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

UMA ABORDAGEM PARA MINERAÇÃO DE DADOS E VISUALIZAÇÃO DE RESULTADOS EM IMAGENS BATIMÉTRICAS

LUIS FERNANDO PLANELLA GONZALEZ

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: Duncan Dubugras Alcoba Ruiz

Porto Alegre
2012

G643a Gonzalez, Luis Fernando Planella
Uma abordagem para mineração de dados e visualização de resultados em imagens batimétricas / Luis Fernando Planella Gonzalez. – Porto Alegre, 2012.
73 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.

1. Informática. 2. Mineração de Dados (Informática). 3. Processamento de Imagens. I. Ruiz, Duncan Dubugras Alcoba. 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 "Uma Abordagem para Mineração de Dados e Visualização de Resultados em Imagens Batimétricas", apresentada por Luis Fernando Planella Gonzalez como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Bioinformática e Modelagem Computacional, aprovada em 27/03/2012 pela Comissão Examinadora:

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

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

Prof. Dr. João Marcelo Medina Ketzer - — CEPAC/PUCRS

Homologada em..08../..06../..2012.., conforme Ata No. ..012... 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

Dedico este trabalho a Deus, eterno pai de amor

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus. És força aos que Te amam e provedor dos que Te buscam.

À minha esposa, Júlia, obrigado por todo o amor e apoio que tens me dedicado, e pela paciência nos momentos que não pude te ajudar com a Gabi.

À minha filha, Gabriela, obrigado por entender quando o papai não pôde estar contigo durante toda esta trajetória.

Aos meus pais, Velarde e Tereza, obrigado pelo incentivo, pelo amor dedicado e pela formação que me permitiram alcançar.

Ao meu orientador, Duncan, obrigado pelos ensinamentos e por acreditar neste trabalho (por vezes mais do que eu).

À Maria Alejandra Gómez Pivel, obrigado por ser a especialista de domínio que acompanhou este trabalho de perto, e prestou seu depoimento, constante na Seção 7.1.4.

Ao pessoal da InStroDI e da STRO, obrigado pelo incentivo prestado, incluindo liberação do trabalho durante as aulas e em alguns outros momentos "críticos".

Ao PPGCC, obrigado por proporcionar a bolsa TAXAS, que financiou meus estudos.

Finalmente, agradeço a todos os que não mencionei, mas que participaram desta trajetória.

UMA ABORDAGEM PARA MINERAÇÃO DE DADOS E VISUALIZAÇÃO DE RESULTADOS EM IMAGENS BATIMÉTRICAS

RESUMO

A batimetria é a medida da profundidade em distintos lugares de uma massa de água, e também a informação derivada de tais medições. Possui diversas aplicações importantes e tem atraído cada vez mais interesse nos últimos anos. Mapas batimétricos podem cobrir toda a extensão do globo terrestre. Entretanto, a análise apenas por inspeção visual destes mapas pode ser difícil, devido a variações sutis na conformação do solo oceânico. Assim, seria interessante a disponibilização de ferramentas computacionais capazes de auxiliar ao especialista de domínio nos mais diversos problemas relativos a imagens batimétricas, analisando-as de forma automática ou semi-automática. A contribuição deste trabalho é uma abordagem para a utilização da mineração de dados para tal análise, e de uma iconografia para a visualização dos resultados da mineração e de características do próprio mapa. São propostas técnicas para o processamento da imagem de entrada, a fim de extrair da mesma registros e atributos que possam ser processados por algoritmos clássicos da mineração de dados. Também é proposta uma iconografia para a visualização dos resultados do processo de descoberta de conhecimento e das características de áreas processadas do mapa. Finalmente a abordagem proposta é testada, aplicando-a sobre uma base de dados real, com supervisão de um especialista de domínio.

Palavras-chave: Mineração de dados; Batimetria; Processamento de imagens; Visualização de informações.

AN APPROACH FOR DATA MINING AND RESULTS VISUALIZATION IN BATHYMETRY IMAGES

ABSTRACT

Bathymetry is the measurement of the depth at various places in a body of water, as well as information derived from such measurements. It has several important applications, and has been attracting increasing interest over the last years. Bathymetry maps may cover the entire extent of the Earth globe. However, the analysis of such maps by visual inspection solely is difficult, due to subtle variations on the seafloor conformation. Thus, it would be interesting to have available computational tools capable of assisting a domain expert in problems related to bathymetry images, by analyzing them automatically or semi-automatically. The contribution of this work is an approach to use data mining for such analysis, and an iconography for results visualization, as well as map characteristics. We propose techniques to process input images, in order to extract records and their features, which can be processed by classic data mining algorithms. We also propose an iconography for visualization of knowledge discovery process results, as well as characteristics of areas in the processed map. Finally, the proposed approach is tested by applying it on a real database, under a domain expert supervision.

Keywords: Data mining; Bathymetry; Image processing; Information visualization.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API	*Application Programming Interface*
BODC	*British Oceanographic Data Centre*
CART	*Classification And Regression Trees*
DWT	*Discrete Wavelet Transform*
GEBCO	*General Bathymetric Chart of the Oceans*
IDWT	*Inverse Discrete Wavelet Transform*
NASA	*National Aeronautics and Space Administration*
NOAA	*National Oceanic and Atmospheric Administration*
SGBD	*Sistema de Gerenciamento de Bancos de Dados*
SQL	*Structured Query Language*
TPC	*Tabela de Probabilidade Condicional*

SUMÁRIO

1. INTRODUÇÃO

A batimetria é a medida da profundidade em distintos lugares de uma massa de água [24]. Nos últimos anos, a batimetria tem sido cada vez mais um objeto de estudo, pois possui diversas aplicações importantes, como exploração de recursos minerais, auxílio da navegação, estudo de correntes submarinas profundas e a movimentação de sedimentos [37].

Atualmente, os mapas batimétricos são construídos pela combinação de dados coletados a partir de navios (através de sonar de multifeixe, por exemplo) e de satélites [37]. Os dados obtidos por satélites não são tão precisos quanto àqueles coletados por navios, porém cobrem toda a extensão do globo terrestre.

Entretanto, a análise de imagens batimétricas de regiões extensas pode ser bastante trabalhosa e sensível a erros, devido a detalhes sutis, como pequenas variações de profundidades ou conformações no fundo oceânico. Esses detalhes são, por vezes, de difícil percepção por inspeção visual. Assim, torna-se interessante a disponibilização de ferramentas computacionais que possam efetuar essa análise de forma automática ou semi-automática, a fim de auxiliar a um especialista de domínio no entendimento e exploração dos mais diversos problemas relativos à batimetria.

Para a realização dessa análise é proposta a utilização da mineração de dados. A mineração de dados é o processo de descoberta automática de informações úteis em grandes repositórios de dados [34], e dispõe de técnicas bem conhecidas para a execução de tarefas descritivas e preditivas [18]. Tarefas descritivas são capazes de explicar o comportamento de um conjunto de dados, e consistem na análise de agrupamentos, análise de associações e detecção de anomalias. Já as tarefas preditivas permitem, a partir do fornecimento de valores conhecidos em alguns registros, a indução de um modelo capaz de classificar ou prever valores do mesmo tipo em registros previamente desconhecidos. Em especial, a classificação e a análise de agrupamentos são as tarefas exploradas neste trabalho.

O objetivo deste trabalho é propor uma abordagem para a aplicação da mineração de dados diretamente sobre imagens batimétricas, não sobre dados numéricos de batimetria. A mineração de dados em imagens possui um desafio adicional de preparação de dados, tendo em vista que os algoritmos clássicos de mineração de dados trabalham sobre dados tabulares. Para minerar imagens, podem ser desenvolvidos novos algoritmos que trabalhassem sobre este formato nativamente (o que não foi detectado em nenhum trabalho pesquisado) ou pré-processar as imagens para transformá-las no formato esperado pelos algoritmos de mineração. Este pré-processamento pode-se dar através da extração de registros e de atributos diretamente do conteúdo da imagem, valendo-se de técnicas bem conhecidas de processamento de imagens, como estatísticas sobre os valores dos pixels, histogramas de cores [26] e coeficientes de *Wavelets* [22]. Também encontra-se na literatura trabalhos que desenvolveram técnicas próprias para este pré-processamento, visto que o processamento de imagens tende a ser bastante dependente do domínio de problema abordado. Neste trabalho ainda são

propostas novas técnicas para a extração de atributos que capturam a morfologia dominante de áreas do mapa.

Entende-se que apenas a extração e a mineração de dados de imagens não são suficientes para que um especialista de domínio possa efetivamente compreender os achados, visto que tradicionalmente, a apresentação dos resultados se dá também em formato tabular, o que pode tornar confuso seu entendimento. Assim, é proposta uma iconografia para a representação dos resultados da mineração, bem como de características do próprio mapa. Uma iconografia intuitiva e consistente pode auxiliar bastante o especialista de domínio na compreensão dos achados e das relações entre as diferentes áreas do mapa.

Finalmente, a abordagem proposta neste trabalho foi aplicada sobre uma base de dados real, de corais de águas profundas [32], e validada junto a um especialista de domínio (oceanógrafa), que emitiu seu parecer sobre os resultados obtidos através da mineração de dados.

Este documento encontra-se estruturado da seguinte forma: O Capítulo 2 apresenta o referencial teórico necessário para o entendimento dos conceitos envolvidos neste trabalho, o Capítulo 3 apresenta os objetivos deste trabalho, bem como os trabalhos relacionados, o Capítulo 4 descreve a abordagem desenvolvida neste trabalho para a aplicação da mineração de dados em imagens batimétricas, o Capítulo 5 descreve a forma proposta para a visualização do processo de mineração de dados e de seus resultados, o Capítulo 6 descreve o desenvolvimento de um protótipo do ambiente de software para poder testar a abordagem proposta, o Capítulo 7 detalha como foi validada a abordagem proposta, aplicando-a sobre uma base de dados real, sob supervisão de um especialista de domínio, e o Capítulo 8 apresenta as considerações finais sobre este trabalho e as possibilidades de pesquisas futuras.

2. REFERENCIAL TEÓRICO

2.1 Batimetria

A batimetria é a medição da profundidade dos oceanos, lagos e rios [6]. É também a informação derivada de tais medições [24]. O primeiro mapa batimétrico com cobertura de toda a extensão terrestre foi publicado em 1903, como o *General Bathymetric Chart of the Oceans* (GEBCO). O GEBCO[1] é um projeto ativo até hoje, oferecendo comercialmente mapas e informações batimétricas.

A batimetria possui diversas aplicações, dentre as quais pode-se citar a exploração de recursos minerais (como petróleo e areia), o auxílio da navegação, o estudo das correntes submarinas profundas e a movimentação de sedimentos [37]. Os mapas batimétricos (ou cartas batimétricas) atuais são construídos a partir da combinação de dados coletados por navios e por satélites.

Para a coleta de dados batimétricos através de navios, são utilizadas ecossondas, que emitem ondas sonoras em direção ao fundo do mar e medem o tempo de retorno das ondas. Devem ser conhecidas as propriedades físicas da coluna de água (por exemplo, a salinidade e a temperatura podem alterar no tempo de resposta da onda [23]). Atualmente são utilizadas sondas capazes de emitir diversas ondas simultaneamente, para aumentar a área de cobertura [25]. Mesmo assim, não elimina-se a necessidade de interpolar os dados em áreas não aferidas. Estima-se que com a tecnologia disponível hoje, seriam necessários 125 anos para amostrar toda a extensão dos oceanos através de ecossondas [37].

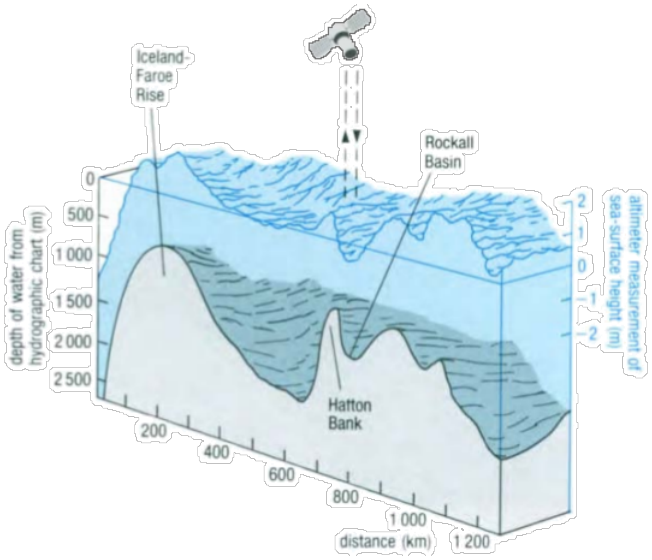


Figura 2.1: Relação entre a superfície e a batimetria oceânica. Fonte: [37]

Os dados obtidos através de satélites não são tão precisos, mas têm a vantagem da cobertura total da superfície terrestre. Os satélites altimétricos não medem a profundidade do oceano, mas a superfície da água. Existe uma relação direta entre a batimetria e a altitude da superfície, devido

[1] http://www.gebco.net

a propriedades gravitacionais da terra. Por exemplo, quando há uma elevação de 2km no fundo do oceano, a superfície da água eleva-se cerca de 2m naquele ponto. Apesar de não ser na mesma magnitude, essa relação (ilustrada na Figura 2.1) é proporcional, permitindo o mapeamento batimétrico aproximado.

2.1.1 Dados batimétricos

Uma base de dados batimétricos deve minimamente conter informações de localização e profundidade. Por exemplo, poderia apresentar-se na forma de um conjunto de tuplas compostas de coordenadas (latitude e longitude) e profundidade. Entretanto, um formato assim pode necessitar de muitos registros para ser suficientemente preciso. Formatos mais eficientes são utilizados para grandes volumes de dados. Por exemplo, a GEBCO disponibiliza através do *British Oceanographic Data Centre* (BODC) [2] uma base dados formatada como uma grade em intervalos fixos de distância, onde cada ponto contém a informação de sua respectiva profundidade.

2.1.2 Imagens batimétricas

Imagens batimétricas tipicamente contêm um mapa enfatizando, através de uma escala de cores, a batimetria. Imagens que representam grandes extensões do globo terrestre são representadas através de uma projeção cartográfica [31]. Existem diversas projeções cartográficas, que podem ser classificadas por superfície, por exemplo, cilíndrica, cônica, azimutal ou poliédrica; ou por preservação de uma propriedade métrica, como forma, área ou distância.

Um exemplo é a projeção geográfica (ou equiretangular, ou *Platte Carré*), que considera meridianos e paralelos como perpendiculares entre si, ambos em espaços equidistantes. Um exemplo dessa projeção é apresentado na Figura 2.2 [3].

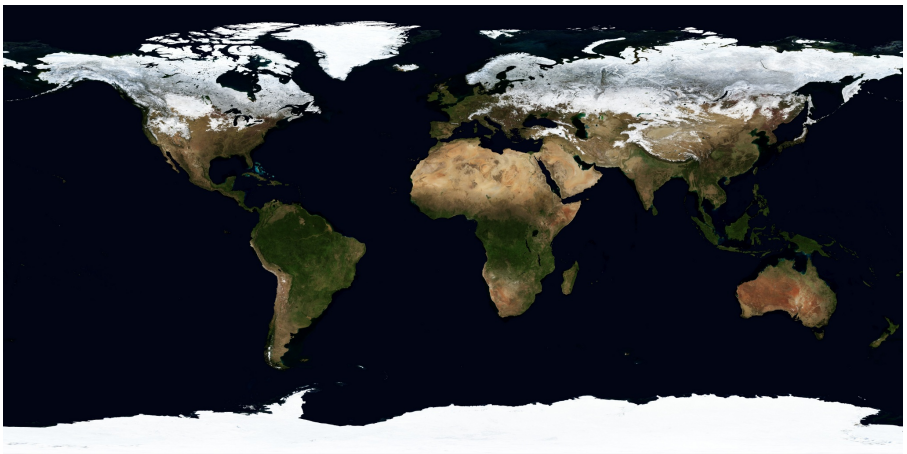


Figura 2.2: Mapa representado na projeção geográfica

[2] `https://www.bodc.ac.uk/data/online_delivery/gebco/`
[3] `http://visibleearth.nasa.gov/view.php?id=73938`

A projeção geográfica introduz uma distorção que se agrava com a proximidade dos polos. Essa projeção, entretanto, é uma das mais utilizadas para representar dados cartográficos em imagens digitais, devido à facilidade do mapeamento de coordenadas para pixels na imagem. Para mapear uma coordenada representada pela latitude no intervalo [90°N, -90°S] e longitude no intervalo [-180°L, 180°O] para um pixel p, utiliza-se

$$x_p = \frac{lon + 180}{360} \times w \qquad y_p = \frac{-lat + 90}{180} \times h$$

onde w é a largura da imagem e h, a altura. Esta fórmula considera (0, 0) o pixel do canto superior esquerdo e (w, h) o pixel do canto inferior direito da imagem. Além disso, há a possibilidade de representar todo o universo de coordenadas na imagem, o que não ocorre, por exemplo, com a projeção de Mercator, que também considera os paralelos e meridianos entre si, porém, sendo uma projeção cilíndrica, os polos teriam uma distância infinita, e a representação é normalmente truncada em um determinado intervalo de latitude. A Figura 2.3 [4] demonstra como funciona a projeção de Mercator.

