% de matriz fina');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(8,4096,'Marga','Sedmentar', 'Clástica', 'Terrígena/Calcária','Siliciclástica/Carbonática','Contém entre 35 e 50% de argila');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(9,32,'Calcilutito','Sedmentar', 'Não-Clástica', 'Química/Biogênica','Calcário','Argila');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(10,64,'Calcissiltito','Sedmentar', 'Não-Clástica', 'Química/Biogênica','Calcário','Silte');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(11,16,'Calcarenito','Sedmentar', 'Não-Clástica', 'Química/Biogênica','Calcário','Areia');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(12,512,'Calcirrudito','Sedmentar', 'Não-Clástica', 'Química/Biogênica','Calcário','Cascalho');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO", "GRANULOMETRIA")
    values(13,131072,'Coquina','Sedmentar', 'Não-Clástica', 'Química/Biogênica','Calcário','Conchas');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO")
    values(14,65536,'Dolomito','Sedmentar', 'Não-Clástica', 'Química/Biogênica','Calcário');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO")
    values(15,16384,'Silexito','Sedmentar', 'Não-Clástica', 'Química/Biogênica','Silica');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO")
    values(16,8192,'Andrita/Gipsita','Sedmentar', 'Não-Clástica', 'Evaporítico','Sulfato de Cálcio');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO","ORIGEM","COMPOSICAO", "SUBCOMPOSICAO")
    values(17,1024,'Halita','Sedmentar', 'Não-Clástica', 'Evaporítico','Cloreto de Sódio');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
```

```