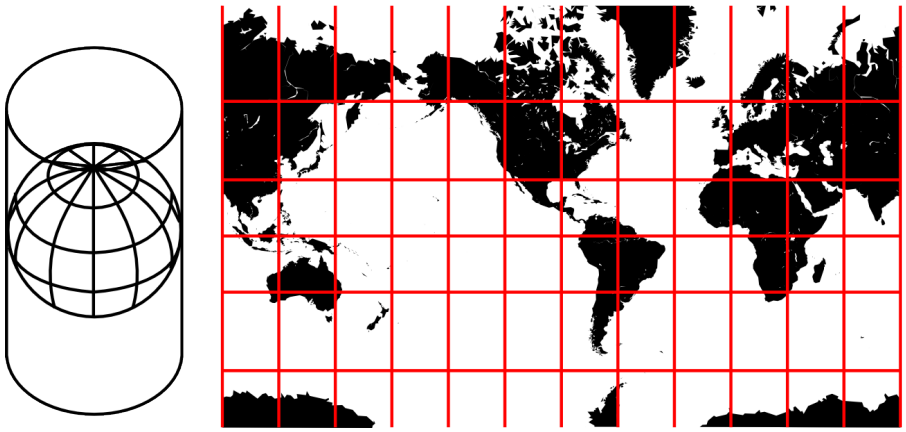
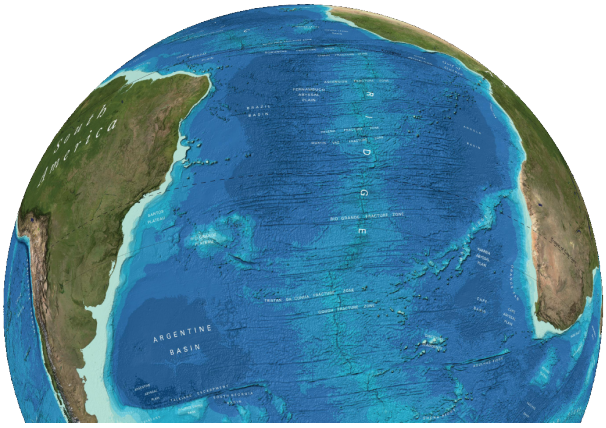


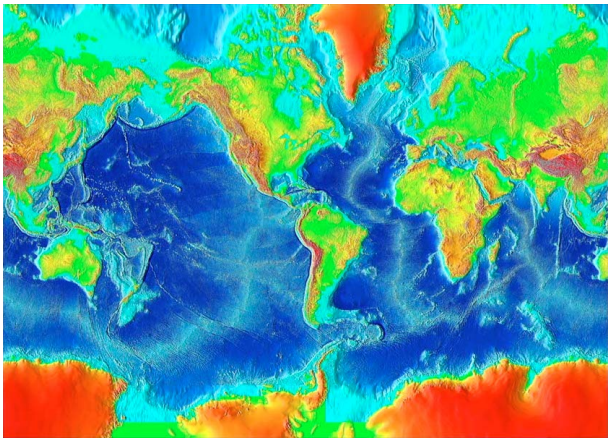
Figura 2.3: Projeção de Mercator

Para representação de imagens batimétricas, pode-se usar qualquer projeção cartográfica. Áreas menores podem inclusive ser apresentadas em perspectiva. Para ilustrar alguns exemplos da diversidade possível de imagens batimétricas, a Figura 2.4a (fornecida pela GEBCO[5]) apresenta a batimetria de uma região terrestre em uma projeção esférica; a Figura 2.4b (fornecida pela NOAA[6]) apresenta um mapa na projeção de Mercator contento não somente a batimetria, mas também a topografia; e a Figura 2.4c (fornecida pela NASA[7]) apresenta uma perspectiva da batimetria na bacia de Los Angeles, nos Estados Unidos da América.

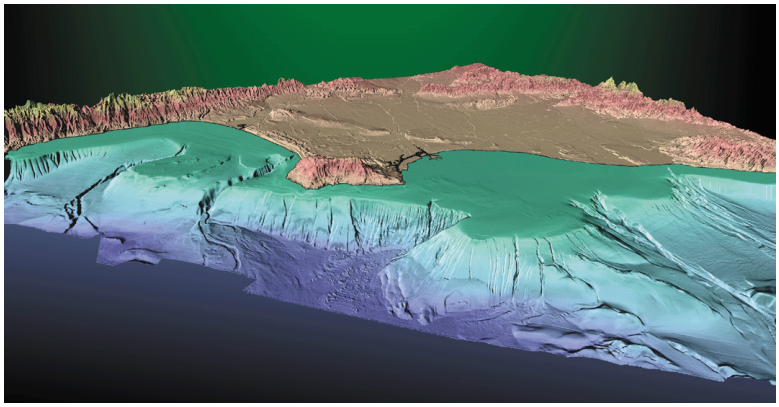
[4] http://upload.wikimedia.org/wikipedia/commons/6/62/Usgs_map_mercator.svg
[5] http://www.gebco.net/general_interest/documents/gebco_08_south_atlantic.pdf
[6] http://www.ngdc.noaa.gov/mgg/image/relief_slides2.html
[7] http://visibleearth.nasa.gov/view.php?id=55316



(a) Imagem batimétrica fornecida pela GEBCO



(b) Imagem batimétrica fornecida pela NOAA



(c) Imagem batimétrica fornecida pela NASA

Figura 2.4: Exemplos de imagens batimétricas

2.2 Mineração de dados

A mineração de dados é o processo de descoberta automática de informações úteis em grandes repositórios de dados [34]. É uma técnica capaz de realizar tarefas dos seguintes tipos:

- Tarefas preditivas: Têm como objetivo a predição do valor de um atributo particular a partir dos valores dos outros atributos. Através da análise de um conjunto de dados que possuem um valor conhecido para um atributo de interesse é induzido um modelo capaz de inferir o valor desse atributo para registros antes desconhecidos. As tarefas preditivas mais utilizadas são a classificação (quando o atributo de interesse é um rótulo de classe) e a regressão (quando esse atributo é um valor numérico).

- Tarefas descritivas: Estas tarefas analisam registros a fim de derivar padrões (como correlações, tendências, grupos ou anomalias) que sumarizem relacionamentos nos dados. Destacam-se neste grupo as tarefas de análise de agrupamentos (que agrupa registros com características semelhantes), de análise de associações (a partir de entradas do tipo "cesta de itens", infere regras de associações entre os itens) e a detecção de anomalias (que aponta registros com comportamento significativamente diferente dos demais).

O processo de mineração de dados tipicamente envolve as seguintes tarefas [18]:

1. Pré-processamento: Composto pelas atividades que precedem a mineração de dados em si. Integração de distintas bases de dados, limpeza, seleção e transformação de dados são atividades possíveis desta etapa.

2. Mineração de dados: Consiste na aplicação de algoritmos de mineração de dados em si.

3. Pós-processamento: Caracterizado pelo processamento de padrões encontrados e pela preparação de visualizações dos mesmos a fim de permitir o entendimento dos achados.

Conforme descrito no Capítulo 3, este trabalho foca-se principalmente na tarefa de classificação, mas também na análise de agrupamento. Assim, ambas as técnicas são descritas a seguir.

2.2.1 Classificação

Segundo Tan, Steinbach & Kumar [34], na tarefa de classificação, a partir de um *conjunto de treino* (registros previamente classificados), aplica-se um algoritmo de aprendizagem que induz um modelo de classificação. Esse modelo é então aplicado a um *conjunto de testes*, que consiste em registros cujo rótulo de classe não é previamente conhecido. Este processo está ilustrado na Figura 2.5.

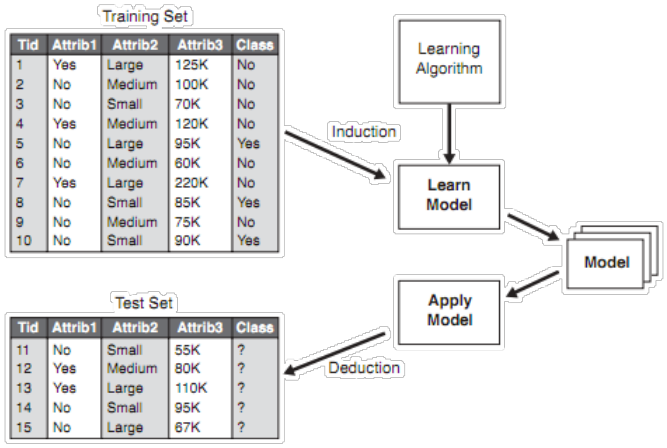


Figura 2.5: Aplicação da tarefa de classificação. Fonte: [34]

Existem diversos modelos de classificação possíveis que, dependendo da característica dos dados e dos objetivos da mineração, podem ser mais apropriados do que outros. Alguns facilitam o entendimento, outros obtêm um melhor desempenho na classificação e outros, ainda, exigem menos recursos computacionais para a construção do modelo. A seguir são apresentados, dentre os modelos de classificação mais utilizados, os que foram aplicados neste trabalho: árvores de decisão e classificação Bayesiana.

Árvores de decisão

Árvores de decisão são árvores onde cada nodo intermediário contém um atributo a ser testado, e cada nodo folha contém um valor de classe. Assim, percorrendo-se um caminho desde o nodo raiz, chega-se ao resultado da classificação. Segundo Tan, Steinbach e Kumar [34], as principais vantagens das árvores de decisão são a facilidade do entendimento do modelo gerado e a robustez quanto a eventuais ruídos nos dados. A Figura 2.6 ilustra uma árvore de decisão.

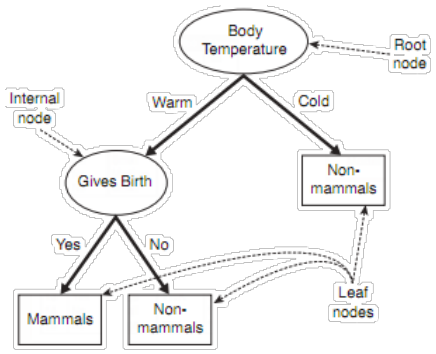


Figura 2.6: Exemplo de árvore de decisão. Fonte: [34]

Como algoritmos clássicos baseados em árvores de decisão, podem ser citados o *CART* [9] e o *C4.5* [29]. Ambos algoritmos baseiam-se na ideia de selecionar um atributo que possua maior seletividade em relação ao atributo classe, a partir da maximização de determinado critério. O teste desse atributo irá gerar um nodo na árvore de decisão. A partir desse nodo, são geradas arestas de acordo com os possíveis valores desse atributo, e o conjunto de treino é dividido em função do valor desse atributo. Cada aresta pode levar a um novo nodo de teste de atributo ou ao valor de classe previsto.

Esta abordagem apresenta dois problemas [5]:

- São algoritmos gulosos, que varrem todo o conjunto de treino, testando cada possível valor de cada atributo;
- Baseiam-se em máximos locais, e não em máximos globais. Isto ocorre porque o conjunto de treino é dividido a cada nodo, ocasionando a possível perda da melhor solução global.

Para contornar o problema dos máximos locais, a abordagem tradicional é a utilização de técnicas de combinação de classificadores (*ensembling*). Estas técnicas geram múltiplas árvores de decisão, combinando-as em um único modelo. Alguns exemplos dessas técnicas são:

- *Bagging* [7]: Esta técnica amostra o conjunto de treino segundo uma distribuição probabilística. Cada amostra é utilizada para a inferência de um modelo base. Para a classificação de um registro, cada modelo base é utilizado para determinar o valor de classe e aquele com maior votação é retornado. O ganho de desempenho de classificação que esta técnica pode trazer depende da instabilidade dos modelos base: quanto mais instáveis, maior é o ganho potencial.

- *Boosting* [14]: A técnica de *boosting* consiste em inferir iterativamente modelos base a partir de amostras do conjunto de treino. A cada registro utilizado é atribuído um peso (em geral, relacionado à precisão da classificação). Na próxima iteração, registros com maior peso possuem maior probabilidade de serem utilizados no novo modelo base. Para a classificação final, é realizada uma votação entre os modelos base, penalizando aqueles com menor precisão.

- *Random Forest* [8]: A partir de uma dada distribuição probabilística, o conjunto de treino é amostrado para a geração de um modelo base. Para determinar a melhor condição de teste de cada nodo, são selecionados aleatoriamente F atributos. Como as outras técnicas de combinação de classificadores, são gerados vários modelos base, e o resultado da classificação é a classe com maior votação entre eles.

As técnicas de combinação de classificadores, em geral, apresentam melhor desempenho de predição em relação a algoritmos tradicionais de inferência de árvores de decisão. Entretanto, também possuem suas limitações, especialmente pela perda da facilidade de entendimento do modelo induzido, devido a um número potencialmente grande de árvores intermediárias (e possivelmente conflitantes entre si). Uma técnica distinta para a indução de árvores de decisão únicas (em contraste à combinação de modelos das técnicas de *ensembling*) e que aproxima-se de máximos globais (e não apenas de máximos locais, como no caso do CART e o C4.5) foi proposta por Basgalupp et al. [5], e utiliza-se de algoritmos genéticos (também chamados de algoritmos evolucionários) para a indução de árvores de decisão.

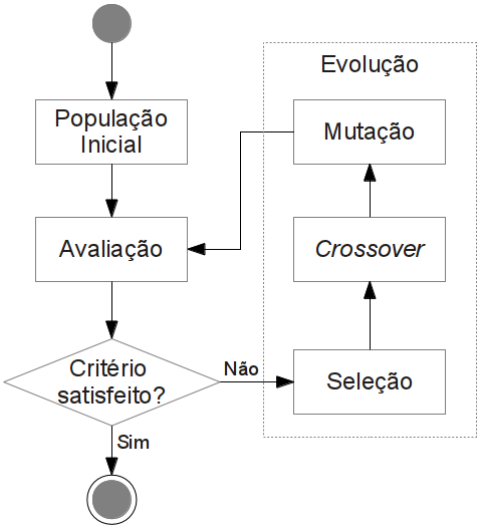


Figura 2.7: Fluxo geral de um algoritmo genético. Fonte: [4]

Algoritmos genéticos são coleções de técnicas de otimização cujo design baseia-se em metáforas do processo biológico [27]. Em geral, algoritmos genéticos trabalham com a ideia de uma população inicial de indivíduos (geração), cada um representando uma possível solução do problema em questão. Essa população evolui em direção à melhor solução, através de operações como *crossover* (combinação de progenitores para a geração da prole) e mutação (alteração súbita e aleatória de

alguma característica da prole, a fim de aumentar a entropia). Cada novo indivíduo é avaliado segundo um determinado critério. Aqueles indivíduos com melhor avaliação têm mais chance de serem selecionados para gerar novos indivíduos na próxima geração. Após um critério de parada (por exemplo, número máximo de gerações ou valor mínimo para avaliação), aquele indivíduo melhor avaliado é selecionado como solução. A Figura 2.7 apresenta o fluxo geral de um algoritmo genético.

Classificação Bayesiana

Conforme descrito por Han & Kamber [18], este tipo de classificação é baseada no teorema de Bayes. Este teorema é um modelo estatístico que permite determinar a probabilidade de hipóteses ocorrerem em um determinado conjunto de registros. Seja *P(H|X)* a probabilidade da hipótese *H* estar correta dado *X* (também chamada de probabilidade *a posteriori*); *P(X|H)* a probabilidade de *X* ocorrer, dada a hipótese; *P(H)* a probabilidade da hipótese ocorrer; e *P(X)* a probabilidade de *X* ocorrer, o teorema de Bayes é definido por

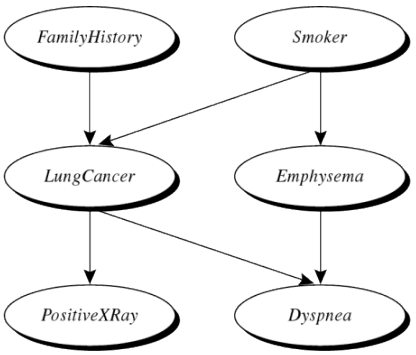
$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Para a classificação utilizando o teorema de Bayes, a hipótese é que um registro com determinado atributo *X* pertença à classe *C*. *P(H)* é a probabilidade *a priori* de um registro qualquer pertencer à classe *C*. Os classificadores baseados no teorema de Bayes buscam maximizar a probabilidade a posteriori de *X* (por isso chamados *maximum a posteriori*), retornando como resultado a classe com maior probabilidade. Alternativamente, pode-se obter como resultado uma distribuição de probabilidades para cada classe.

Uma característica dos classificadores Bayesianos é que o modelo pode ser facilmente atualizado com um novo registro de treino, bastando atualizar as probabilidades correspondentes ao novo registro.