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(20,32768,'Basalto','Ígnea');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(21,262144,'Diabásio','Ígnea');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(22,2097152,'Ígnea Ácida','Ígnea');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(23,4194304,'Ígnea Alcalina','Ígnea');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(24,1048576,'Ígnea não especificada','Ígnea');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(25,8388608,'Crosta Oceânica','Ígnea');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(26,16777216,'SRD','Ígnea');
insert into ROCHAS ("IDROCHAS","VALOR","LITOLOGIA","TIPO")
    values(27,524288,'Metamórfica não especificada','Metamórfica');
```

Apêndice E

Ferramenta de ETL (PaleoGeoDB Tool)


PaleoGeoDB Tool
O 3º campo de log, armazena
as tentativas de armazenamento no banco.
Help para carga
de planilhas
Help
As planilhas já devem estar previamente prontas.
Para cara radio button, faça o(s) upload(s) seguindo as regras abaixo:
Linha da costa
Um Upload
Arq: Costa Brasileira 360 On.xlsx
lat <- Col[1]; Lon <- col[0];
Profundidade <- 0;
Obs <- "Linha da Costa"
Resulta em: 4972 Inserts.
Linha de 3Km
Um Upload
Arq: Linha 3000M Costa.xlsx
lat <- Col[1]; Lon <- col[0];
Profundidade <- -3000;
Obs <- "Linha de 3Km"
Resulta em: 6772 Inserts.
Quebra de Plataforma
Um Upload
Arq: Quebra da plataforma.xlsx
lat <- Col[1]; Lon <- col[0];
Profundidade <- col[2];
Obs <- "Quebra de Plataforma"
Resulta em: 3270 Inserts.
Coord. Gerais e Gravidade
Vários Upload
21 arq. Batimetria + 18 de Gravimetria
lat <- Col[1]; Lon <- col[0];
Profundidade <- -3000;
Obs <- "Linha de 3Km"
Total de regs pós Operações = 4 146 142
Upload
1 - Tab-Coordenadas
Linha da Costa
Quebra de Plataforma
Linha de 3km
Coordenadas Gerais
Gravidade
ADD
Preenche!
Commit
Testes

Apêndice F

Resultados obtidos pelo algoritmo criado

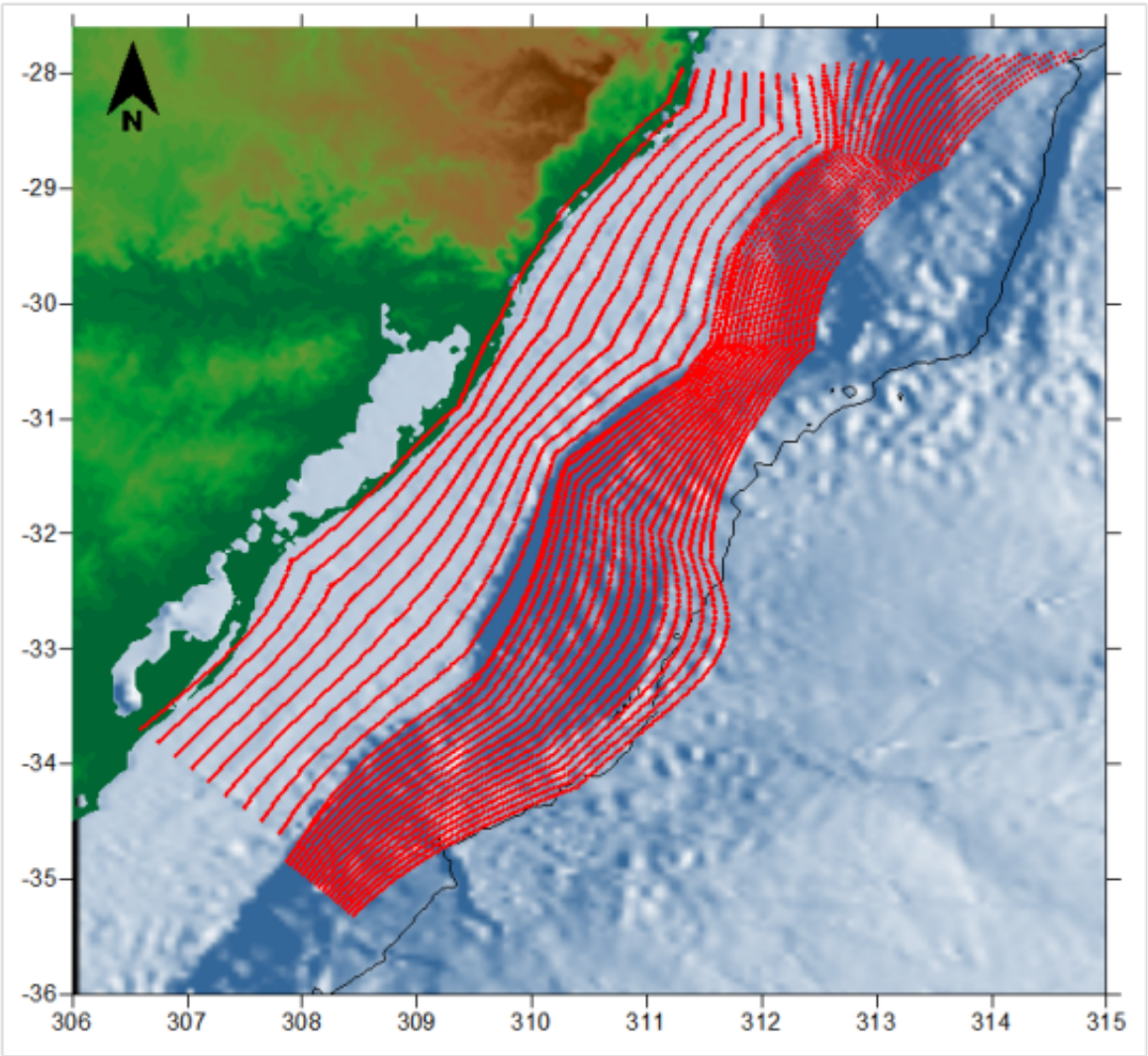


Figura F.1: Gráfico da posição dos dados na bacia de Pelotas

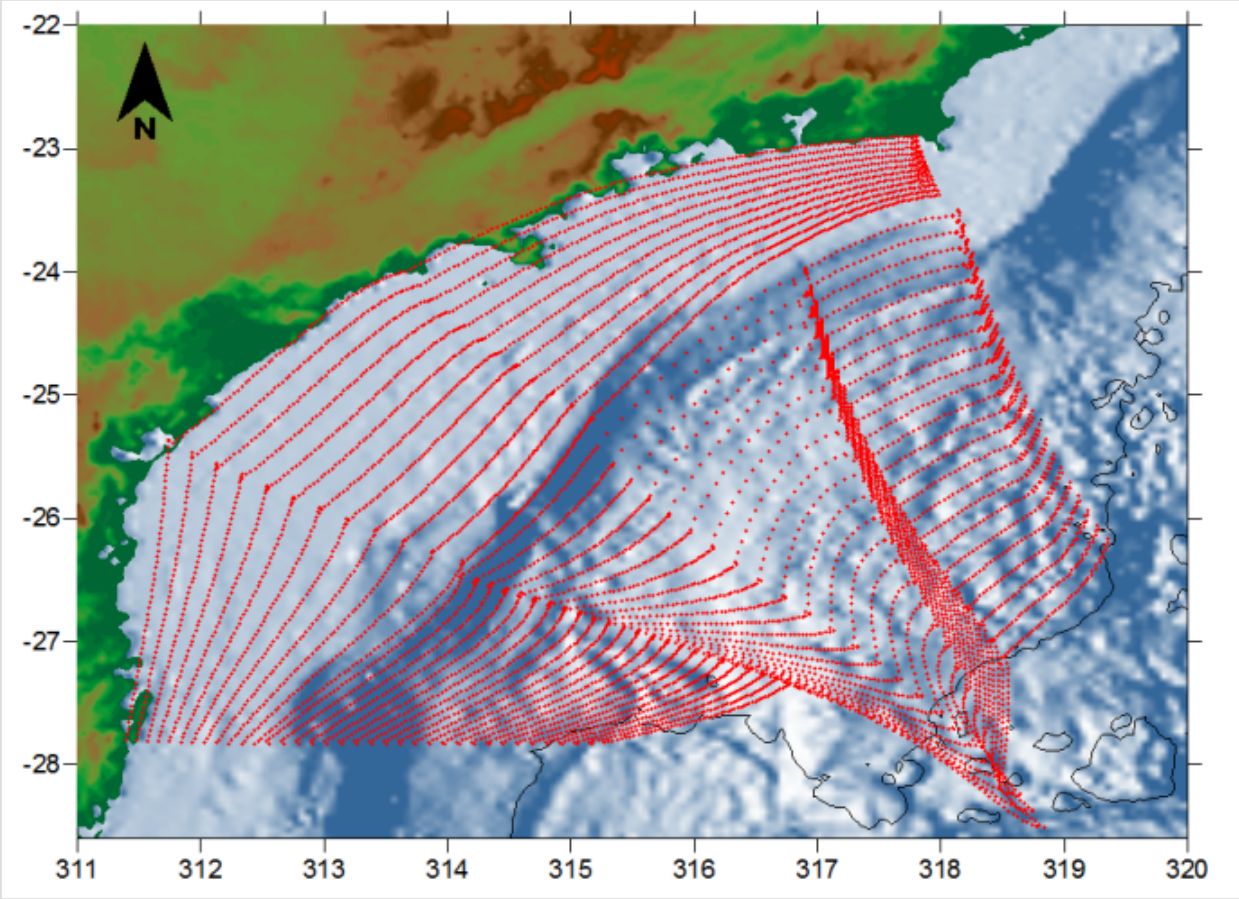


Figura F.2: Gráfico da posição dos dados na bacia de Santos

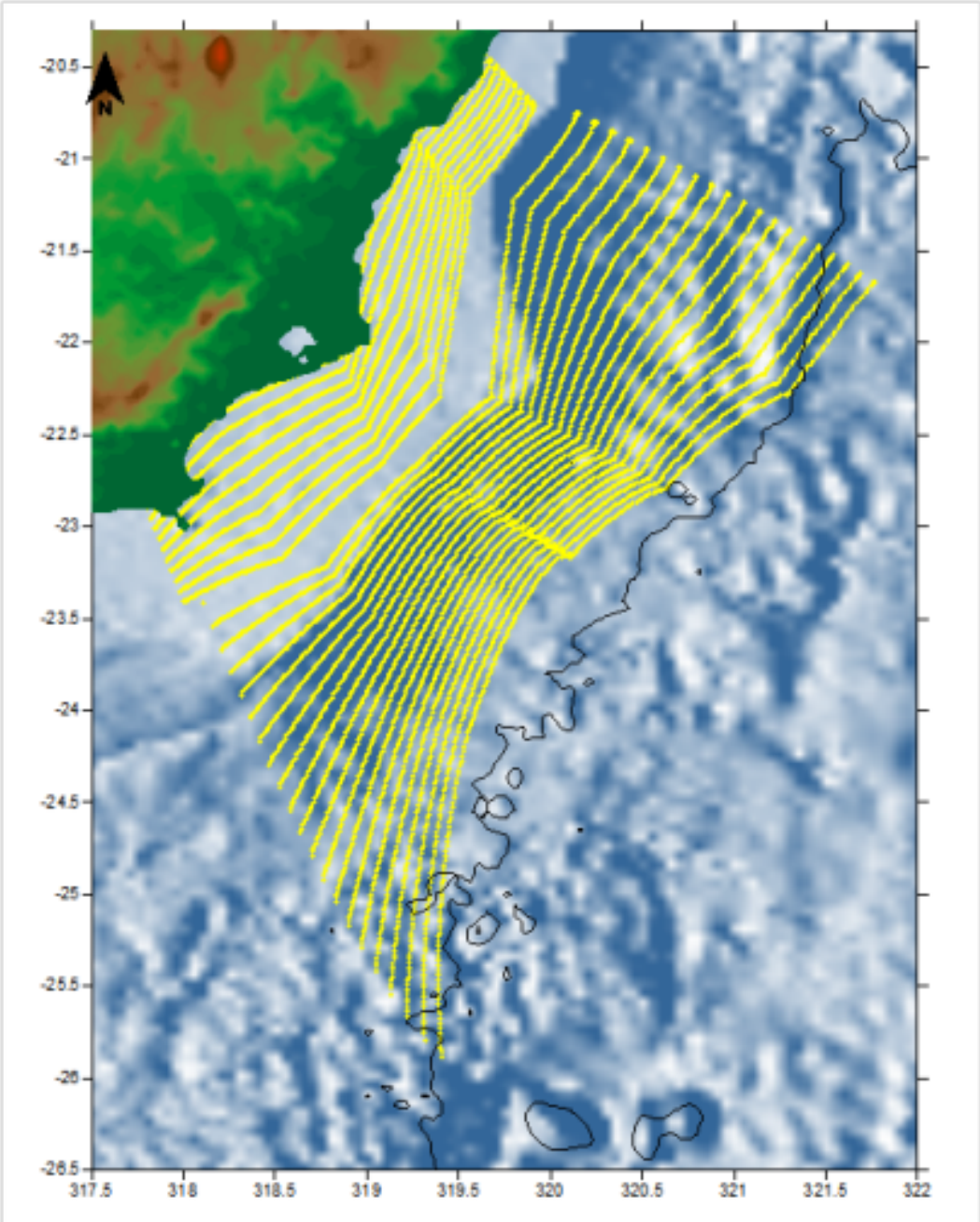


Figura F.3: Gráfico da posição dos dados na bacia de Campos.