O classificador Bayesiano mais simples é o *naïve Bayes* (ou Bayes ingênuo). Ele possui esse nome porque desconsidera que possam ocorrer correlações entre atributos, ou seja, dados quaisquer atributos *A* e *B*, *A* é condicionalmente independente de *B*. O *naïve Bayes* é computacionalmente barato, pois apenas mantém contadores para cada atributo, e necessita executar operações aritméticas básicas.

Na prática, as correlações entre atributos são comuns (por exemplo, quanto maior a idade, maior a escolaridade). Assim, pode ocorrer a perda de precisão na classificação do *naïve Bayes* por causa dessa característica. Para superar essa limitação, foi desenvolvido o algoritmo de redes Bayesianas [28] (também chamado de redes de crenças Bayesianas, redes de crenças ou redes probabilísticas). Uma rede Bayesiana é composta de dois componentes [18]: um grafo acíclico dirigido (Figura 2.8a) e um conjunto de tabelas de probabilidades condicionais (Figura 2.8b). Quando há uma aresta do nodo *Y* ao nodo *Z*, *Y* é o pai ou predecessor imediato de *Z*, e *Z* é o descendente de *Y*.



(a) Grafo acíclico dirigido

	FH, S	*FH, ~S*	*~FH, S*	*~FH, ~S*
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

(b) Tabela de probabilidades condicionais

Figura 2.8: Rede Bayesiana. Fonte: [18]

Em uma rede Bayesiana, cada variável é condicionalmente independente das variáveis que não são seus descendente, dados seus pais. No Exemplo da Figura 2.8, a variável *LungCancer* é condicionalmente dependente de *FamilyHistory* e de *Smoker*, mas é condicionalmente independente de *Emphysema*.

Para cada variável, é mantida uma tabela de probabilidade condicional (TPC), que contém a probabilidade de uma variável dados os valores possíveis de seus pais, calculada por *P(Y|Pais(Y))*. No exemplo mostrado na Figura 2.8b, a TPC para a variável *LungCancer* contém a probabilidade de *LungCancer* ser verdadeiro ou falso para cada valor possível de seus antecessores (*FamilyHistory* e *Smoker*). Neste exemplo, todas as variáveis são booleanas, mas para outros domínios, a tabela pode ser criada utilizando-se cada valor possível (ou discretizações no caso de valores contínuos).

Para calcular a probabilidade de uma determinada combinação de variáveis em uma rede Bayesiana, sendo X_i cada um dos atributos e x_i os valores possíveis dos atributos, pode-se aplicar

$$P(x1,\ldots,xn) = \prod_{i=1}^{n} P(x_i|Pais(Y_i))$$

onde *P(x1, . . . , xn)* é a probabilidade de uma combinação particular de valores de X e os valores de $P(x_i|Pais(Y_i))$ correspondem às entradas da TPC para Y_i. No caso da classificação, utiliza-se como X o atributo classe, calculando-se a probabilidade para cada possível rótulo de classe dados os valores das TPCs para os outros atributos dados seus pais.

Outros modelos de classificação

Além dos modelos de classificação descritos acima, há diversos outros que podem ser relacionados. Alguns dos principais exemplos, segundo Han & Kamber [18], são os classificadores baseados em regras, retro-propagação (redes neurais), máquinas de suporte vetoriais (SVMs), associação e aprendizagem preguiçosa (vizinhos mais próximos) e algoritmos genéticos.

Avaliação dos resultados da classificação

Tipicamente, o conjunto de dados de entrada (com registros previamente classificados) é utilizado tanto para a indução do modelo de classificação (conjunto de treino) quanto para o teste do mesmo (conjunto de teste). Porém, para que o modelo seja avaliado de maneira mais próxima à realidade, é desejável que os conjuntos de treino e de teste não sejam os mesmos. Existem algumas técnicas bem conhecidas para o particionamento dos dados originais, como [34]:

- *Holdout*: O conjunto de dados é dividido em 2 conjuntos disjuntos: um deles utilizado como treino e outro como teste.
- Sub-amostragem aleatória: Esta técnica consiste na repetição do método *holdout* *n* vezes, selecionando-se registros aleatoriamente a cada iteração. A precisão final é a média de todas as iterações.
- Validação cruzada: Semelhante à sub-amostragem aleatória, porém mais controlada. Os dados originais são divididos em *n* subconjuntos disjuntos. Cada um desses subconjuntos é utilizado uma vez como teste, enquanto os outros subconjuntos são utilizados como treino. Este processo é repetido para todos os subconjuntos. Finalmente, a precisão atribuída ao modelo é a média das precisões de cada uma das *n* iterações. Um caso particular desta técnica é o *leave one out* (deixe-um-fora), quando *n* é igual ao tamanho do conjunto de dados de entrada.

A avaliação do modelo de classificação geralmente ocorre a partir da matriz de confusão [34]. Ela é uma matriz cuja célula I_{ij} denota a contagem de registros sabidamente da classe *i* que foram classificados pelo modelo como a classe *j*. As linhas denotam as classes reais, enquanto as colunas, as classes determinadas a partir do modelo. Assim, a diagonal principal (células onde $i = j$) representa os acertos, enquanto todas as outras células são erros de classificação. A Figura 2.9 representa uma matriz de confusão.

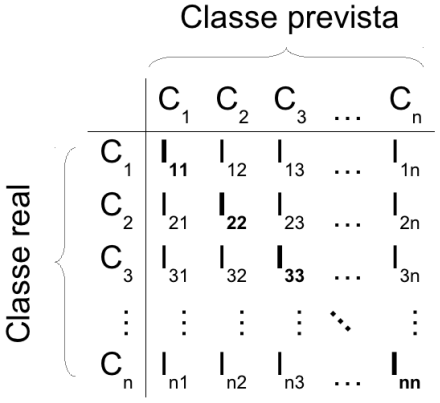


Figura 2.9: Representação de uma matriz de confusão

Outra forma comum de representar a avaliação de desempenho de um classificador é através da precisão, que é a porcentagem de registros que tiveram sua classe corretamente prevista pelo

modelo. A precisão é calculada a partir da divisão do número de registros corretamente classificados pelo número total de registros no conjunto de testes. Pode-se também representar a precisão de cada classe, sendo calculada pela divisão do número de registros corretamente classificados na classe pelo número total de registros daquela classe.

2.2.2 Análise de agrupamentos

A análise de agrupamentos é uma tarefa não supervisionada que busca agrupar os registros de um conjunto de dados de forma automática, sendo útil por agrupar os dados em grupos previamente desconhecidos [18].

Um algoritmo bem conhecido para a análise de agrupamentos é o *k-means* (*k*-médias), que segundo Alpaydin [1] tem como parâmetro de entrada *k*, que é o número de agrupamentos esperados, e inicia selecionando *k* objetos como centroides iniciais. Então, todos os outros objetos são atribuídos ao centroide mais próximo, avaliados conforme uma função de similaridade (por exemplo, pela distância euclidiana) para cada atributo. Cada centroide representa um grupo. Em seguida, atualiza-se o centroide de cada grupo (por exemplo, através da média dos elementos do grupo). Então um novo centroide é selecionado para cada grupo, e o algoritmo, reaplicado, até que os centroides não mudem, ou satisfaçam a um critério de parada.

A Figura 2.10 apresenta sucessivas iterações do algoritmo, com 3 centroides sorteados aleatoriamente, com os demais pontos sendo atribuídos ao centroide mais próximo e o novo centroide sendo atualizado com a média dos demais elementos do grupo.

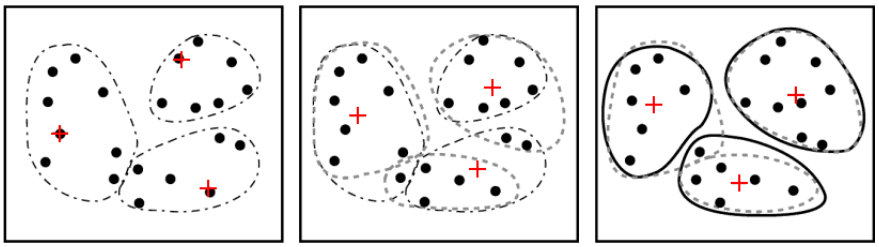


Figura 2.10: Aplicação do algoritmo *k-means*. Fonte: [18]

2.3 Mineração de dados em imagens

Os algoritmos clássicos de mineração de dados utilizam dados tabulares como entrada, onde cada registro é descrito por um conjunto fixo de atributos. Quando outros tipos de dados de entrada são utilizados (como imagens, sons, vídeos ou textos) apresentam-se dois caminhos possíveis: desenvolver novos algoritmos capazes de minerar nativamente esses dados ou pré-processar os dados de entrada a fim de obter registros compatíveis com os algoritmos da mineração de dados.

Após uma revisão de literatura, realizada nas bases de dados *IEEEXplore*[8] e *ACM Digital Library*[9], utilizando-se as palavras-chave *image data mining*, entre os primeiros 25 resultados de cada busca

[8] http://ieeexplore.ieee.org/
[9] http://dl.acm.org/

não foram encontrados artigos que nativamente apliquem algoritmos de mineração de dados em imagens. Dentre os resultados citados, existem trabalhos que utilizam buscas por conteúdo baseados em atributos extraídos de imagens. Pode-se comparar essas buscas com critérios do tipo: "cidades com n°de habitantes $\geq$ 10.000". Porém, o conceito de mineração de dados como um processo de "descoberta automática de informações" [34] não admite caracterizar um critério de busca como mineração de dados. Entretanto, existem diversos trabalhos que utilizam-se do pré-processamento de imagens, a fim de extrair dados em formato compatível com algoritmos de mineração de dados. Alguns desses trabalhos são descritos na Seção 3.2.

Para a extração de registros, deve-se considerar que uma imagem pode conter uma infinidade de informações (distintos objetos, regiões ou textos). Dependendo do tipo de problema abordado, apresentam-se as seguintes opções: utilização da imagem inteira como um único registro ou particionamento da imagem em diversos registros distintos. Tipicamente, trabalhos que buscam classificar imagens de acordo com o seu conteúdo utilizam cada imagem como um único registro [12, 20]. Quando há a necessidade do particionamento da imagem em distintos registros, pode-se adotar regiões de tamanho fixo [17], detecção de objetos salientes (reconhecidos através de técnicas de visão computacional) [13] ou mesmo ter cada pixel da imagem como um registro distinto [21].

Quanto à extração de atributos, existem técnicas bem conhecidas de processamento de imagens que podem ser utilizadas, além de técnicas específicas para o domínio do problema abordado. Como técnicas bem conhecidas, podem ser citadas:

- **Histograma de cores**: Consiste na utilização de um histograma, pressupondo como parâmetro de entrada o número de intervalos a serem representados. Então, cada pixel da imagem é contado no intervalo correspondente à discretização de seu valor de cor [26]. Essa discretização pode dar-se de diferentes formas. Um exemplo seria converter o modo de cores da imagem de entrada para uma paleta de cores limitada, contando os pixels de cada cor da paleta na respectiva entrada do histograma.

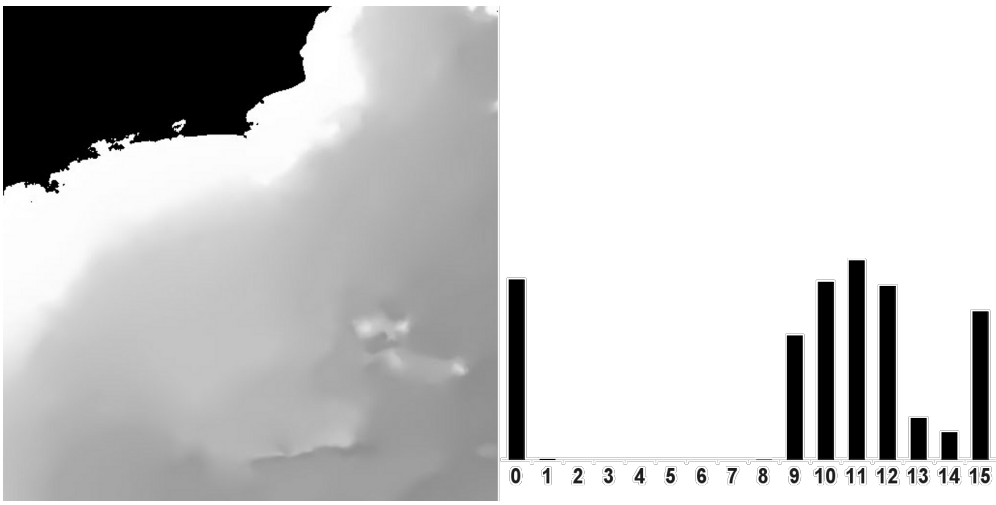


Figura 2.11: Histograma de cores

Outra opção seria converter a imagem para a escala de cinza (256 níveis), contando cada pixel na entrada calculada pela divisão do nível de cinza do pixel pelo número de entradas no histograma. Cada entrada do histograma obtido é utilizado como um atributo para a mineração de dados. A Figura 2.11 apresenta visualmente um histograma de 16 entradas para a respectiva figura (em escala de cinza).

- **Estatísticas dos pixels**: Consiste no mapeamento de atributos estatísticos dos pixels. Podem ser utilizados os componentes de cor individualmente (vermelho, verde, azul ou matiz, luz, saturação) ou o nível de cinza (para imagens em escala de cinza). Sobre estes, podem ser extraídos atributos estatísticos como média, desvio padrão, variância, mediana ou moda.

- **Coeficientes de *Wavelets*** [22]: Nesta abordagem, são utilizados os coeficientes dos *wavelets* dominantes, que capturam formas, texturas e localizações. Quando aplicada sobre uma imagem, a *transformada discreta de Wavelet* (DWT) gera uma nova imagem, contendo 4 componentes, representados na Figura 2.12: aproximação A, detalhes horizontais H, detalhes verticais V e detalhes diagonais D [16]. A aproximação é uma imagem com metade da resolução original, enquanto os outros componentes contêm coeficientes que representam as diferenças em cada um dos sentidos (horizontal, vertical e diagonal) da aproximação em relação à imagem original. Este processo pode ser reaplicado no componente da aproximação, resultando em uma nova imagem com os 4 componentes. Neste caso, foi aplicada uma DWT de 2 níveis. O processo pode ser repetido iterativamente, aumentando-se o nível da DWT. É possível reconstruir a imagem original, sem perdas, a partir dos 4 componentes, através da aplicação da transformada inversa discreta de Wavelet (IDWT).

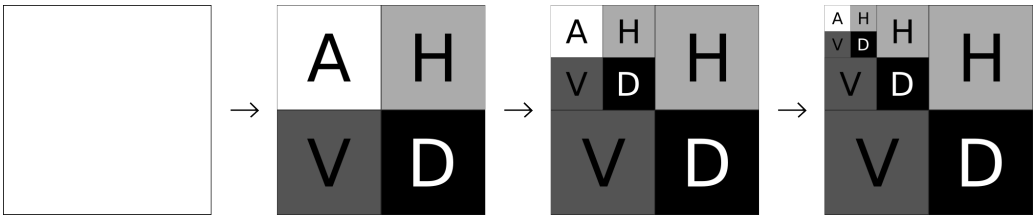


Figura 2.12: Aplicação da transformada discreta de Wavelet em imagens

Diversos trabalhos utilizam essas técnicas, e também outras, desenvolvidos pelos próprios autores, de acordo com o domínio do problema abordado. Por exemplo, quando o objetivo da mineração de dados é o reconhecimento de objetos, diversos trabalhos têm aplicado tanto o histograma de cores quanto coeficientes de *Wavelets*, como Fan et al. [13] e Ding et al. [12]. Ainda outros trabalhos, como Barnes et al. [2, 3], desenvolveram técnicas novas, específicas para o domínio de problemas abordado. No caso de Barnes et al., o domínio abordado é a busca de fragmentos de imagens dentro de um mapa.

2.4 Considerações finais do capítulo

Neste capítulo foi apresentada a revisão bibliográfica necessária para o entendimento deste trabalho. Inicialmente foi apresentado o conceito de batimetria e como os dados batimétricos são aferidos. Especialmente, destaca-se a medição da superfície da água, através de satélites altimétricos, visto que este trabalho, conforme descrito no Capítulo 4, utiliza imagens batimétricas satelitais.

Foi também apresentado o conceito de mineração de dados e os tipos de tarefa que esta é capaz de realizar. Este trabalho foca-se na tarefa de classificação, e, secundariamente, análise de agrupamentos. Assim, foram apresentados os modelos de classificação utilizados por este trabalho, especialmente árvores de decisão e classificadores Bayesianos.

Finalmente, visto que os algoritmos clássicos de mineração de dados utilizam dados tabulares como entrada (e não imagens), foram apresentadas técnicas bem conhecidas para a extração de registros e de atributos a partir das imagens. Esta é a abordagem utilizada na prática para minerar imagens, conforme resultado de busca na literatura.

3. OBJETIVOS DESTE TRABALHO

Conforme apresentado na Seção 2.1, a batimetria possui diversas aplicações importantes, da preparação de rotas de navegação à prospecção de recursos naturais. Recentemente, diversas pesquisas estão sendo direcionadas para a batimetria, e têm-se buscado dados batimétricos mais detalhados [37].