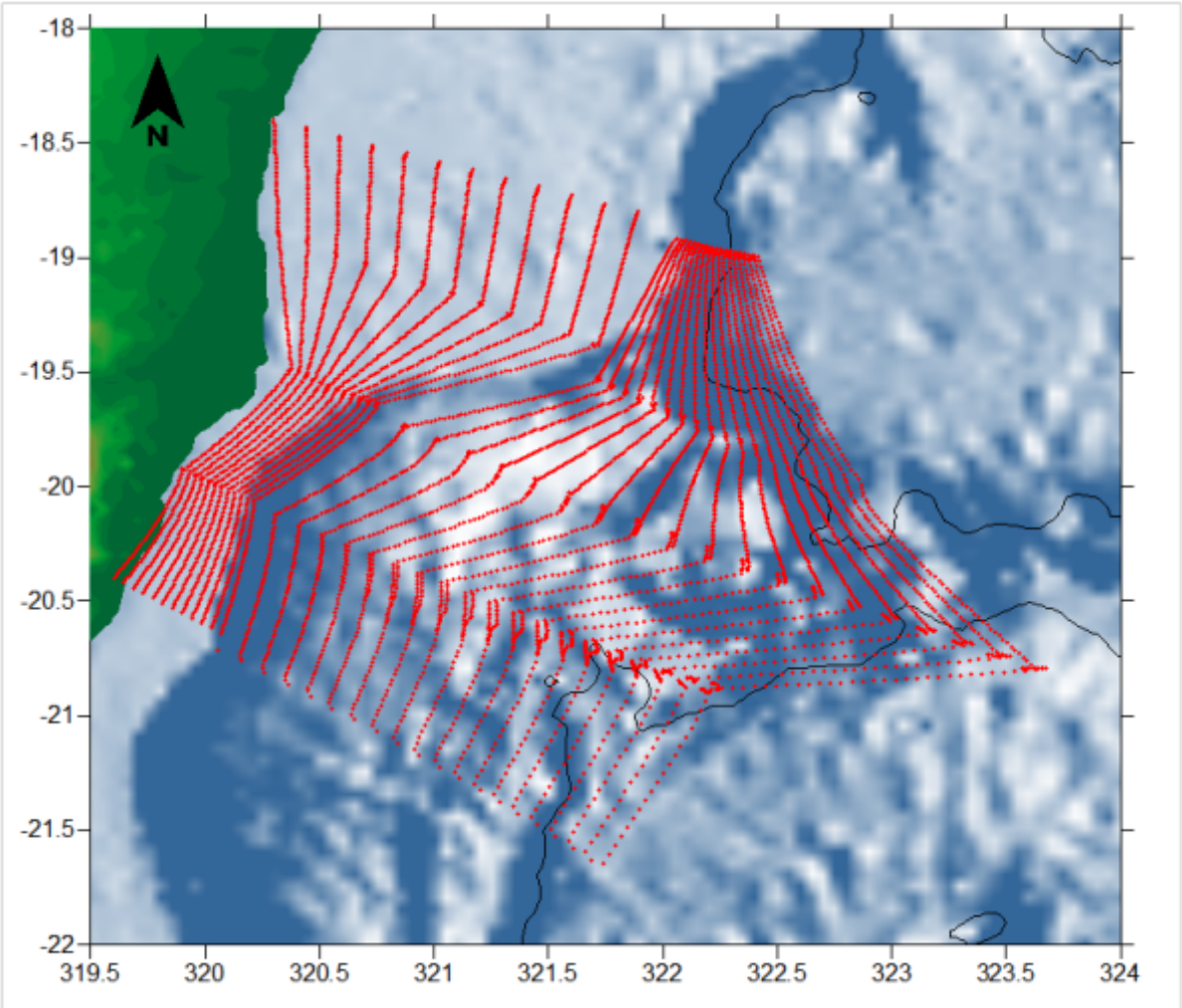


Figura F.4: Gráfico da posição dos dados na bacia de Espírito Santo.

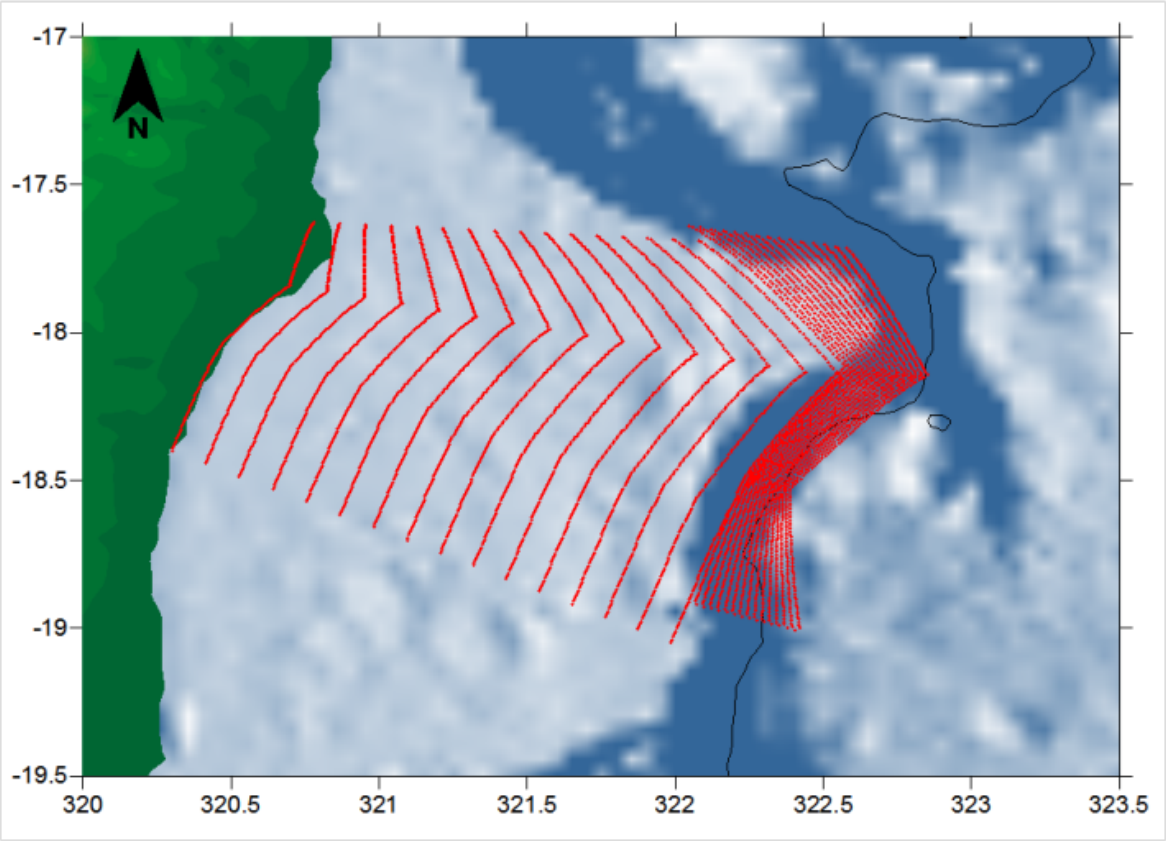


Figura F.5: Gráfico da posição dos dados na bacia de Mucuri.