Imagens batimétricas globais de alta resolução podem ser facilmente encontradas na Internet, disponibilizadas por instituições como a GEBCO[1], NASA[2] e NOAA[3]. Entretanto, a análise apenas por inspeção visual de áreas extensas pode ser bastante trabalhosa e sensível a erros. Por exemplo, podem ocorrer variações sutis na conformação do fundo oceânico, bem como semelhanças de conformações em distintas regiões no mapa. Assim, seria interessante a disponibilização de ferramentas computacionais que possam analisar imagens batimétricas de forma automática ou semi-automática, a fim de auxiliar a um especialista de domínio na análise dessas imagens e no entendimento dos mais diversos problemas que estejam a elas relacionados.

A seguinte questão de pesquisa foi enunciada para este trabalho: "Dada uma imagem batimétrica, de acordo com um formato esperado, tendo como pressuposto um conjunto de regiões rotuladas por um especialista de domínio, é possível encontrar um modelo computacional capaz de rotular de forma consistente as outras regiões do mapa, com uma precisão satisfatória?"

Para a realização dessa análise é proposta a utilização da mineração de dados. A justificativa é que esta dispõe de técnicas bem conhecidas que são potencialmente úteis para realizar essa análise, especialmente a tarefa de classificação. A análise de agrupamentos também é uma técnica possível para a realização de análise automática da imagem batimétrica. A aplicação de outras tarefas da mineração de dados está fora do escopo deste trabalho.

Minerar imagens possui algumas vantagens em relação a minerar dados convencionais, como por exemplo:

- Existem imagens batimétricas de alta resolução (por exemplo, [4] disponibiliza imagens de até 240 pixels / °) disponíveis gratuitamente na Internet, representando a batimetria de toda a extensão do globo terrestre;

- Utilizando uma imagem, as relações (similaridades e dissimilaridades) entre áreas próximas ficam mais evidentes, se comparadas à utilização de dados tabulares;

- A utilização de uma imagem como entrada permite a apresentação visual dos resultados da mineração dentro da própria imagem, de forma natural.

[1] http://www.gebco.net
[2] http://www.nasa.gov
[3] http://www.noaa.gov
[4] http://visibleearth.nasa.gov/view_cat.php?categoryID=1484

Entretanto, a mineração de imagens possui alguns desafios, como:

- A necessidade do pré-processamento da imagem para a extração de registros e seus atributos, conforme descrito na Seção 2.3. Estes atributos devem ser extraídos diretamente do conteúdo da imagem, e representá-lo com precisão satisfatória;
- A extração de atributos a partir de imagens normalmente utiliza coeficientes e atributos numéricos, que não encontram correspondentes diretos no domínio do problema abordado. Esta característica é conhecida como lacuna semântica.

3.1 Requisitos de uma solução

Entende-se que uma solução para a análise de imagens batimétricas através da mineração de dados deve, ao menos, cumprir os seguintes requisitos:

- A imagem batimétrica deve ser separada em regiões. A cada região pode ser manualmente atribuído um rótulo de classe. Então, pode-se executar a tarefa de classificação para inferir o valor de classe das outras regiões do mapa, que não foram manualmente classificadas.
- Semelhantemente, pode-se executar a tarefa de análise de agrupamentos, para que as regiões possam ser agrupadas de forma automática.
- A visualização do processo de mineração de dados e de seus achados é fundamental para o entendimento dos mesmos por parte de um especialista de domínio. Assim, os resultados devem ser apresentados com uma iconografia intuitiva.

3.2 Trabalhos relacionados

Após uma busca realizada nas bases de dados *IEEEXplore*[5] e *ACM Digital Library*[6], utilizando-se as palavras-chave *bathymetry data mining*, não foram encontrados trabalhos que utilizam mineração de dados para a análise de imagens ou dados batimétricos.

Então, buscou-se publicações que utilizam mineração de dados em imagens, através das palavras-chave *image data mining*. Alguns exemplos de trabalhos encontrados são:

- Ding et al. [12] apresentam o *Annotating Images by Semantic Corpus (AISC)* que é capaz de classificar imagens de forma autônoma, através da mineração de dados. O conjunto possível de classes foi extraído do *Large-Scale Concept Ontology for Multimedia (LSCOM)*, que é uma ontologia contendo diversos termos semânticos comumente encontrados em conteúdos multimídia. Cada termo semântico foi utilizado em mecanismos de buscas de imagens na Internet. Foi baixado um determinado número de imagem para cada termo. Então, cada

[5] `http://ieeexplore.ieee.org/`
[6] `http://dl.acm.org/`

imagem foi processada com técnicas de extração de atributos e armazenada como um registro do conjunto de treino, criando um modelo de classificação. Todo esse processo foi realizado offline. Assim, dada uma nova imagem, o sistema extrai seus atributos e a classifica utilizando o modelo previamente inferido.

- Barnes et al. [3] apresentam uma técnica que utiliza um modelo, construído a partir de uma base de dados de blocos de pixels manualmente classificados. Utilizando uma nova imagem, são separados novos blocos de pixels e comparados com o modelo, permitindo a classificação de regiões dessa imagem. Os mesmos autores aplicaram essa técnica na análise de imagens de satélite a fim de detectar danos causados por desastres naturais como furacões e enchentes [2]. Foram demonstrados resultados da técnica em imagens da destruição causada pelo furacão Katrina, que atingiu a costa sul dos Estados Unidos em 2005, classificando regiões do mapa como danificadas.

- Lu & Yang [21] propõem o uso da mineração de dados em imagens, a fim de classificá-las utilizando árvores de decisão. Neste trabalho, cada pixel da imagem é considerado um registro distinto para a mineração. Como exemplo de aplicação, os autores utilizam imagens de letras e números, para o reconhecimento de caracteres. Foram comparados os resultados com aplicações tradicionais de OCR (*Optical Character Recognition*), tanto com textos normais quanto desfocados ou com ruído. A aplicação da mineração de dados mostrou-se superior, especialmente nos casos onde o texto estava desfocado ou havia ruído.

- Gueguen & Datcu [17] propõem um método para a análise de séries temporais de imagens de satélite (SITS), que é uma sequência de imagens de satélite de um mesmo local, ao longo do tempo. O algoritmo separa cada imagem em células quadradas no formato de matriz (a matriz tem o mesmo formato para todas as imagens), e realiza a análise de agrupamentos sobre cada célula. O número de grupos é um parâmetro do algoritmo, e as células podem ser atribuídas a grupos distintos ao longo da série.

- Fan et al. [13] apresentam um algoritmo chamado *product of mixture-experts (PoM)*, que realiza a classificação de uma imagem através de uma ontologia de conceitos, sendo os mais gerais, aplicáveis à imagem como um todo, e os mais específicos aplicáveis aos "objetos salientes", que são delimitados através de técnicas de visão computacional. Para a classificação dos objetos salientes, são extraídas suas caraterísticas através de atributos estatísticos (média e variância), histograma de cores e coeficientes de *wavelets*. Finalmente, a referida ontologia é utilizada para auxiliar na classificação, considerando co-ocorrências de objetos em cenas comuns para aumentar a precisão da predição.

3.3 Considerações finais do capítulo

A batimetria tem recebido atenção nos últimos anos, por possuir diversas áreas de aplicação. Assim, evidencia-se a necessidade de ferramentas computacionais para analisar dados batimétricos. Para tal, este trabalho propõe a utilização da mineração de dados, não aplicada a bases de dados convencionais com informações batimétricas, mas diretamente a imagens batimétricas.

Após uma pesquisa na literatura, não foram encontrados trabalhos que utilizem a mineração de dados para resolver problemas de batimetria. Consequentemente, também não foram encontrados trabalhos que minerem imagens batimétricas.

Foram, então, pesquisados trabalhos que utilizam a mineração de dados em imagens, para distintos domínios de imagens e problemas. Uma boa parte deles se propõem a classificar imagens a partir de seu conteúdo. Outros dividem a imagem em objetos ou blocos e buscam classificá-los ou agrupá-los. Nenhum deles, entretanto, destina-se especificamente à mineração de imagens batimétricas, onde não apenas características do conteúdo das imagens são importantes, mas também similaridades entre as conformações de distintas regiões do mapa devem ser levadas em conta para a mineração de dados. No Capítulo 4 é proposta uma abordagem para a aplicação da mineração de dados em imagens batimétricas.

Outro elemento importante para o processo de descoberta de conhecimento em imagens batimétricas é a visualização dos resultados. Uma forma intuitiva de apresentação gráfica do processo de descoberta de conhecimento, e, principalmente, dos achados desse processo pode ajudar de forma significativa ao especialista de domínio que esteja envolvido no processo. Assim, no Capítulo 5 é proposta uma forma de visualização para facilitar o entendimento dos achados no processo de mineração.

4. UMA ABORDAGEM PARA MINERAR IMAGENS BATIMÉTRICAS

Há uma infinidade de variações e possibilidades para imagens batimétricas. Alguns exemplos são apresentados na Figura 2.4. O tratamento de distintas projeções cartográficas e perspectivas do mapa pode dificultar muito o processamento das imagens, e está fora do escopo deste trabalho, podendo ser um novo tema de pesquisa. Assim, a abordagem proposta pressupõe que as imagens de entrada possuam características específicas, descritas a seguir.

4.1 Características esperadas da imagem

Neste trabalho, a entrada esperada é uma imagem de um mapa na projeção geográfica. Esta foi a projeção escolhida porque, conforme descrito mais adiante, a abordagem proposta divide o mapa em células quadradas, e a projeção geográfica define os paralelos e meridianos como perpendiculares entre si. Além disso, essa projeção permite facilmente mapear todo o espaço de coordenadas latitude e longitude para pixels na imagem, conforme descrito na Seção 2.1.2.

Somente é necessário um dado por pixel, a batimetria. É importante que o dado batimétrico lido respeite uma escala linear, mantendo as proporções. Assim, espera-se que a imagem de entrada possua, como modelo de cores, a escala de cinza. Há necessidade de diferenciar regiões que contenham terra seca, visto que estas não possuem utilidade para a batimetria. Por convenção, assume-se que os pixels de valor zero (preto absoluto) representam terra seca e os demais níveis de cinza (1 a 255) representam as distintas batimetrias. Os pixels mais claros representam menores profundidades (mais rasas) e os pixels mais escuros, as maiores profundidades.

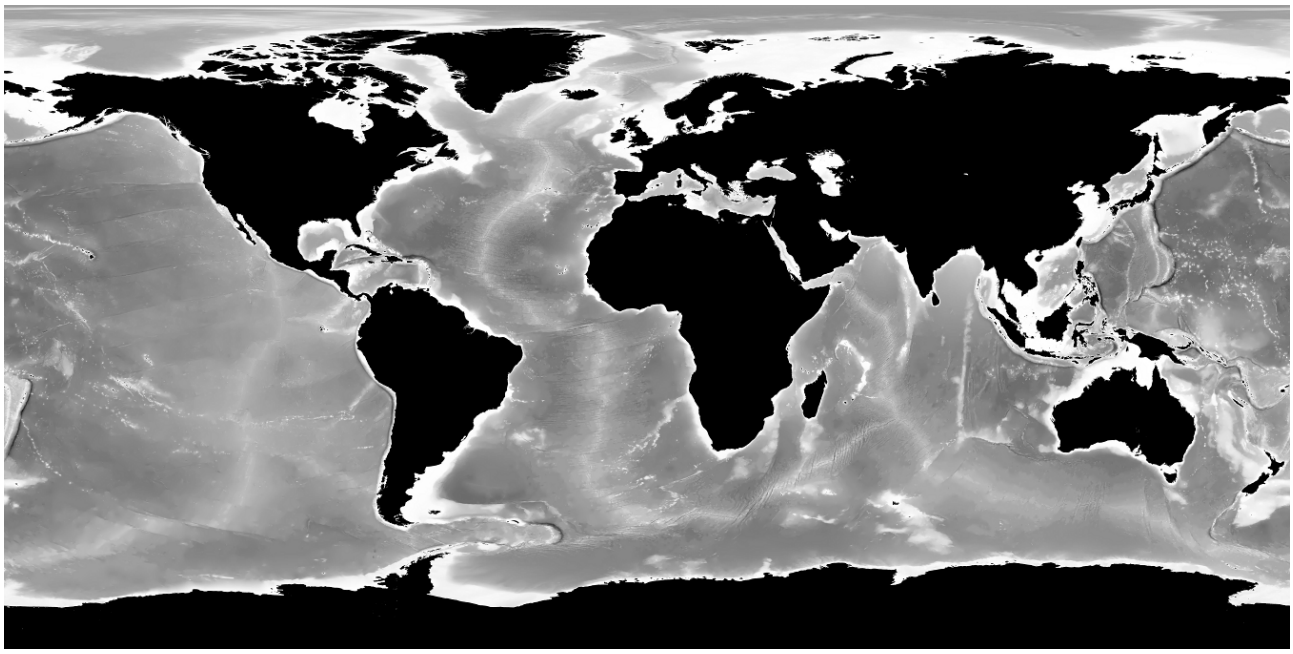


Figura 4.1: Mapa batimétrico utilizado neste trabalho

Uma imagem no formato descrito acima encontra-se disponível na Internet pelo projeto *Visible Earth*, da NASA, no endereço `http://visibleearth.nasa.gov/view_rec.php?id=8392`. São disponibilizadas imagens em alta resolução (21.601 $\times$ 10.801 pixels) e média resolução (5.400 $\times$ 2.700). O referido mapa é apresentado na Figura 4.1. É possível também utilizar um recorte do mapa, quando, por exemplo, há interesse apenas em uma região específica.

4.2 Preparação dos dados

Como o objetivo deste trabalho é a mineração de dados em imagens, a preparação dos dados é fundamental (conforme descrito na Seção 2.3). A seguir, são descritas técnicas para transformar a imagem batimétrica em registros que possam ser passados para os algoritmos de mineração de dados.

4.2.1 Extração de registros

O mapa é dividido em áreas quadradas, em pixels, de tamanho arbitrário, denominadas *células*. Cada célula é um registro para a mineração de dados. Visto que este trabalho busca minerar dados batimétricos, não há interesse em considerar regiões que contenham terra seca. Assim, células que contenham pixels totalmente pretos são descartadas.

4.2.2 Extração de atributos

Para a extração de atributos, são utilizadas diversas técnicas, algumas bem conhecidas e outras propostas neste trabalho. As técnicas bem conhecidas (ver Seção 2.3) empregadas neste trabalho são:

- **Atributos estatísticos**: Dois atributos estatísticos sobre os valores (nível de cinza) dos pixels da célula: média e desvio padrão. De forma sintética, são capturadas a batimetria média e o quão variada é a batimetria dentro da célula (através do desvio padrão). Por exemplo, células com pouca variação, como plataformas continentais, possuem um baixo desvio padrão, enquanto taludes continentais possuem um desvio padrão elevado.

- **Histograma**: Um histograma com um número arbitrário de entradas, onde cada atributo passado para o algoritmo representa a contagem de pixels cujo nível de cinza é aproximado para o intervalo. O histograma captura a distribuição da célula em relação às distintas faixas de profundidade. Por exemplo, se forem usadas 4 entradas no histograma, células que possuam um número elevado de pixels na entrada 1, alguns pixels na entrada 2 e nenhum nas demais, significa que uma boa parte da mesma é muito profunda, mas também possui áreas um pouco menos profundas. Através do histograma pode-se, por exemplo, determinar se uma faixa específica de profundidade é relevante para o modelo de classificação.

- **Coeficientes de *Wavelets***: Existe uma infinidade de abordagens para a utilização de *Wavelets* na extração de atributos. Neste trabalho, utiliza-se uma variação da técnica descrita por Wang et al. [36]. Não é exatamente a mesma técnica porque os autores utilizam imagens coloridas, com transformação dos componentes de cor, além de outros atributos complementares. A técnica utilizada aqui possui 2 variáveis: o tamanho base para a célula, que deve ser uma potência de 2 (isto é requisito da DWT); e o número de níveis da DWT que será aplicado a ela. Os componentes resultantes desse processo são compostos de coeficientes, sendo cada coeficiente utilizado como um atributo para mineração de dados. No teste da solução, apresentado no Capítulo 7, as células são redimensionadas para 128 $\times$ 128 pixels, e são aplicados 4 níveis da DWT, resultando em 256 atributos para a mineração de dados. Estas variáveis são as mesmas utilizadas por Wang et al. [36].

Além dessas, outras técnicas novas são propostas neste trabalho. Para a mineração de imagens batimétricas, são desejáveis atributos que descrevam, de forma aproximada, a morfologia da célula. Semelhanças de conformações de distintas células podem ser um indicador importante para o modelo de mineração, a fim de posicionar as células em determinada classe ou grupo. Estas técnicas de extração de atributos são descritas a seguir:

- **Regiões**: Os atributos das regiões capturam o formato aproximado da célula. Esta técnica consiste em dividir a célula em uma grade composta por regiões (sub-áreas da célula). Cada uma delas armazena a diferença da média dos valores dos seus pixels em relação à média dos pixels de toda a célula. Assim, é capturado o formato aproximado, independentemente de sua batimetria. São admitidos tanto valores positivos quanto negativos. A Figura 4.2 exemplifica os valores das regiões.