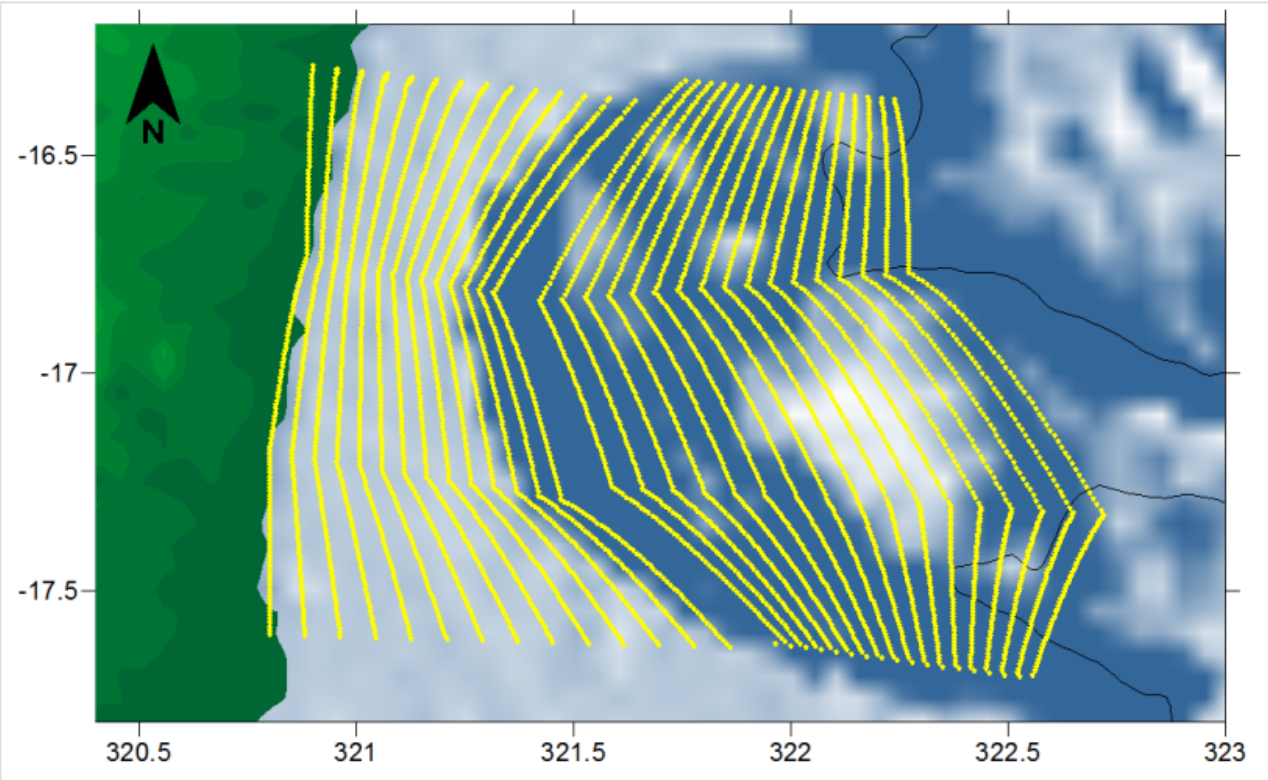


Figura F.6: Gráfico da posição dos dados na bacia de Cumuruxatiba.

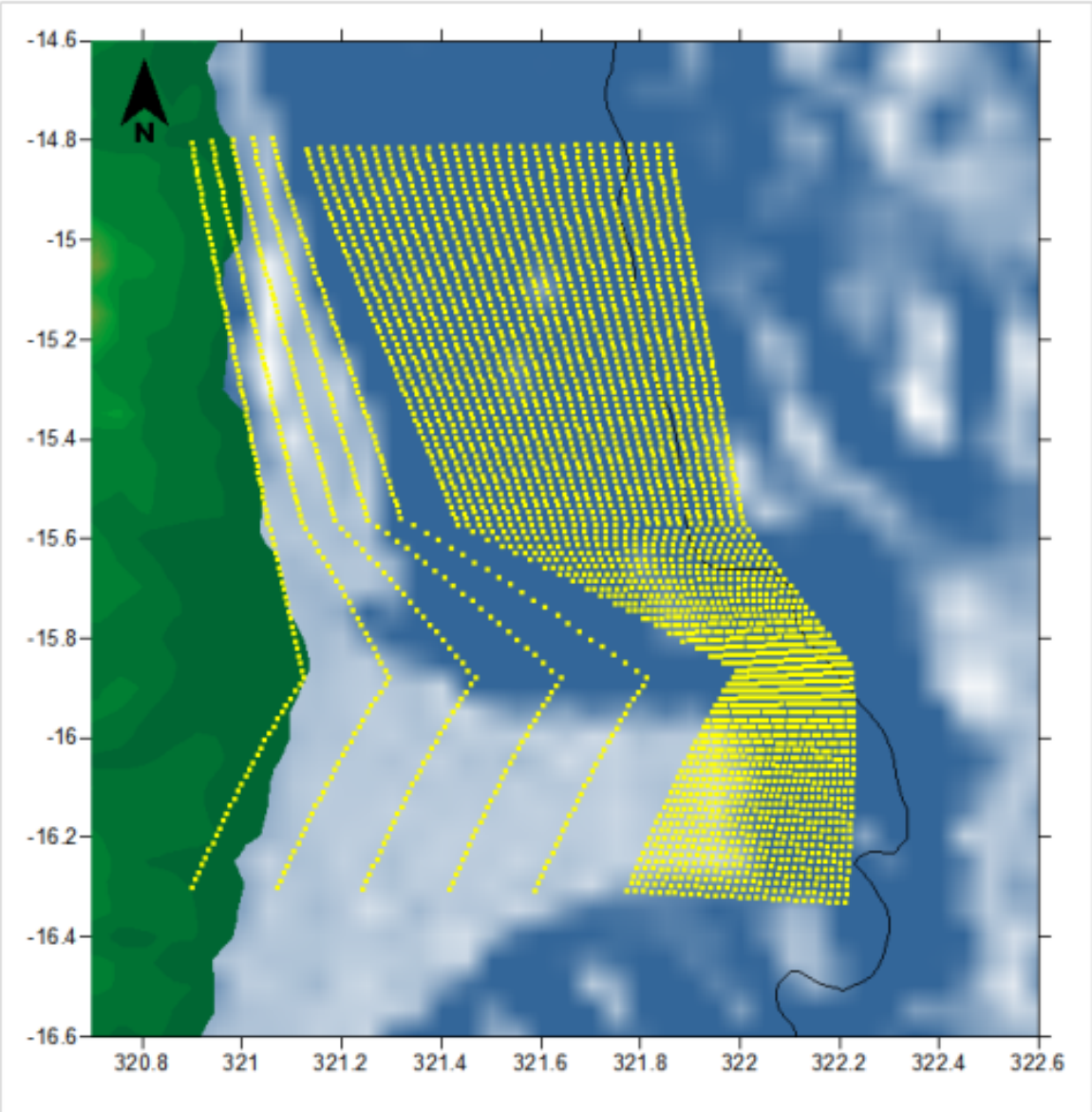


Figura F.7: Gráfico da posição dos dados na bacia de Jequitinhonha.

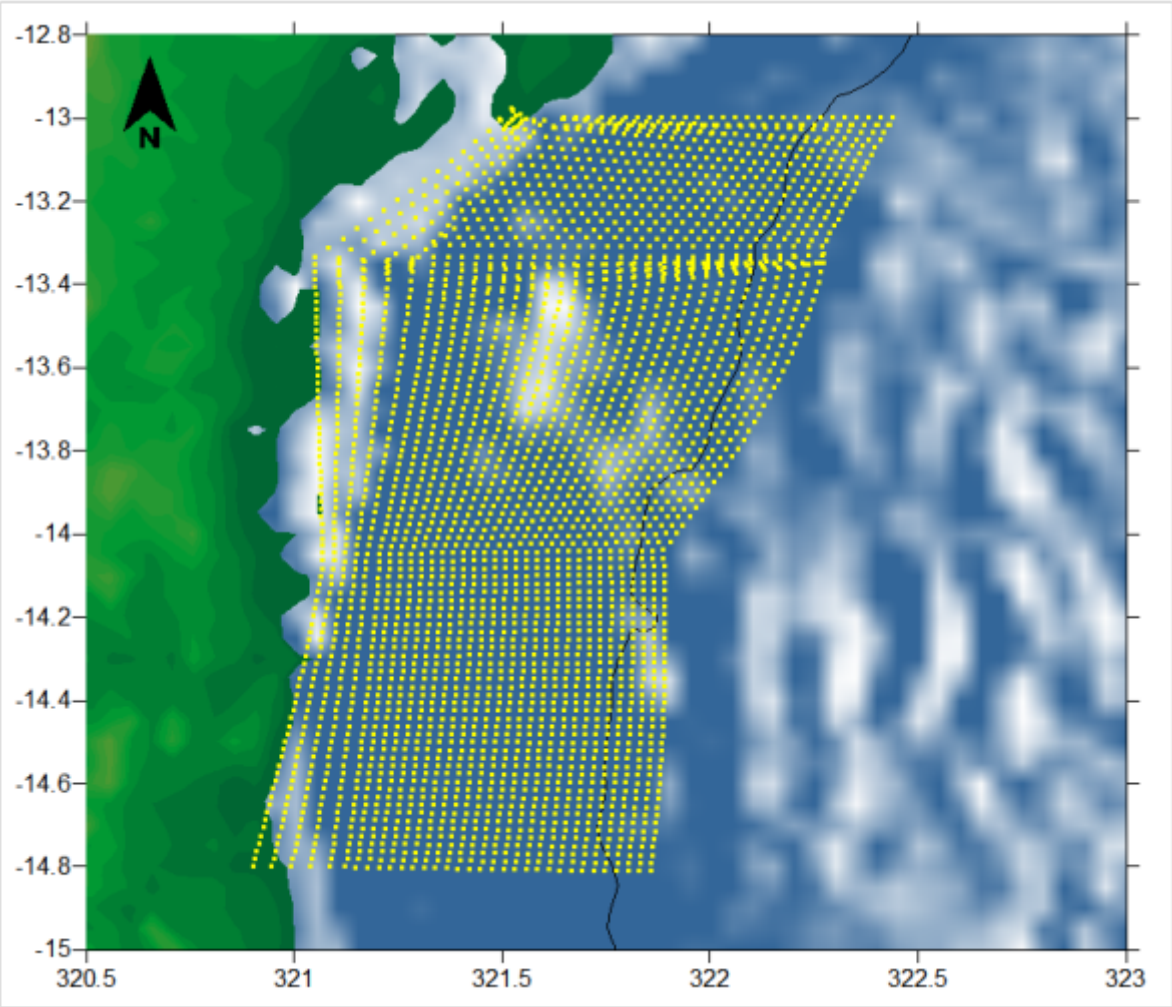


Figura F.8: Gráfico da posição dos dados na bacia de Camamu-Alamada.