-48,1	-36,1	-20,1	-12,1
-20,1	-28,1	-34,1	-39,1
48,7	5,9	-9,1	-26,1
104,9	92,9	22,9	-7,1

Figura 4.2: Regiões

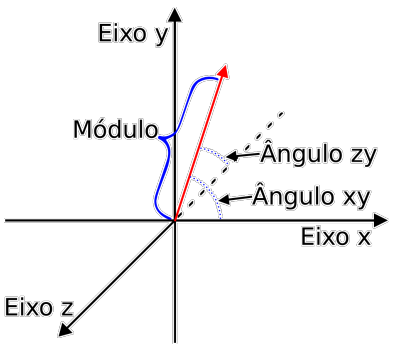
Como formas semelhantes podem ocorrer em distintas rotações, pode-se fazer com que aquela de maior valor dentre os 4 cantos seja posicionada no canto superior-esquerdo através de rotação em ângulos de 90°, 180°ou 270°; ou de espelhamento horizontal e / ou vertical. Para a finalidade de semelhanças de conformações não há diferença se utilizada rotação ou espelhamento. Assim, de uma forma fácil, pode-se encontrar regiões semelhantes em ângulos de aproximadamente 90°. Entretanto, como a rotação não é livre, semelhanças de conformações

que ocorram em ângulos intermediários podem não ser capturadas pelas regiões. A definição do número de regiões é importante, podendo variar de 2×2 até o tamanho da célula. Entretanto, o uso de valores muito altos pode dificultar o reconhecimento das formas, por serem muitos pontos, enquanto 2×2 dispõe de capacidade bastante limitada para descrever o formato da célula.

- **Vetor**: A fim de capturar a tendência da elevação e ângulos dominantes da célula, é utilizado um vetor. O vetor é calculado em função das regiões, que são médias, e não de pixels individuais, para evitar que pixels fora do padrão (com valor muito alto ou muito baixo) influenciem demasiadamente o vetor. Sejam *min* a região com a menor média de valores, e *max* a região com a maior média dos valores, o vetor possui as seguintes características:

$$x = x_{max} - x_{min}$$
$$y = y_{max} - y_{min}$$
$$z = valor_{max} - valor_{min}$$
$$\hat{a}ngulo_{xy} = atan(\frac{y}{x})$$
$$\hat{a}ngulo_{zy} = atan(\frac{y}{z})$$
$$módulo = \sqrt{x^2 + y^2 + z^2}$$

O ângulos são descritos em graus. Como as posições x e y já estão presentes nas regiões, z é simplesmente a diferença entre a maior e a menor região, estes não são passados para os algoritmos de mineração de dados. Os atributos efetivamente utilizados são o módulo do vetor e os ângulos xy e zy. A Figura 4.3a apresenta o vetor desenhado em um sistema de coordenadas tridimensional. Já a Figura 4.3b posiciona o vetor sobre as regiões.



(a) Vetor no sistema de coordenadas

-29,6	0,4	65,4	87,4
-28,6	-40,6	18,4	74,4
-26,6	-37,6	-21,6	32,4
-21,6	-26,6	-38,6	-13,6

(b) Vetor posicionado sobre as regiões

Figura 4.3: Vetor

Tanto para os atributos de regiões quanto do vetor, pode-se definir um limiar que delimita o valor mínimo do módulo do vetor para que a célula não seja considerada totalmente plana. Se o módulo não ultrapassa esse valor, a célula é considerada totalmente plana, e todos os atributos de regiões e do vetor são passados com o valor zero para o algoritmo de mineração.

4.2.3 Normalização dos dados

Dependendo do algoritmo de mineração ou da tarefa que está sendo executada, pode ser importante normalizar os valores dos atributos, por possuírem distintos domínios. A normalização dos dados é especialmente importante para algoritmos de análise de agrupamentos e de alguns algoritmos de classificação (como vizinhos mais próximos ou redes neurais) [18].

A Tabela 4.1 descreve os intervalos de valores encontrados para cada tipo de atributo na imagem descrita na Seção 4.1, utilizando-se células de tamanho 32 $\times$ 32 pixels e regiões de tamanho 4 $\times$ 4.

Tabela 4.1: Domínios dos atributos

Conjunto de atributos	**Atributo**	**Mínimo**	**Máximo**
Atributos estatísticos	Média	100,2	255,0
	Desvio padrão	0,0	65,6
Histograma	Histograma	0	1024
Regiões	Região	-107,8	108,5
Vetor	Módulo	0	186
	Ângulo XY	-161,6	180,0
	Ângulo ZY	-71,6	71,6
Coeficientes de Wavelets	Aproximação	1154,1	4326,0
	Diferenças verticais	-536,8	599,0
	Diferenças horizontais	-538,3	464,3
	Diferenças diagonais	-511,1	444,7

Pela diversidade de domínios entre os distintos atributos, algoritmos que consideram a semelhança entre dois registros como sendo a soma das distâncias de cada atributo (por exemplo, o *k-means* quando utilizada a distância euclidiana), poderiam considerar apenas os coeficientes de *wavelets* como atributos dominantes. A solução nestes casos é a normalização dos valores.

Existem diversas formas possíveis de normalizar dados. Neste trabalho optou-se pela normalização mínimo-máximo [18], por ser uma das técnicas mais simples. O resultado é um valor entre 0 e 1 obtido através da seguinte função: sejam *min* e *max* o menor e o maior valores do atributo (ou conjunto de atributos), respectivamente, e *val* o valor real, o valor normalizado *norm* é calculado por:

$$norm = \frac{val + |min|}{max + |min|}$$

4.3 Considerações finais do capítulo

Conforme apresentado na Seção 2.3, minerar imagens envolve uma série de decisões. Especialmente o processamento da imagem necessita a definição de como serão formados os registros e quais serão seus atributos, que devem descrever o conteúdo da imagem de forma relevante para o domínio de problema sendo tratado.

Neste trabalho optou-se por dividir a imagem batimétrica em células quadradas em pixels, sendo cada uma utilizada como registro. Para a extração de atributos, dispõe-se de algumas técnicas bem conhecidas, como estatísticas dos pixels, histograma de cores e coeficientes de *Wavelets*. Além destes, foram propostos outros atributos para capturar o formato aproximado das células: regiões e vetor.

Quanto ao mapa batimétrico utilizado, naturalmente deve-se utilizar uma projeção do globo terrestre que possa ser descrita com células quadradas, e que permita o mapeamento de coordenadas latitude - longitude para pixels na imagem. Neste caso, a imagem utilizada é de um mapa na projeção geográfica. Esta projeção considera os meridianos e paralelos como perpendiculares entre si, e preserva as distâncias entre os mesmos. Entretanto, ela cria uma distorção que se agrava à medida que as células se distanciam da linha do Equador. Esta distorção pode influenciar no desempenho dos algoritmos de mineração, visto que células semelhantes podem estar tão distorcidas que não sejam reconhecidas como tal. Entretanto, neste trabalho não é realizado nenhum procedimento de compensação dessa distorção, o que pode ser um tema de pesquisa futuro.

5. VISUALIZAÇÃO DO PROCESSO DE MINERAÇÃO EM IMAGENS BATIMÉTRICAS

O processo de descoberta de conhecimento em imagens batimétricas pode naturalmente utilizar a própria imagem original como base para a exibição de dados e resultados. Em comparação à apresentação de dados tabulares, a apresentação gráfica pode facilitar o entendimento do especialista de domínio. Mais do que isso, a apresentação dos resultados através de uma forma visual intuitiva e consistente pode fazer a diferença para que o especialista de domínio possa compreender os achados da mineração de dados. Sendo assim, é proposta uma forma de representação gráfica para o processo de mineração de dados em imagens batimétricas. Os exemplos aqui apresentados seguem o formato esperado da imagem batimétrica descrito na Seção 4.1.

5.1 Colorização das células

Ao analisar o mapa como um todo, são necessários elementos que facilitem o entendimento global dos dados do mapa. Inicialmente, as células são delimitadas no mapa através de linhas contínuas. Cada classe ou grupo é representado por uma cor. É importante diferenciar os dados reais dos dados induzidos. Assim, para células previamente classificadas são utilizadas cores com menor transparência (resultando em um efeito mais saturado), enquanto que para classes ou grupos atribuídos automaticamente são utilizadas cores com maior transparência. Células que foram descartadas por possuírem terra seca não são modificadas. A Figura 5.1 mostra exemplos de cada uma dessas situações.

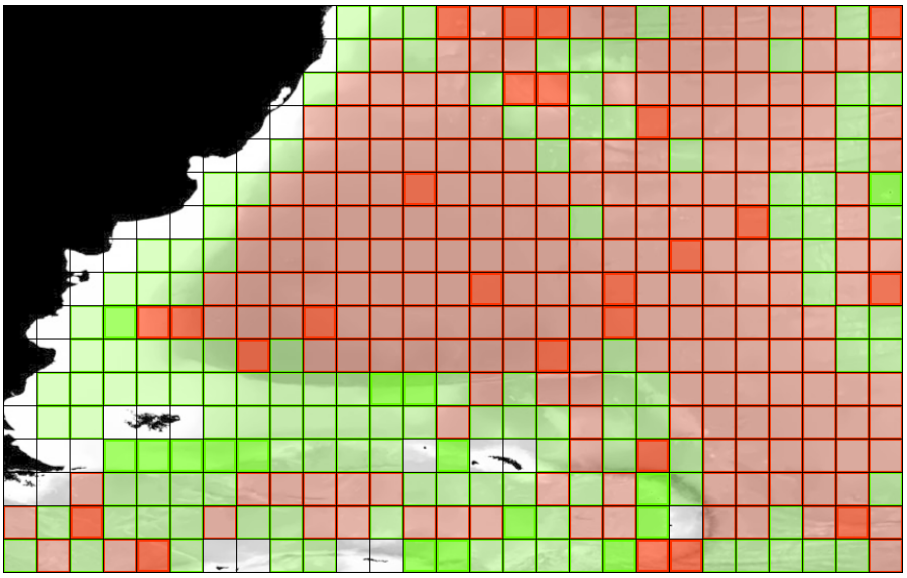
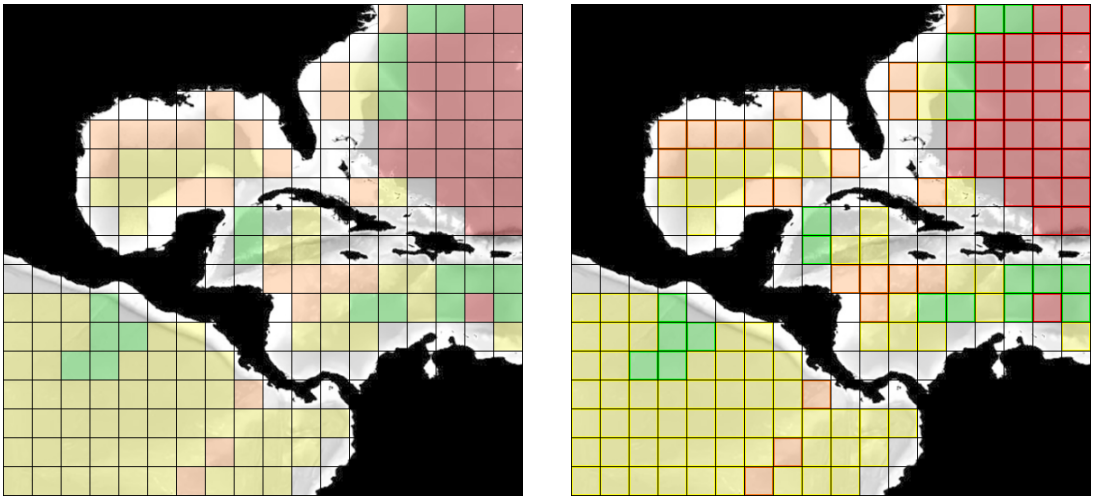


Figura 5.1: Delimitação de células e colorização de acordo com as classes / grupos

Ao aplicar a transparência a cor original da classe ou grupo, a célula pode ser exibida em cores mais claras quando a célula é dominantemente rasa ou mais escuras, quando a célula é dominantemente profunda. Quando são utilizadas cores próximas, como vermelho, laranja e amarelo, poderia ser dificultada a diferenciação visual entre células destas cores. Para criar um maior contraste e delimitar de forma mais visível as células, é também desenhada uma borda interna com a cor sólida da classe ou grupo, que é levemente mais espessa em células previamente classificadas. A Figura 5.2 mostra a mesma área com ou sem bordas internas, evidenciando que na Figura 5.2b fica mais visível a diferença entre células de distintas classes / grupos em relação à Figura 5.2a.



(a) Sem a utilização de bordas internas (b) Com a utilização de bordas internas

Figura 5.2: Diferenças entre a utilização ou não de bordas internas

5.2 Iconografia para representação dos atributos das células

Também é útil representar os atributos extraídos de cada célula, e não somente sua classe ou grupo. Dos atributos extraídos das células (conforme descrito na Seção 4.2.2), o mais evidente para ser desenhado em cada célula no mapa é o vetor. Sua iconografia será descrita a seguir.

5.2.1 Vetores

O vetor possui 3 componentes: módulo, ângulo xy e ângulo zy. Para sua representação no plano, dois desses componentes são naturalmente representáveis: o módulo, através do comprimento do vetor e o ângulo xy, dado pela rotação do vetor. Entretanto, para representar o ângulo zy, foi utilizada uma nova abordagem: a espessura do traço e o tamanho da seta do vetor. Como não existem valores de referência na imagem, apenas proporções (seria necessária uma informação externa à imagem para uma escala com as magnitudes referentes a cada nível de cinza), estes elementos são discretizados para 4 níveis. Sendo A a média e σ o desvio padrão do ângulo zy em todo o conjunto de dados, as faixas são: $-\infty$ a A - 2σ, A - 2σ a A - σ, A - σ a $A + \sigma$ e $A + \sigma$ a ∞.

Para não prejudicar o entendimento, quando o módulo do vetor é menor que o tamanho da seta, a linha não é desenhada. A Figura 5.3 apresenta os mais variados tipos de vetores. Esta figura também apresenta algumas células, demarcadas em amarelo, cujos vetores não foram desenhados por serem consideradas totalmente planas, segundo o limiar descrito na Seção 4.2.2. As demais células que não possuem o vetor desenhado contêm ao menos um pixel totalmente preto, e, portanto, foram descartadas.

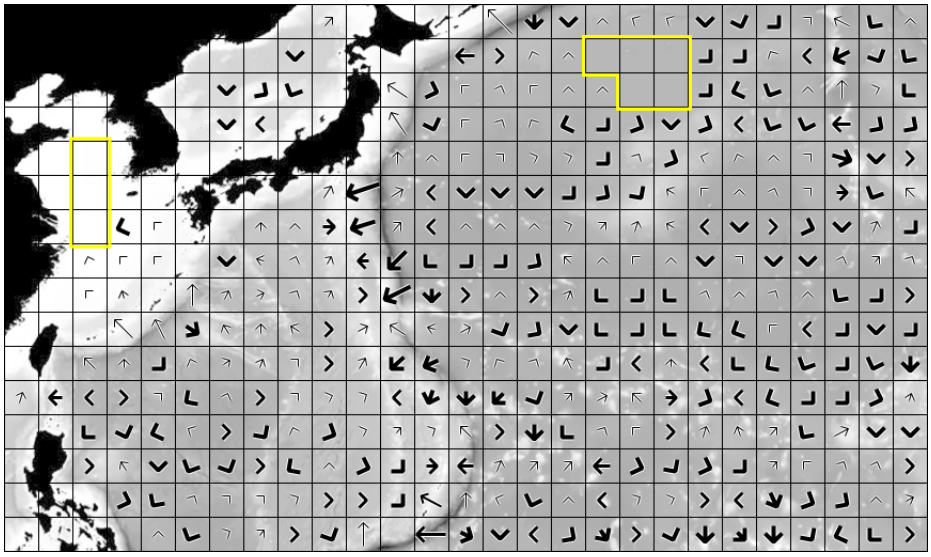
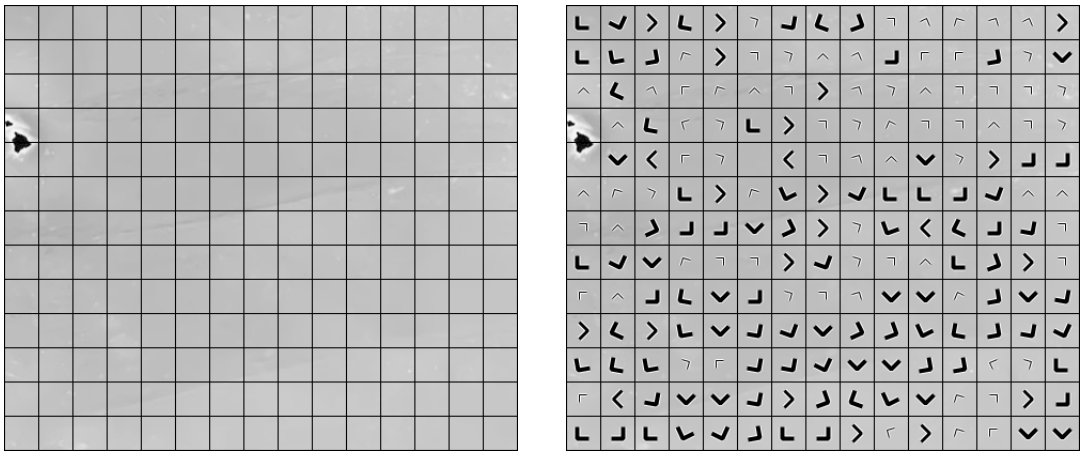


Figura 5.3: Distintos tipos de vetores representados no mapa

O vetor auxilia no entendimento da direção dominante da célula, especialmente em casos onde a variação é sutil. A Figura 5.4a apresenta uma área do mapa desenhada sem os vetores, com diversas células contendo pouca aclividade e pouca variação entre elas. Já a Figura 5.4b apresenta as mesmas células com seus respectivos vetores. Ficam evidentes as diferenças entre células visualmente parecidas, mas que podem possuir diferenças significativas no que se refere à direção e à inclinação dominantes.