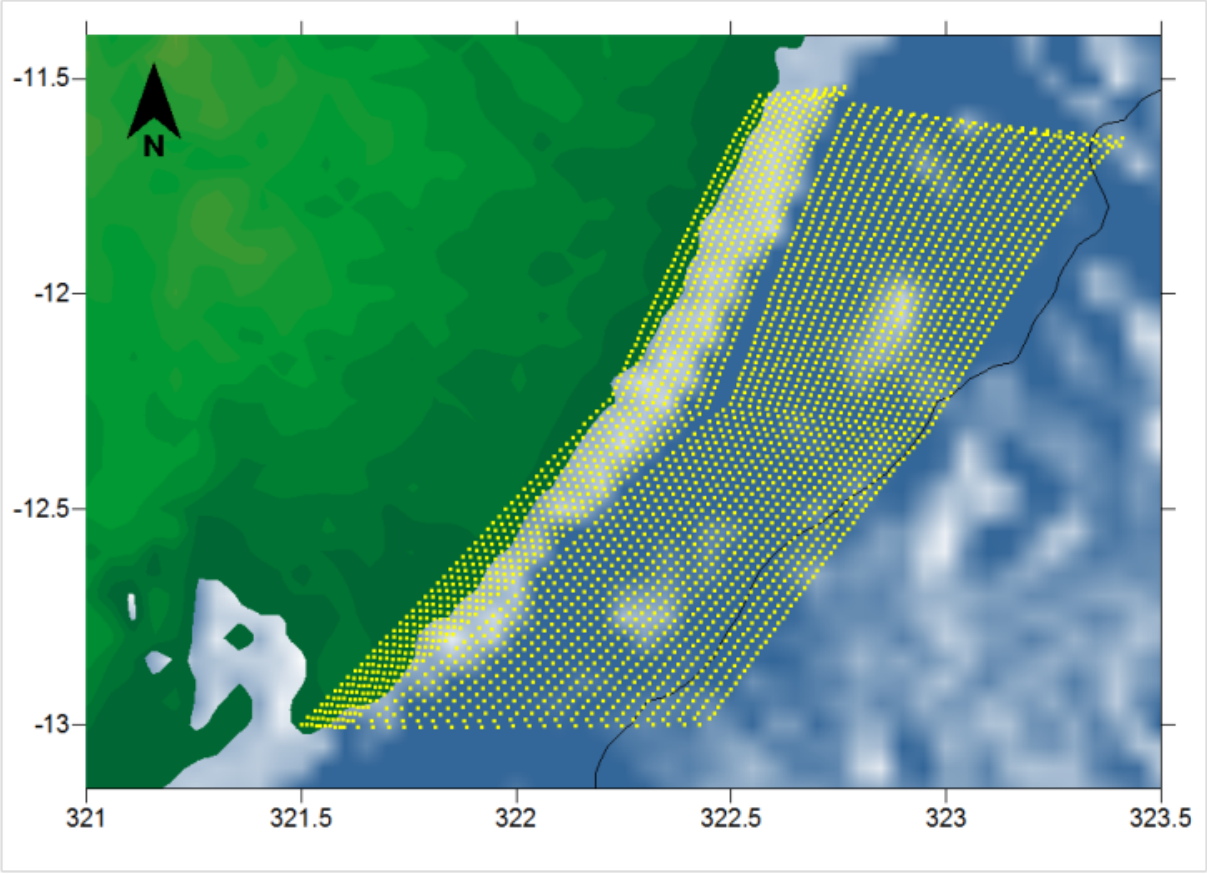


Figura F.9: Gráfico da posição dos dados na bacia de Jacuípe.

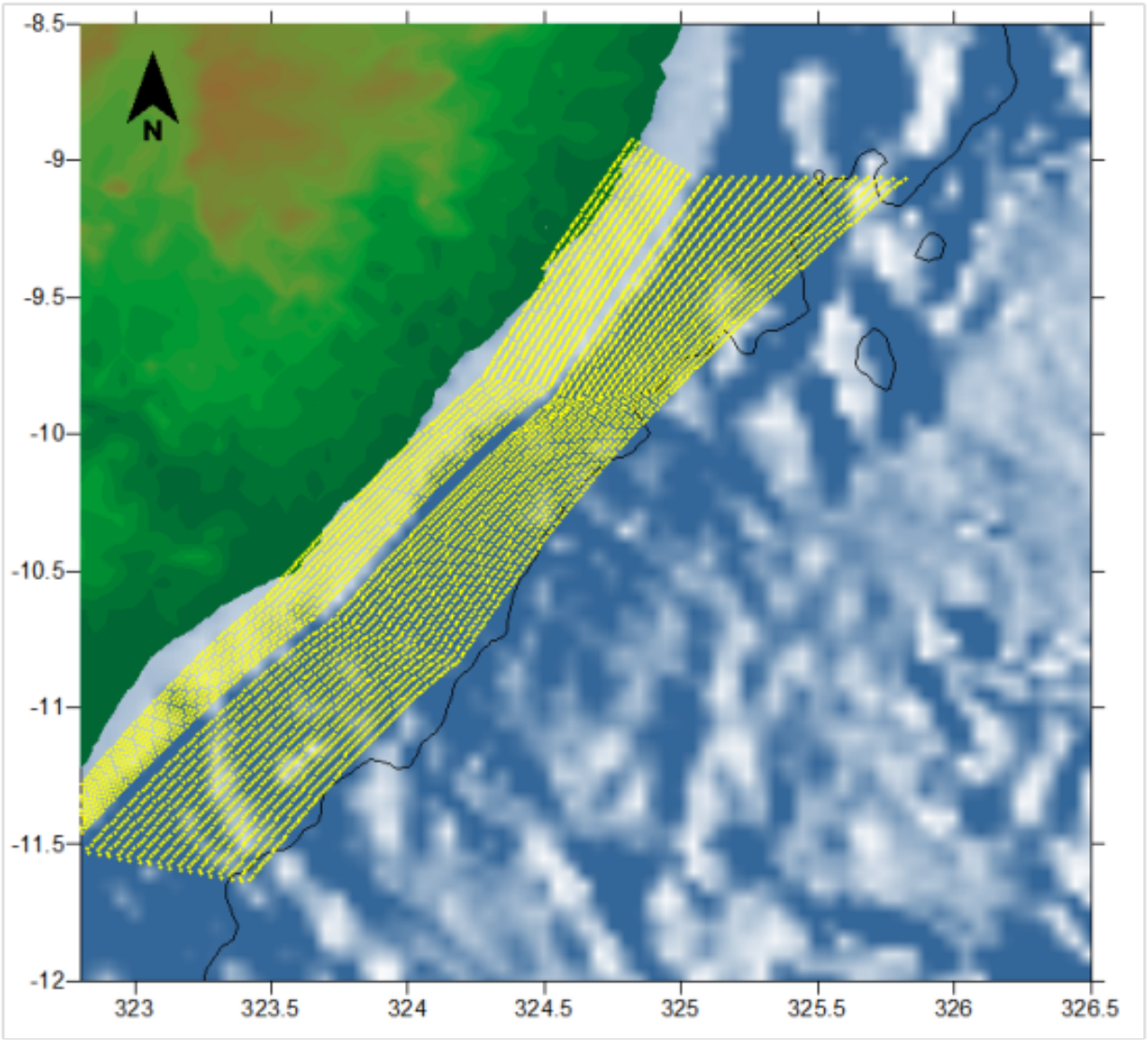


Figura F.10: Gráfico da posição dos dados na bacia de Sergipe-Alagoas.