(a) Sem a utilização dos vetores

(b) Com a utilização dos vetores

Figura 5.4: Diferenças entre a utilização ou não de vetores

5.2.2 Outros conjuntos de atributos

Apesar deste trabalho explorar somente a representação gráfica dos vetores, também outros atributos poderiam ser desenhados no mapa para facilitar o entendimento em relação a eles. Como temas de pesquisas futuras, poderiam ser avaliados os elementos gráficos para os seguintes atributos:

- Atributos estatísticos: desenho de uma curva normal sobre cada célula, destacando a média e o desvio padrão;
- Histograma: o histograma pode ser desenhado como um gráfico de barras em cada célula;
- Regiões: as regiões pode ser demarcadas no interior de cada célula, por exemplo, com cores mais escuras para representar as regiões mais profundas e cores mais claras para representar regiões mais rasas;

Já para os coeficientes de *Wavelets*, dificilmente haveria alguma iconografia intuitiva que possa ser desenhada sobre cada célula.

5.3 Considerações finais do capítulo

Tradicionalmente, a mineração de dados é aplicada sobre dados tabulares, sendo que todo o processo de preparação de dados e de apresentação de resultados também ocorre neste formato. Entretanto, como neste trabalho são processadas imagens batimétricas, o formato tabular não é o mais natural para a exibição de dados e resultados para o usuário.

Uma forma de visualização intuitiva e consistente dos resultados da mineração de dados em imagens batimétricas é muito importante para o entendimento dos achados. Por exemplo, a percepção de proximidade entre as células, suas classes iniciais ou inferidas e seus atributos não ficariam explícitas ao analisar um resultado tabular, que provavelmente contenha todos os atributos de cada célula, além de seu rótulo de classe.

Neste capítulo foram descritos e discutidos alguns dos aspectos para a construção de uma iconografia. Aspectos como a diferenciação através de cores para classes iniciais ou atribuídas, bem como de grupos, e a representação iconográfica de atributos extraídos das células, especialmente o vetor podem contribuir para o entendimento da imagem batimétrica e dos resultados da mineração de dados.

6. TESTE DA ABORDAGEM

Com o objetivo de validar a proposta deste trabalho, de mineração de dados em imagens batimétricas, foi desenvolvido um protótipo do ambiente de software.

6.1 Funcionalidades implementadas

O protótipo tem como base os requisitos descritos na Seção 3.1, e são implementadas as seguintes funcionalidades:

- Seleção da imagem batimétrica:
  - Importar uma nova imagem batimétrica, selecionando parâmetros para a importação, como o tamanho da célula e número de regiões. Mais detalhes na Seção 6.4.2;
  - Abrir uma imagem batimétrica previamente importada. Ver Seção 6.4.3;
- Para a tarefa de classificação (mais detalhes na Seção 6.4.4):
  - Criar / modificar / remover um rótulo de classe;
  - Atribuir um rótulo de classe a uma célula individualmente ou a um conjunto de células, através de uma importação de arquivo;
  - Executar a tarefa de classificação, selecionando o algoritmo e seus parâmetros, bem como quais atributos serão utilizados;
  - Exibir os resultados de uma classificação previamente executada.
- Para a tarefa de análise de agrupamentos (mais detalhes na Seção 6.4.5):
  - Executar a tarefa de análise de agrupamentos, selecionando o algoritmo e seus parâmetros, bem como quais atributos serão utilizados;
  - Exibir os resultados de uma análise de agrupamentos previamente executada;
  - Transformar os grupos encontrados pelo algoritmo em classes.

6.2 Modelo de dados

Para não restringir o tempo de vida dos dados obtidos e gerados durante a operação do ambiente de software a uma única execução, todas as operações são persistidas em uma base de dados. Foi definido um modelo de dados que contempla as funcionalidades descritas acima, apresentado na Figura 6.1.

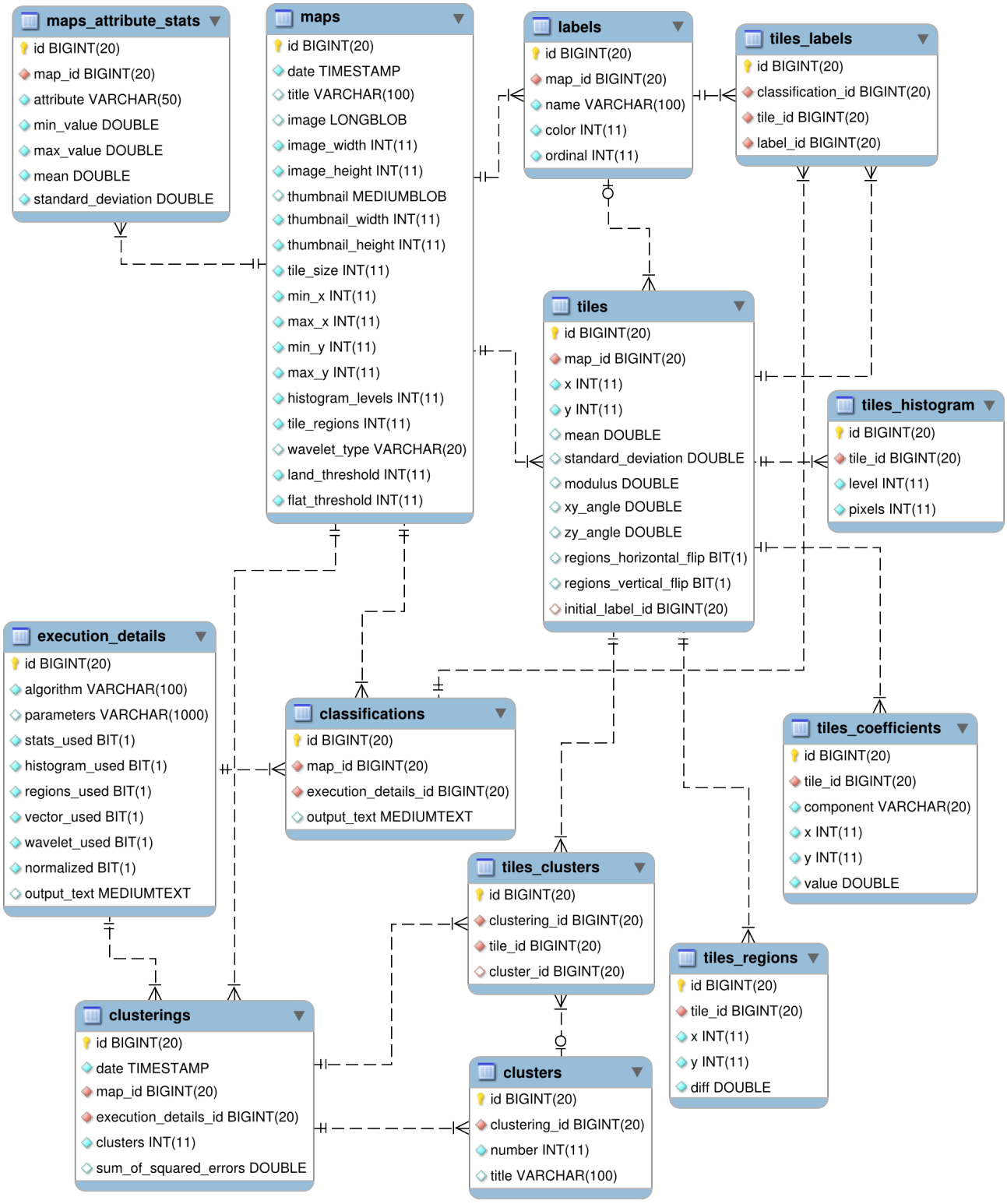


Figura 6.1: Modelo de dados, na notação da ferramenta MySQL Workbench

A Tabela 6.1 descreve em linhas gerais o que é armazenado em cada uma das tabelas do modelo de dados.

Tabela 6.1: Descrição do modelo de dados

Tabela	**Descrição**
maps	Armazena os dados gerais dos mapas, como a imagem, o tamanho das células, os intervalos de células importadas, o número de entradas no histograma, o número de regiões e o tipo de *Wavelet*.
tiles	Armazena os dados gerais das células, como suas coordenadas (x, y), os atributos estatísticos e os do vetor, bem como a classe manualmente atribuída.
labels	Contém as classes. Cada classe pertence a um único mapa.
maps_attribute_stats	Armazena estatísticas sobre os valores de um tipo de atributo de um determinado mapa: média, desvio padrão, mínimo e máximo.
tiles_histogram	Contém as entradas do histograma para cada célula.
tiles_regions	Armazena os dados das regiões, como a célula, a coordenada (x,y) (relativa à celula) e o valor.
tiles_coefficients	Armazena o valor dos coeficientes dos wavelets para cada célula.
execution_details	Armazena os parâmetros comuns da execução de uma tarefa de mineração de dados, como os atributos utilizados, o algoritmo e seus parâmetros e se os valores foram ou não normalizados.
classifications	Contém os dados de uma execução da tarefa de classificação, basicamente as chaves estrangeiras para as tabelas *maps* e *execution_details*.
tiles_labels	Contém, para cada classificação, a classe atribuída a cada célula.
clusterings	Contém os dados de uma execução da tarefa de análise de agrupamentos, basicamente as chaves estrangeiras para as tabelas *maps* e *execution_details*, além do número de agrupamentos gerados pelo algoritmo.
clusters	Contém cada um dos grupos gerados pelo algoritmo de análise de agrupamentos.
tiles_clusters	Armazena, para uma execução de análise de agrupamentos, o grupo atribuído a cada célula.

6.3 Detalhes de implementação

O protótipo foi desenvolvido com a linguagem de programação Java[1], utilizando as seguintes ferramentas (todas software livre): Weka[2], um ambiente de aprendizagem de máquina e mineração de dados que disponibiliza uma API para uso de seus algoritmos; JWave[3], biblioteca que implementa diversos tipos de *Wavelets* e funções como a DWT; e QueryDSL[4], biblioteca que facilita a manipulação da base de dados e execução de consultas SQL. Para a persistência dos dados, foi utilizado o SGBD MySQL[5].

6.4 O protótipo em operação

A seguir são apresentadas as capturas de tela de cada funcionalidade descrita na Seção 6.1.

6.4.1 Janela principal

A Figura 6.2, apresenta a janela principal do protótipo, que possui, além da imagem do mapa, um painel lateral com as informações e operações possíveis e uma barra de status abaixo. As informações e operações são:

- Importar uma nova imagem batimétrica;
- Carregar uma imagem batimétrica previamente importada;
- Informações gerais sobre o mapa atual;
- Opções de visualização: nível de zoom e seleção de elementos gráficos que serão desenhados sobre o mapa (grade delimitando as células e vetores);
- Opções para exportação: exportar os dados como ARFF para análise na ferramenta Weka e exportar a imagem completa, exatamente como exibida (incluindo os elementos gráficos adicionais);
- Aplicação de tarefas de mineração de dados: classificação (conforme descrito na Seção 6.4.4) e análise de agrupamentos (conforme descrito na Seção 6.4.5).

[1] http://www.java.com
[2] http://www.cs.waikato.ac.nz/ml/weka/
[3] http://code.google.com/p/jwave/
[4] http://www.querydsl.com
[5] http://www.mysql.com

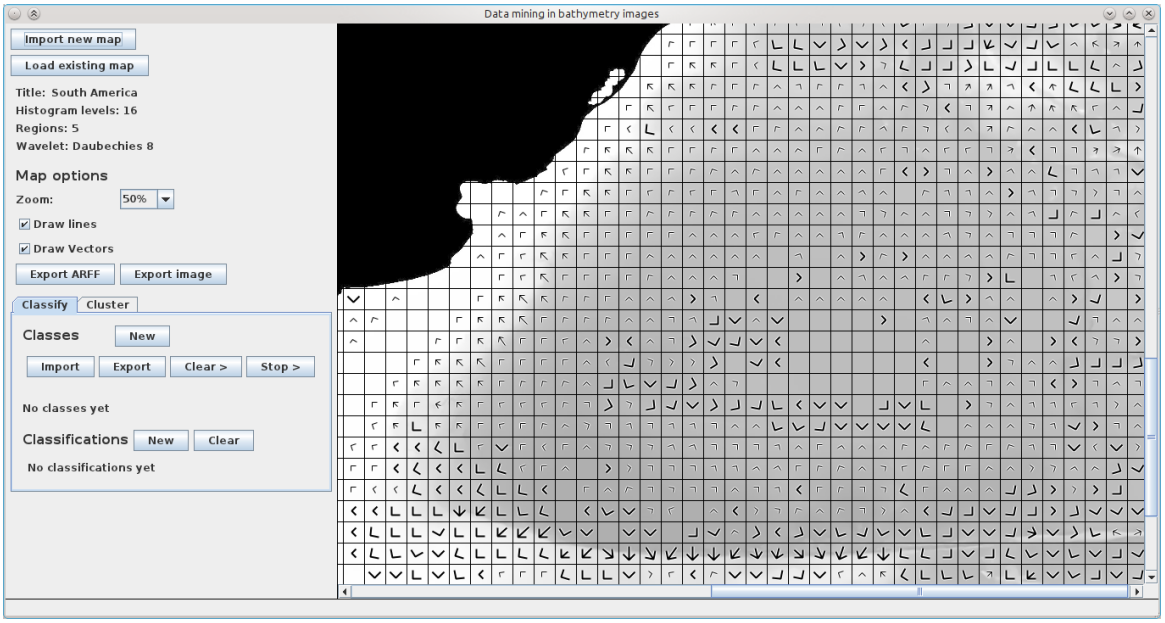


Figura 6.2: Protótipo: janela principal

6.4.2 Importação de uma imagem batimétrica

A Figura 6.3 mostra a janela de importação de uma imagem batimétrica. Além da seleção do arquivo de imagem, no topo da tela e da previsão da imagem no canto esquerdo, são informados os parâmetros da importação:

- Tamanho da célula, em pixels;
- Célula inicial a ser considerada no eixo x;
- Célula final a ser considerada no eixo x;
- Célula inicial a ser considerada no eixo y;
- Célula final a ser considerada no eixo y;
- Limiar de terra: número de pixels totalmente pretos (terra seca) para que a célula seja ignorada;
- Limiar do plano: módulo do vetor para que a célula seja considerada totalmente plana;
- Número de entradas no histograma;
- Número de regiões na célula;
- Tipo de *Wavelet* a ser utilizado.

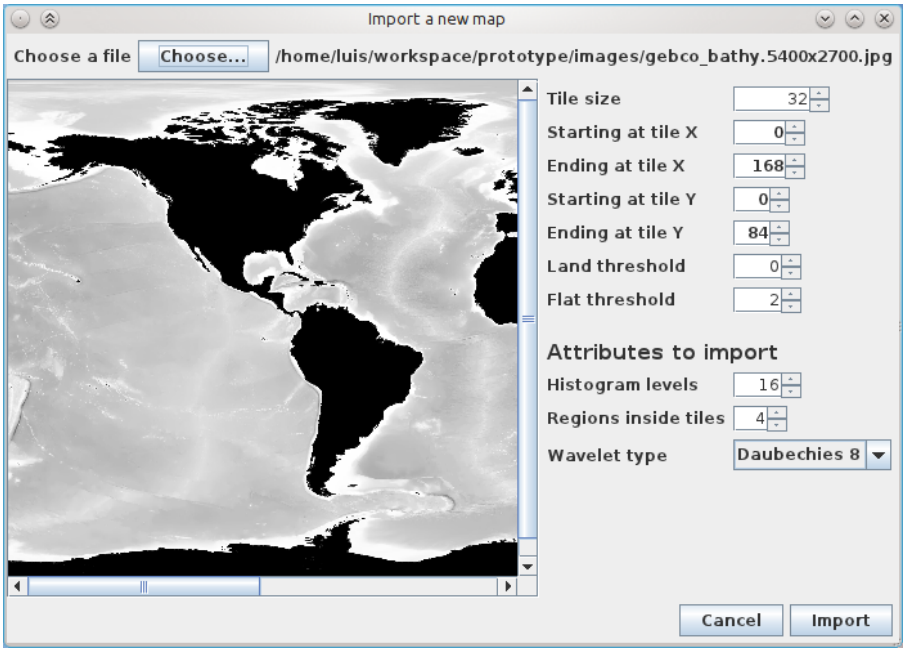


Figura 6.3: Protótipo: importação de uma imagem batimétrica

6.4.3 Carga de imagem batimétrica previamente importada

Na Figura 6.4, pode-se observar a janela de carga de uma imagem batimétrica previamente importada. São exibidas informações gerais sobre cada imagem, como a data e hora da importação, o tamanho da célula, os intervalos de células selecionados, o número de entradas no histograma, o número de regiões, e o tipo de *Wavelet*.