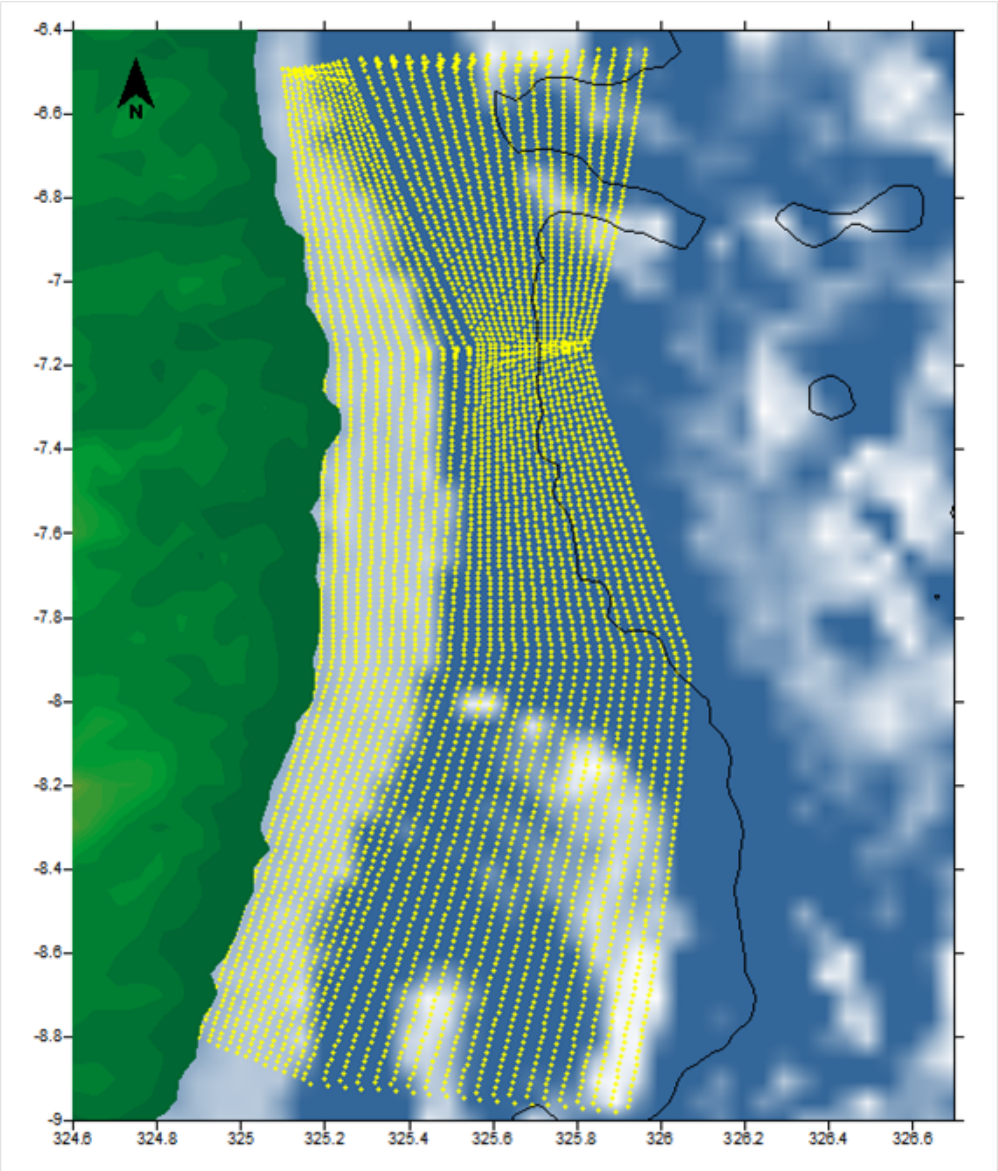


Figura F.11: Gráfico da posição dos dados na bacia de Pernambuco-Paraíba.

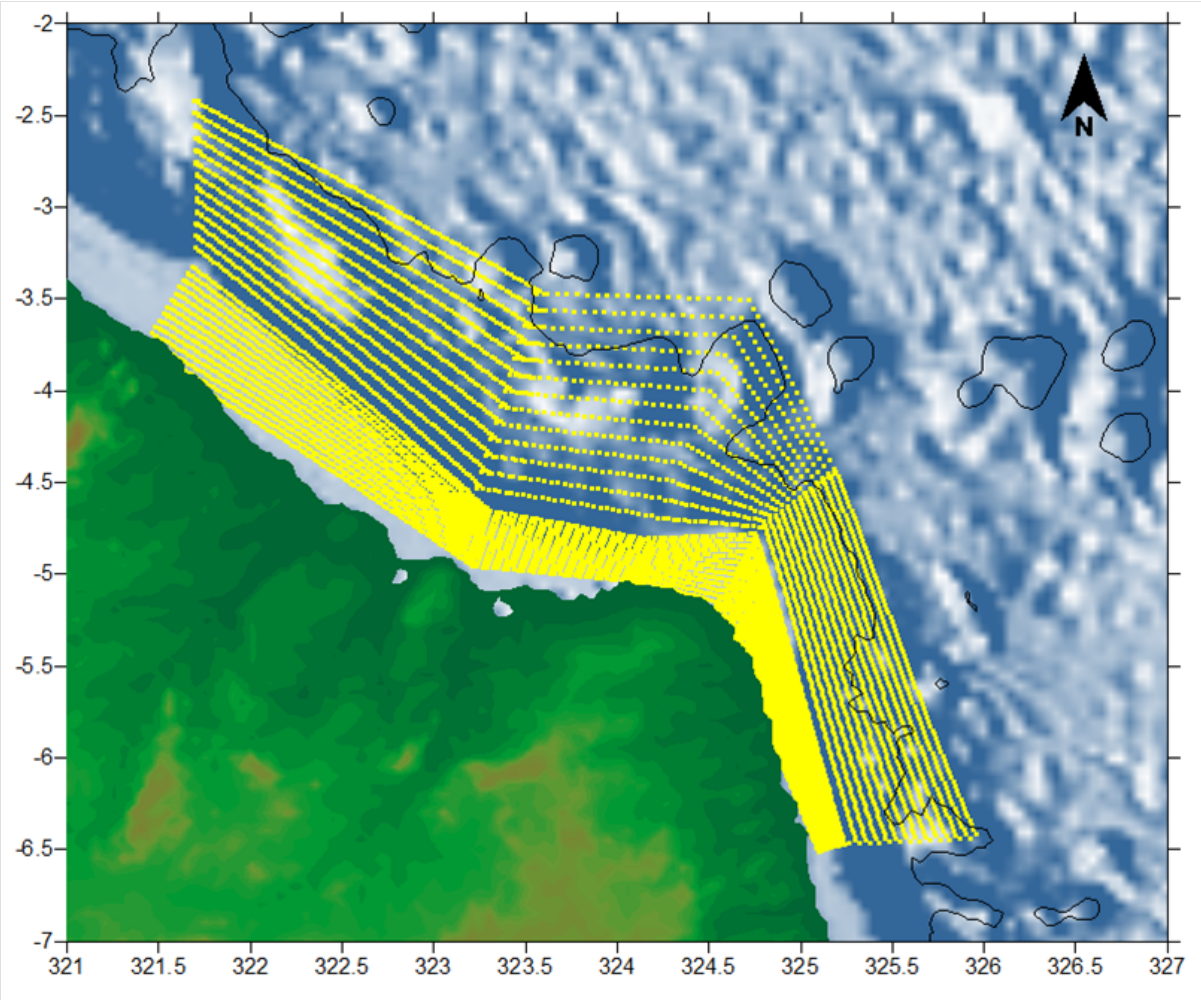


Figura F.12: Gráfico da posição dos dados na bacia de Potiguar.

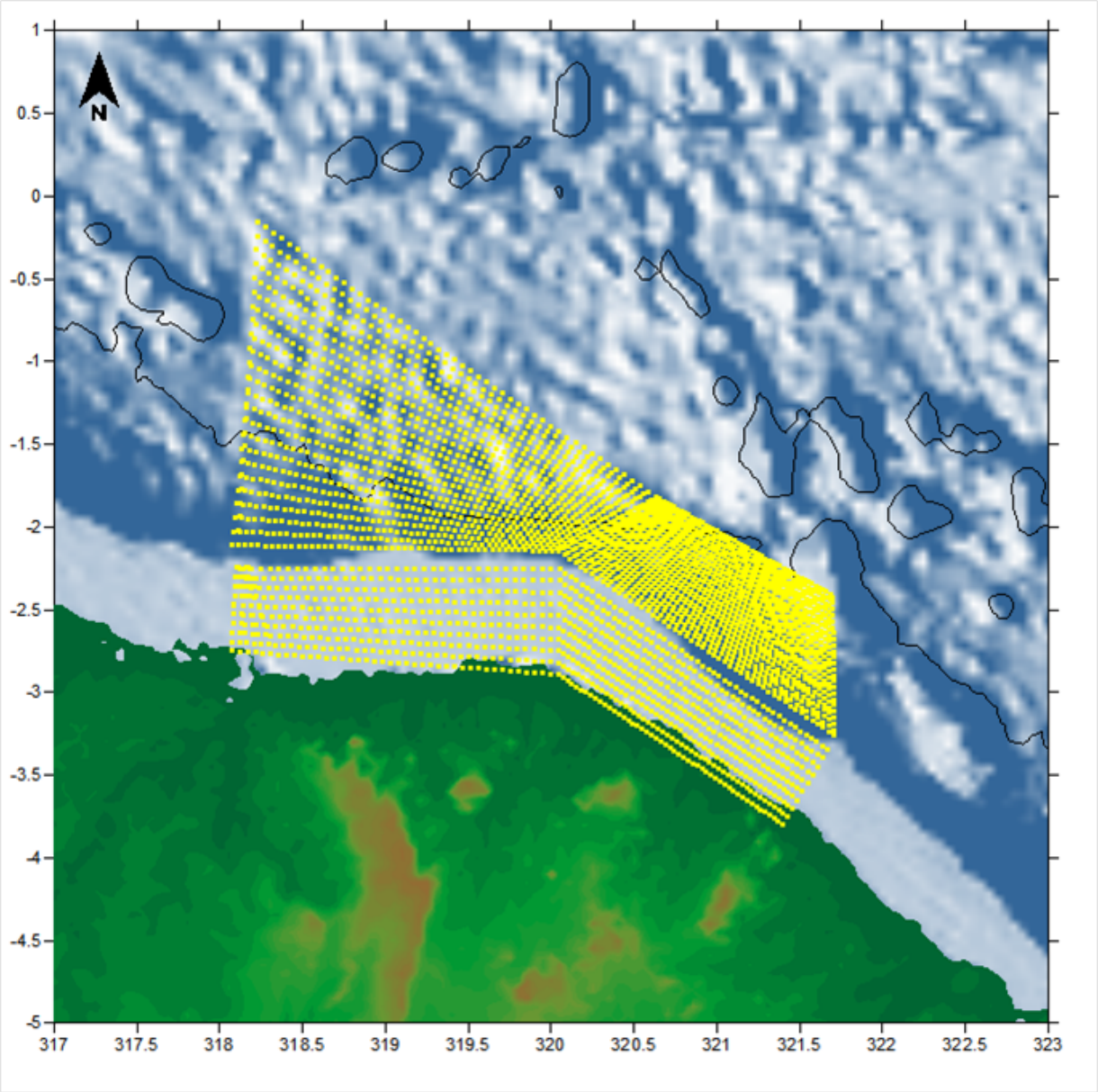


Figura F.13: Gráfico da posição dos dados na bacia de Ceara.

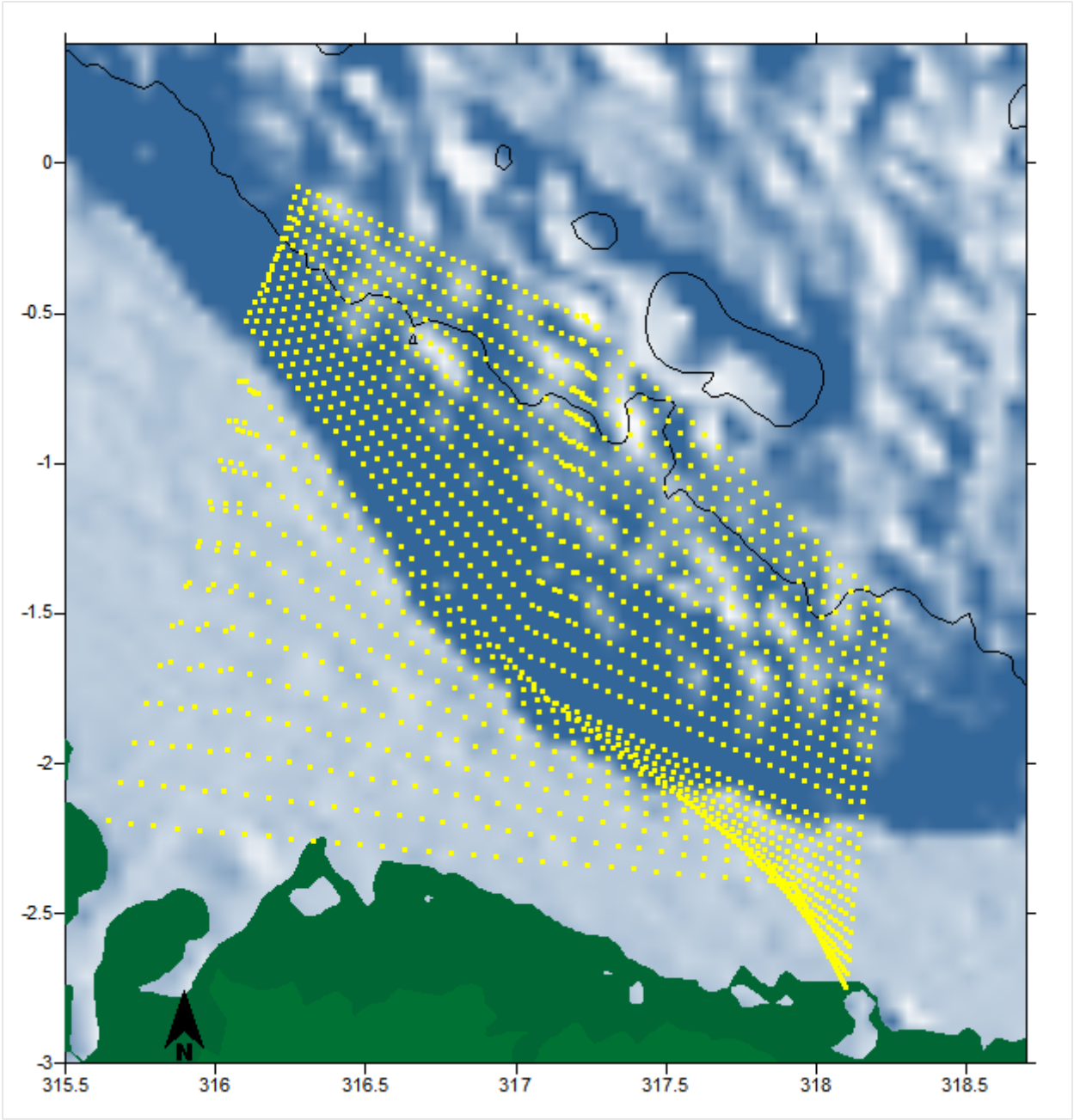


Figura F.14: Gráfico da posição dos dados na bacia de Barreirinhas.