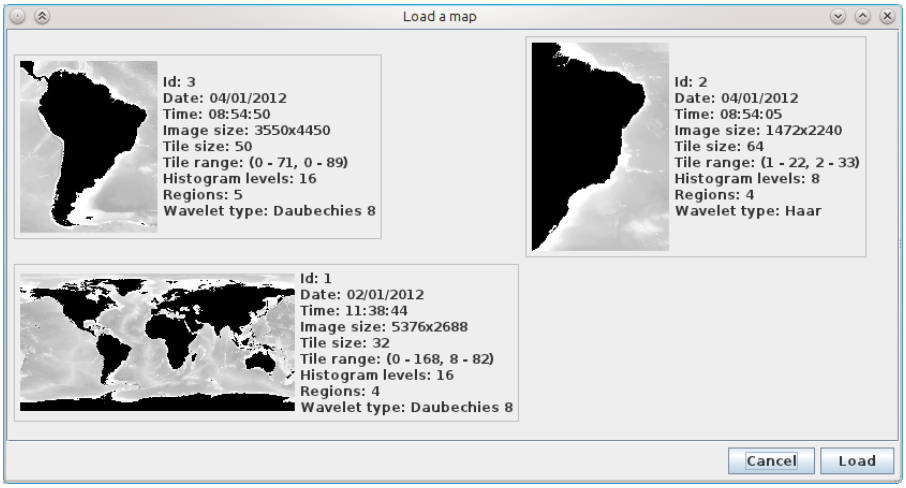


Figura 6.4: Protótipo: carga de imagem batimétrica previamente importada

6.4.4 Execução da tarefa de classificação

Para executar a classificação, inicialmente deve-se, através das operações disponíveis na seção *Classes*, criar os rótulos de classe disponíveis, informando o nome e a cor de cada um. Após, pode-se selecionar a seta ao lado da classe para manualmente atribuí-la, com o mouse, a células no mapa.

Alternativamente, é possível importar um arquivo contendo as coordenadas (x,y) da célula e a classe correspondente, para definir quais as células são previamente classificadas. Pode-se então executar a classificação, selecionando na janela auxiliar o algoritmo a ser executado, seus parâmetros (os mesmos algoritmos e parâmetros possíveis na ferramenta Weka), quais atributos serão usados na classificação e se os atributos serão ou não normalizados.

Também é exibida uma lista com as classificações já executadas, contento o nome do algoritmo executado e quais atributos foram utilizados, representados pelas seguintes letras: S (atributos estatísticos), H (histograma), R (regiões), V (vetor) e W (coeficientes de *Wavelets*), além de apresentar a letra N quando os atributos foram normalizados.

A Figura 6.5 exibe a janela principal com o resultado de uma execução de classificação, bem como a janela auxiliar para executar uma nova classificação. No exemplo, foram criadas 3 possíveis classes, *Low* (vermelho), *Medium* (amarelo) e *High* (verde). Algumas células foram manualmente rotuladas. Através de uma execução de classificação, as outras células foram classificadas. A colorização das células segue as definições da Seção 5.1, utilizando cores mais saturadas e bordas internas mais espessas em células manualmente rotuladas, e cores menos saturadas e bordas mais finas nas células que tiveram sua classe inferida.

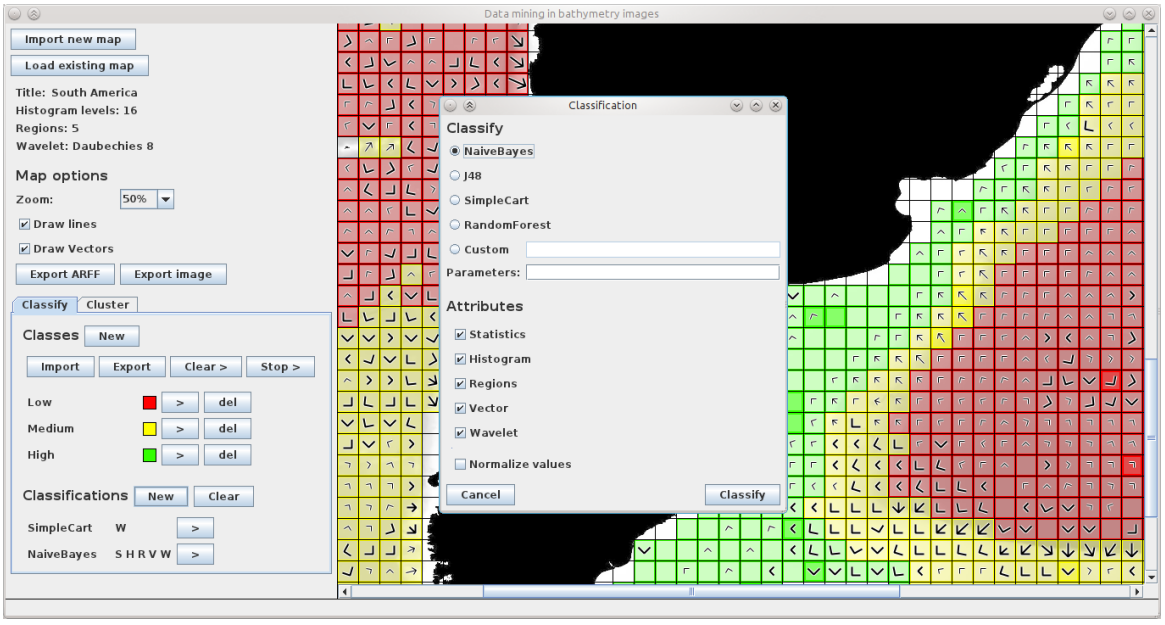


Figura 6.5: Protótipo: execução da classificação

6.4.5 Execução da tarefa de análise de agrupamentos

Ao iniciar a tarefa de análise de agrupamentos, é exibida uma janela auxiliar bastante semelhante à da classificação, porém contendo como opções algoritmos de análise de agrupamentos. Além do algoritmo e seus parâmetros, seleciona-se quais atributos serão usados na análise de agrupamentos e se os mesmos serão normalizados. Também é exibida uma lista com os agrupamentos já realizados,

contendo o nome algoritmo e quais atributos foram utilizados, da mesma forma como na classificação (descrita na Seção 6.4.4).

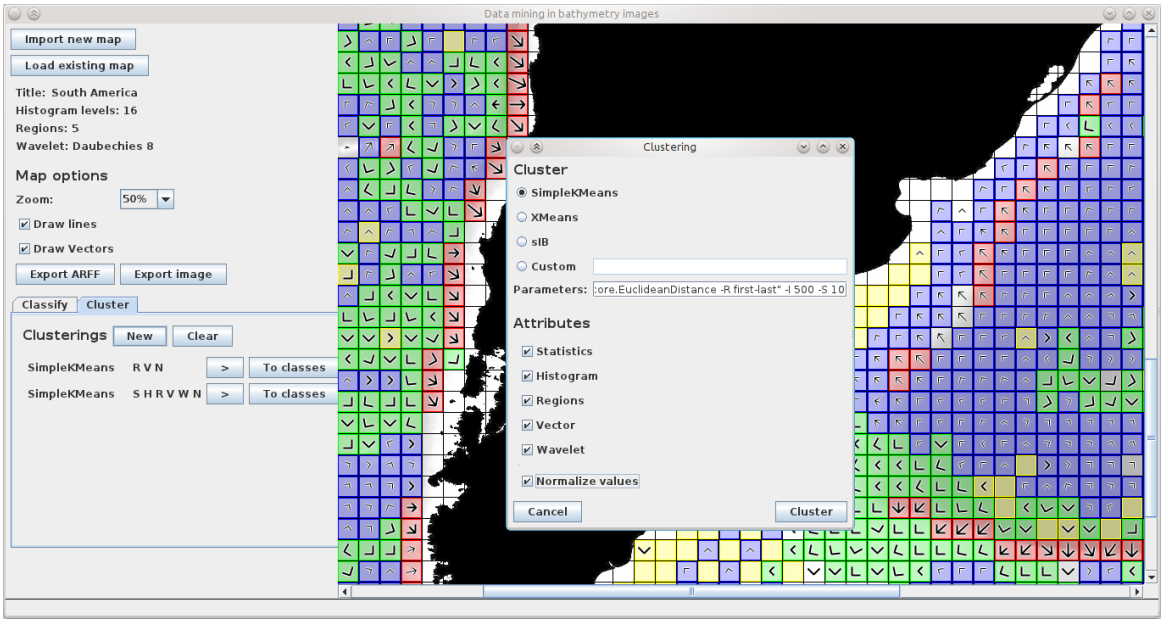


Figura 6.6: Protótipo: execução da análise de agrupamentos

A Figura 6.6 mostra a janela principal com o resultado de uma análise de agrupamentos e também a janela auxiliar para executar uma nova análise de agrupamentos.

6.5 Considerações finais do capítulo

Neste capítulo foi descrito o desenvolvimento de um protótipo que implementa a análise de imagens batimétricas através da mineração de dados, com o objetivo de validar a abordagem proposta no Capítulo 4. Esta abordagem define o formato esperado da imagem batimétrica, bem como as técnicas utilizadas para a preparação dos dados (extração de registros e atributos). Assim é obtido um conjunto de dados que descreve a imagem batimétrica em questão, porém no formato esperado pelos algoritmos clássicos de mineração de dados.

Sendo a análise realizada sobre imagens batimétricas, foram também implementadas as técnicas de visualização do processo de mineração descritas no Capítulo 5. Entende-se que para que um especialista de domínio possa utilizar o sistema de forma proveitosa, que facilite o entendimento dos dados representados e dos resultados da mineração, é fundamental o emprego de uma visualização consistente e intuitiva.

As funcionalidades implementadas pelo protótipo foram as que considera-se importantes para que um especialista de domínio explore o mapa batimétrico, e obtenha a referida análise computacional automática (através da análise de agrupamentos) ou semi-automática (através da classificação). Ambas as operações são possíveis no protótipo, através de algoritmos de mineração de dados implementados pela ferramenta Weka.

Adicionalmente, foi implementada no protótipo a possibilidade de seleção dos atributos extraídos que serão utilizados em cada tarefa, permitindo a experimentação por parte do usuário. O mesmo pode, não somente selecionar e parametrizar o algoritmo de mineração de dados, mas também selecionar quais atributos serão utilizados, permitindo uma diversificação de resultados.

O objetivo do protótipo não é replicar todo o ambiente do Weka, com todas as suas opções. Entretanto, é disponibilizada a opção para exportar o conjunto de dados para o Weka, permitindo a utilização do mesmo para a análise dos dados, de forma complementar àquilo que é oferecido no protótipo.

7. AVALIAÇÃO DA ABORDAGEM

Para avaliar os resultados da mineração de dados em imagens batimétricas, foi solicitada a ajuda de uma especialista de domínio, uma oceanógrafa. Como o foco deste trabalho é a tarefa de classificação, buscou-se uma base de dados que possua informações sobre a classificação de algumas células do mapa a fim de, por mineração de dados, classificar as outras células sobre as quais não se tem informações.

7.1 Base de dados de corais de águas profundas

Foi proposto pela oceanógrafa uma base de dados sobre a presença de corais de águas profundas, que é descrita a seguir.

7.1.1 Introdução

Os corais de águas profundas foram descobertos no século XVIII. Porém, apenas recentemente despertaram real interesse e a sua pesquisa avançou de forma significativa. O interesse está relacionado ao seu papel de geradores de habitats para comunidades de peixes e por seu papel como importantes registros paleoceanográficos de alta resolução [33].

Para a indústria pesqueira, o conhecimento da distribuição e do funcionamento dos ecossistemas associados aos corais de águas profundas tornou-se relevante após o declínio dos estoques costeiros. Já, como registros paleoceanográficos, os corais são importantes por incorporarem nos seus esqueletos diferentes proporções de elementos químicos e de isótopos que variam ao longo do seu crescimento, acompanhados de variações climáticas e oceanográficas. Desta forma, os esqueletos de corais constituem valiosos arquivos de informações paleoclimáticas e paleoceanográficas.

Os corais de águas profundas são comuns em montes submarinos e no talude das margens continentais e ilhas [30]. A sua distribuição está intimamente relacionada à geologia já que esta condiciona o tipo de substrato e influencia na configuração das correntes. A relação entre o tipo de substrato e o tipo de coral que pode se desenvolver é tão estreita que os registros de ocorrência de diferentes espécies de corais podem ser utilizados como indicadores do tipo de substrato [19].

Embora o ecossistema de mar profundo seja um dos mais extensos do planeta cobrindo em torno de 60% da superfície sólida da Terra (considerando mar profundo como aquele com mais de mil metros de profundidade), o conhecimento deste ambiente é limitado quando comparado a outros ecossistemas marinhos [15]. Desta forma, o conhecimento da biogeografia de espécies de águas profundas ainda é muito limitado. Por exemplo, os mapas globais de distribuição dos corais de águas profundas, sugerem uma maior densidade destes no Atlântico Norte. No entanto, isto muito provavelmente seja um artefato relacionado ao esforço amostral, uma vez que o Atlântico Norte é a bacia mais estudada e densamente amostrada.

Assim, a base de dados de corais de águas profundas torna-se um excelente exemplo para a aplicação do método de mineração de dados em imagens batimétricas, visto a dificuldade da amostragem de dados. Outros trabalhos (por ex. [10, 11, 35]) têm estimado a distribuição de corais utilizando uma gama de parâmetros mais extensa, não apenas a batimetria, mas também parâmetros de oceanografia física, química e biológica.

No entanto, a principal contribuição deste trabalho é a apresentação de uma abordagem inovadora utilizando-se como entrada apenas a imagem batimétrica. Caso necessário, os resultados do método podem ser combinados com outros bancos de dados externos, enriquecendo-os ainda mais.

7.1.2 Descrição dos dados e preparação para a mineração

A base de dados original utilizada neste trabalho é da distribuição de corais de águas profundas de Rogers et al. [32]. Ela foi extraída do *Ocean Biogeographic Information System*[1]. A base contém 6.553 registros obtidos entre os anos de 1869 e 2005. Cada registro contém as coordenadas latitude-longitude com a respectiva espécie de corais encontrada naquele ponto.

Para aplicar esta base de dados neste trabalho, foram definidas 2 classes: *Sim* e *Não*, indicando a presença de alguma espécie de coral na célula. Através das coordenadas de latitude e longitude, os dados originais foram mapeados para suas respectivas células no mapa, sendo que àquelas que possuem ao menos uma espécie de coral foi inicialmente atribuída a classe *Sim*.