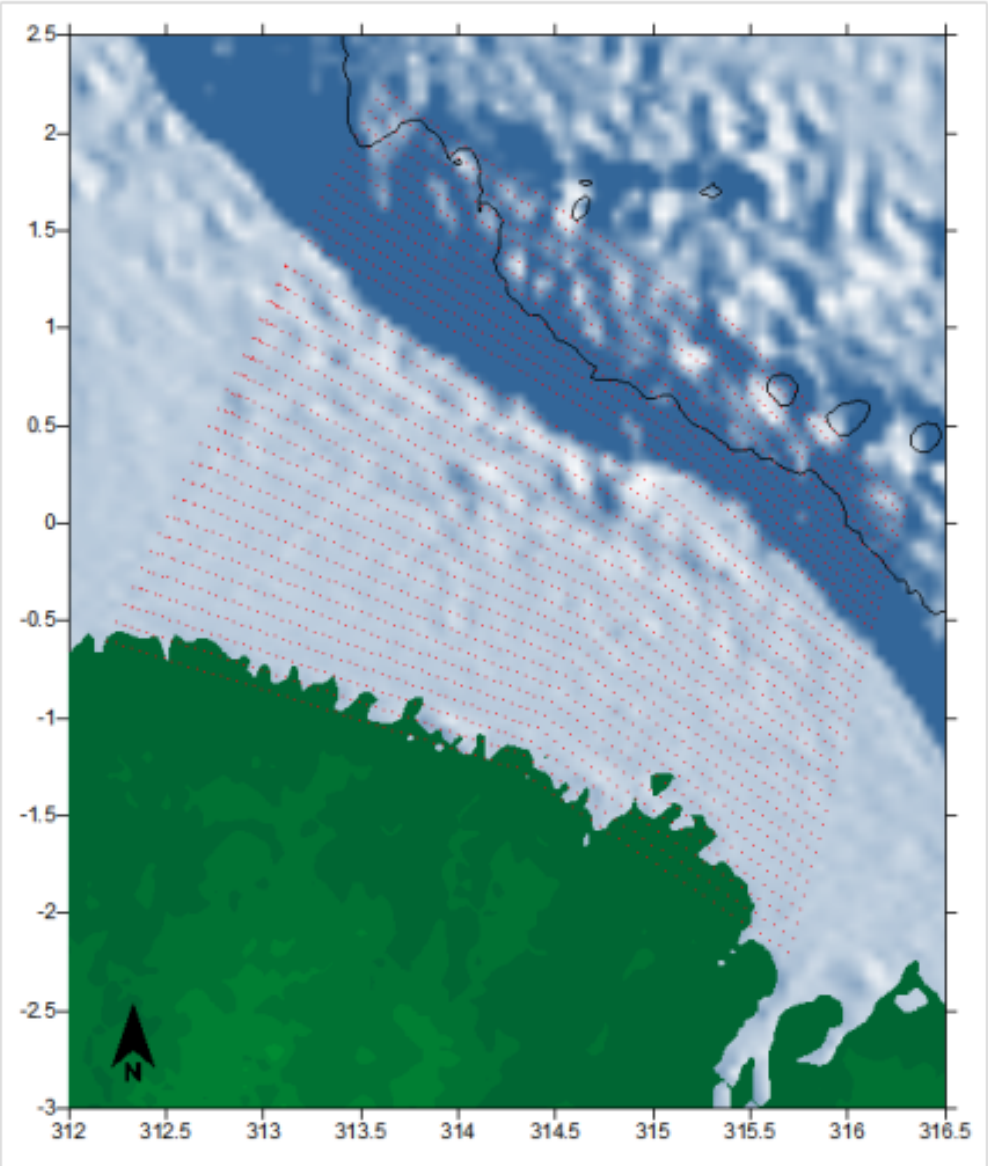


Figura F.15: Gráfico da posição dos dados na bacia de Pará-Maranhão.

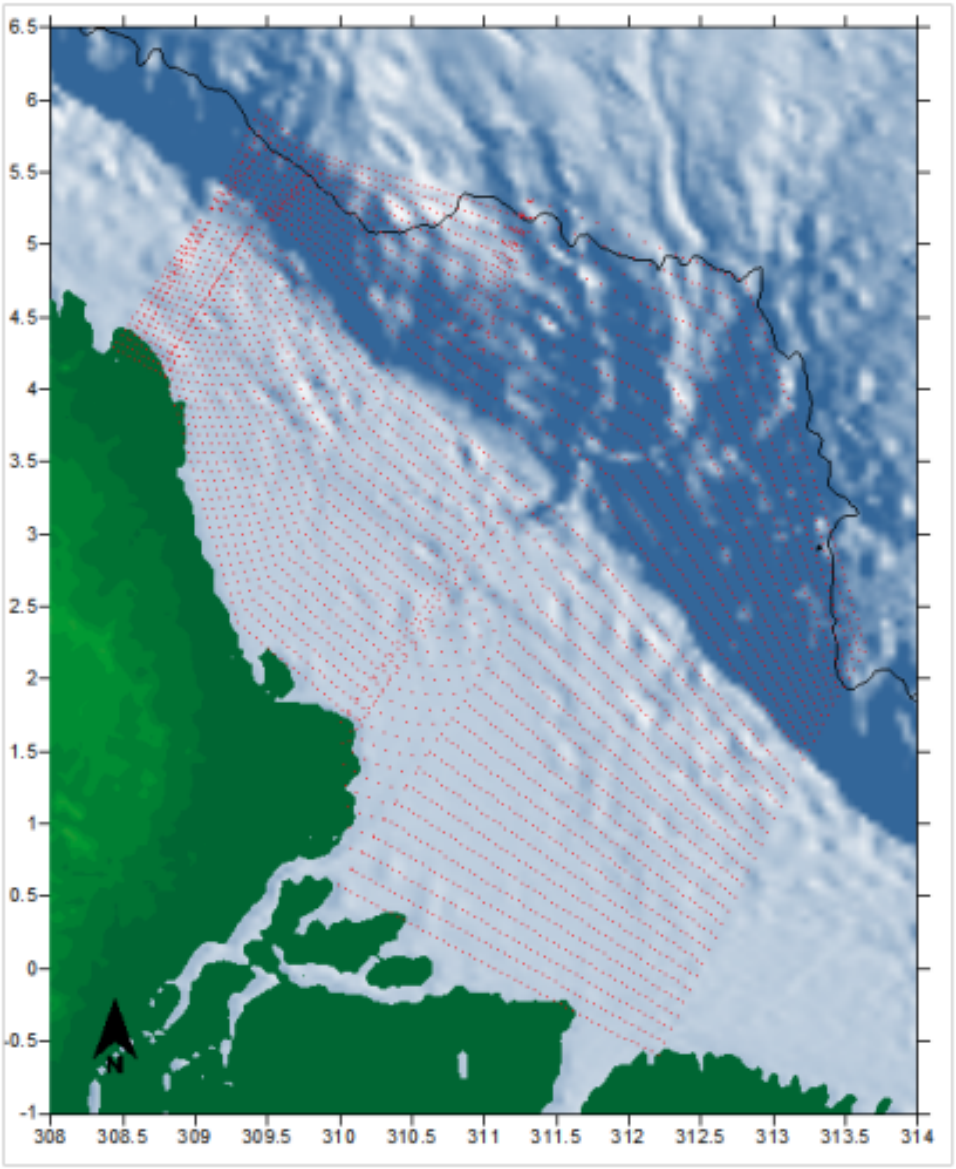


Figura F.16: Gráfico da posição dos dados na bacia de Foz do Amazonas.