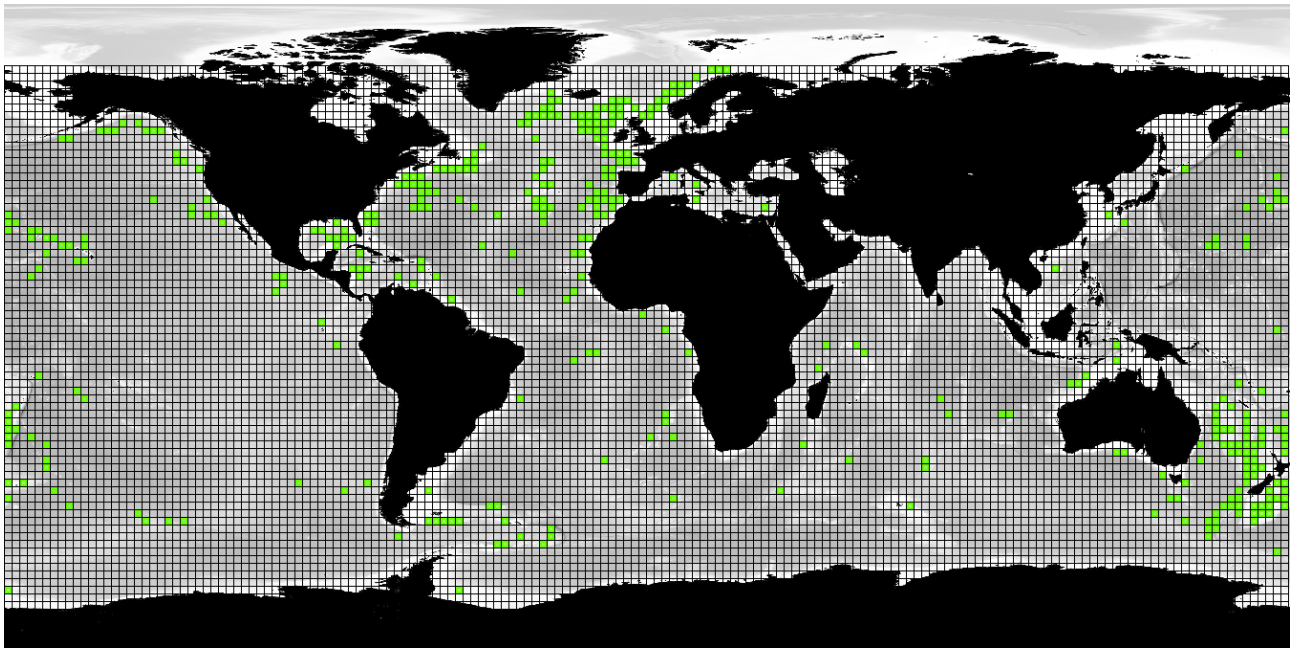


Figura 7.1: Mapeamento da base de dados de corais

Como a base de dados é global, foi utilizada a imagem original, sem recortes, na resolução de 5.400 $\times$ 2.700 pixels. Foi utilizado um tamanho de célula de 32 $\times$ 32 pixels. Como nos dados originais não constavam quaisquer registros a respeito da região ártica, foram apenas consideradas

[1] `http://www.iobis.org`

células a partir da primeira linha (no sentido norte-sul) que possuía dados. Tendo em vista que também células que possuem pixels pretos são descartadas, resultaram 7.153 células válidas, das quais 379 foram inicialmente classificadas como *Sim*. Pelos dados originais, 199 células que contêm terra seca, e portanto, descartadas, possuem alguma espécie de coral. A Figura 7.1 destaca no mapa as células que foram inicialmente classificadas como *Sim*.

Na base de dados não constam informações sobre locais que não contêm corais. Portanto, não há a informação se células não mapeadas contêm ou não corais. Para a correta execução da classificação, é fundamental que também constem na base de dados células inicialmente classificadas como *Não*. Para resolver essa situação foi aplicada uma heurística: das células não classificadas, foi aleatoriamente selecionado o mesmo número de células classificadas como *Sim*, e a elas foi atribuída a classe *Não*. Assim, o conjunto de treino manteve o balanceamento das classes.

7.1.3 Aplicação da tarefa de classificação

Para determinar o algoritmo a ser utilizado, foram selecionados os principais algoritmos de classificação dos modelos apresentados na Seção 2.2.1: o C4.5, o CART, o Random Forest, o naïve Bayes e as redes Bayesianas. Entretanto, o algoritmo de redes Bayesianas não pôde ser executado sobre o conjunto de dados por requerer mais memória RAM do que o disponível (em um computador com 4GB de memória). Portanto, este algoritmo não foi considerado. O desempenho de cada algoritmo foi medido pelo método de validação cruzada (*cross validation*) com 10 partições. O resultado sumarizado está descrito na Tabela 7.1.

Tabela 7.1: Precisão de cada execução

Algoritmo	**Precisão Geral**	**Precisão do Sim**	**Precisão do Não**
C4.5	65,17%	64,64%	65,70%
CART	72,03%	74,93%	69,13%
Random Forest	71,11%	77,84%	64,38%
Naïve Bayes	71,64%	69,13%	74,14%

Além da melhor precisão geral, um critério importante para a seleção de um algoritmo para o problema abordado é a minimização dos falsos positivos. Isto se deve a que nos dados originais constam somente locais que contêm corais. Assim, minimizando-se os falsos positivos aumenta-se a chance de encontrar corais em outros locais não mapeados.

Pelos critérios estabelecidos, o algoritmo que apresentou o melhor desempenho foi o Random Forest. A Figura 7.2 apresenta o mapa na iconografia descrita no Capítulo 5, onde as células de cor mais saturadas são aquelas previamente classificadas, e as menos saturadas são as inferidas pelo algoritmo. A cor verde foi utilizada para representar a classe Sim (presença de corais) e a cor vermelha para representar a classe Não (ausência de corais).

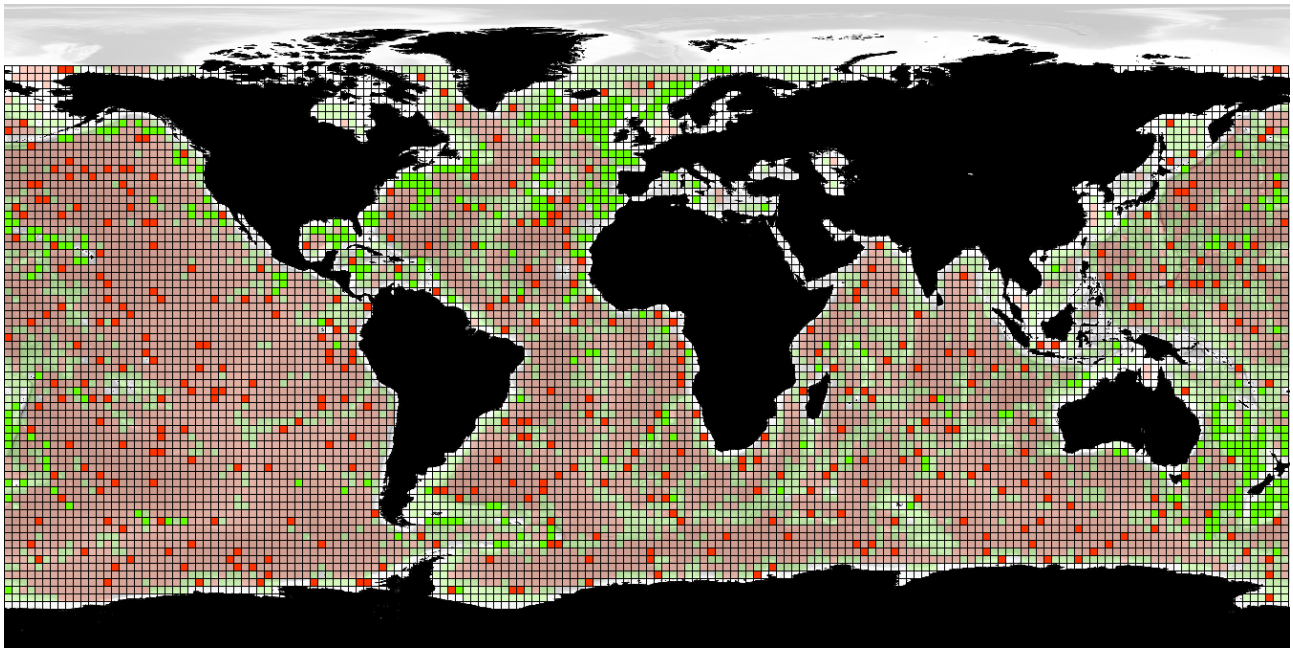


Figura 7.2: Resultado da aplicação do algoritmo Random Forest sobre a base de dados de corais de águas profundas

7.1.4 Depoimento da oceanógrafa

A fim de constatar a validade dos resultados obtidos através da aplicação do algoritmo de classificação na base de dados de corais de águas profundas, foi solicitado à oceanógrafa que analisasse os resultados e emitisse seu parecer sobre os mesmos. O seu depoimento é apresentado a seguir.

"A distribuição prevista mostra um padrão coerente com o esperado, tanto em termos de distribuição geográfica como de localização em relação à fisiografia do fundo oceânico.

"Ao compararmos a distribuição comprovada (locais onde foram registrados corais) com a distribuição prevista, observa-se que a segunda apresenta uma distribuição das ocorrências mais homogênea, diferentemente da primeira que sofre o claro efeito da concentração do esforço amostral no Atlântico Norte.

"Já do ponto de vista do efeito da fisiografia, sabe-se que a distribuição dos corais está intimamente relacionada à geologia, visto que esta condiciona o tipo de substrato e o relevo, influenciando assim a configuração das correntes. Por tal motivo, os corais costumam ser comuns em montes submarinos e no talude das margens continentais e ilhas [30] o que aparece bem representado no mapa de previsão da distribuição.

"Além da geologia e da configuração do relevo, existem outros parâmetros de oceanografia física, química e biológica que condicionam a distribuição dos corais de águas profundas. No entanto, os resultados obtidos neste trabalho a partir da carta batimétrica são comparáveis àqueles obtidos em abordagens mais abrangentes que consideram diversos parâmetros ambientais. Por exemplo, os resultados obtidos por Davies e Guinotte [10], embora tenham uma maior resolução espacial, mostram um padrão de distribuição semelhante ao obtido neste trabalho. As principais diferenças com o trabalho de Davies e Guinotte ocorrem nas altas latitudes e no Pacífico Norte oriental, ao

longo da margem oeste da América do Norte - regiões para as quais o trabalho dos autores não prevê a ocorrência de corais. No entanto, ao menos no caso do Pacífico NE, os resultados obtidos neste trabalho parecem consistentes, uma vez que realmente existem registros da ocorrência de corais nesta região".

7.2 Considerações finais do capítulo

Este trabalho propõe uma abordagem (Capítulo 4) para a análise de imagens batimétricas através da mineração de dados. Para validar a proposta e possibilitar a experimentação por parte do usuário na construção de uma solução, foi desenvolvido um protótipo (Capítulo 6).

Entretanto, para efetivamente validar a proposta, é necessária uma base de dados real. Então, buscou-se junto a uma especialista de domínio, uma oceanógrafa, uma base de dados com áreas do mapa previamente classificadas. Assim, através da tarefa de classificação, outras áreas do mapa, sobre as quais não se tinha informação, podem ter sua classe inferida.

Para a validação, foi utilizada uma base de dados da distribuição de corais de águas profundas, de Rogers et al. [32]. Esta base de dados é composta de registros de espécies de corais e as respectivas coordenadas (latitude e longitude) de onde estas sabidamente são encontradas.

No formato esperado, segundo a abordagem proposta, a granularidade de registro é definida por uma célula no mapa (Seção 4.2.1). Cada célula pode conter somente uma classe. Entretanto, na base de dados original podem haver diversas espécies de corais de águas profundas mapeadas para a mesma célula. Assim, a espécie do coral não pode ser utilizada como classe.

Para a utilização dessa base de dados foram consideradas 2 classes, *Sim* e *Não*, indicando a presença ou ausência de corais na célula. O conjunto de dados original foi mapeado para as células correspondentes. Aquelas que continham ao menos um registro de coral foram inicialmente classificadas como *Sim*. Uma parcela dos registros foram mapeados para células que continham terra seca, sendo, portanto, descartados.

Na base de dados não há informação sobre pontos onde não há corais. Assim, optou-se por selecionar aleatoriamente um número de células não classificadas igual ao número de células classificadas, para que seja a elas atribuídas a classe *Não*. Sabidamente esta abordagem pode trazer uma perda de desempenho, por poder aplicar a classe *Não* a registros que deveriam conter corais.

Mesmo assim, verificou-se com a oceanógrafa que as células que receberam sua classificação através da mineração de dados foram "coerente com o esperado, tanto em termos de distribuição geográfica como de localização em relação à fisiografia do fundo oceânico".

Em uma análise somente com os registros inicialmente classificados, o algoritmo Random Forest foi selecionado dentre os algoritmos aplicados, por apresentar precisão geral de 71,1%, e 77,8% de precisão para a classe *Sim*. Para o domínio do problema abordado, a precisão da classe *Sim* é a mais importante, pois o objetivo neste problema é encontrar áreas no mapa que potencialmente possuem corais. Dentre os algoritmos executados, este foi o que obteve a menor precisão para a classe *Não* (ver Tabela 7.1). Entretanto, isto não desqualificou o algoritmo porque a qualidade da classificação

inicial das células da classe *Não* tende a ser menor, visto que estes não constam no conjunto de dados originais, e foram atribuídos de forma aleatória. Idealmente seria necessário um conjunto de dados com áreas que sabidamente não possuam corais, para aumentar ainda mais a qualidade da predição.

8. CONSIDERAÇÕES FINAIS

Devido à gama de aplicações da batimetria, como, por exemplo, o planejamento de rotas de navegação, o estudo de correntes marítimas e a prospecção de recursos minerais, esta tem despertado cada vez mais interesse. Nos últimos anos a comunidade científica tem buscado dados batimétricos mais precisos, e reconhecido cada vez mais a importância da batimetria como um todo.

Este trabalho foca-se em imagens de mapas batimétricos, que podem conter áreas extensas, inclusive, cobrindo todo o globo terrestre. Semelhanças de conformações em distintas áreas do mapa, ou variações sutis entre áreas próximas podem ser difíceis de serem detectadas, especialmente quando a única forma de análise das imagens batimétricas é a inspeção visual. Assim, é interessante a disponibilização de uma ferramenta computacional capaz de analisar tais imagens, levando em conta similaridades e dissimilaridades entre distintas áreas.

Para tal análise, é proposta a utilização da mineração de dados. Com tarefas preditivas, especialmente a classificação, e tarefas descritivas, principalmente a análise de agrupamentos, a mineração de dados possui técnicas úteis para a análise de imagens batimétricas. Entretanto, os algoritmos clássicos da mineração de dados aceitam como entrada somente dados tabulares, compostos de registros e seus atributos, impossibilitando seu uso diretamente em imagens.

A contribuição deste trabalho consiste em uma abordagem para minerar imagens batimétricas e também de uma representação iconográfica para a visualização dos resultados da mineração e de características do próprio mapa.

A abordagem para minerar imagens batimétricas contempla a extração de registros e de seus atributos. A extração de registros é obtida a partir da divisão da imagem de entrada em células quadradas, em pixels. Cada uma dessas células é considerada como um registro para a mineração de dados. Para a extração de atributos, são utilizadas técnicas de processamento de imagens bem conhecidas, bem como são propostas novas técnicas. Estatísticas dos valores dos pixels, histogramas de cores [26] e coeficientes de *Wavelets* [22] são as referidas técnicas bem conhecidas. Como técnicas novas, são propostas as regiões, que capturam a morfologia aproximada da célula, e os vetores, que capturam a aclividade e a orientação dominante da célula. Assim, os registros e atributos extraídos podem ser utilizados pelos algoritmos clássicos de mineração de dados.

A visualização dos resultados da mineração de dados é um fator importante dentro da análise de imagens batimétricas, podendo impactar positivamente no entendimento dos achados, por parte do especialista de domínio. Assim, é proposta uma iconografia para a visualização desses resultados, bem como de características do próprio mapa. Esta iconografia enfatiza as células que foram classificadas manualmente pelo especialista de domínio, em relação às células que tiveram seu rótulo de classe inferido pelos algoritmos de mineração. Além disso, características das células, especialmente os vetores, são indicadores importantes para o entendimento do mapa, fornecendo ao especialista de domínio maior clareza ao analisar a imagem. Na iconografia proposta, os vetores são desenhados sobre cada célula, enfatizando a orientação e a aclividade dominantes.

Para a validação da abordagem proposta, foi consultado uma especialista de domínio, uma oceanógrafa, para o acompanhamento da aplicação das técnicas descritas neste trabalho em uma base de dados real. Foi realizado um teste com uma base de dados de registros de corais de águas profundas. Os dados foram foram mapeados para células no mapa, e a abordagem proposta foi aplicada sobre a imagem. Conforme depoimento da oceanógrafa, os achados foram consistentes com a expectativa, sendo que as características das células classificadas como contendo corais são coerentes com as de células que sabidamente os possuem. Assim, a questão de pesquisa, descrita no Capítulo 3 pôde ser respondida de forma afirmativa.

Para pesquisas futuras, apresentam-se algumas possibilidades, descritas a seguir:

- A abordagem pode ser aplicada sobre outros conjuntos de dados, a fim de verificar seu desempenho em distintos problemas relativos à batimetria, incluindo problemas que utilizem mais de 2 classes;

- A iconografia proposta pode ser expandida, através da representação visual de outros atributos sobre cada célula correspondente. Por exemplo, uma curva normal poderia ser exibida sobre cada célula, a fim de representar os atributos estatísticos. Outro exemplo é o histograma, que pode ser representado como um gráfico de barras sobre cada célula;

- Sabidamente a projeção geográfica, que foi utilizada neste trabalho, introduz distorções no mapa, que se agravam em áreas mais distantes da linha do Equador. Pode-se pesquisar formas para a compensação dessa distorção nos atributos extraídos da célula;

- Finalmente, a abordagem proposta pode ser adaptada para outros domínios de imagens, que não a batimetria. Por exemplo, sem alterações na abordagem, poderiam ser mineradas imagens do relevo terrestre, considerando-se que pixels totalmente pretos representam água, fazendo com que células que os contenham sejam descartadas. Assim, o domínio do problema abordados seria o da topografia. Podem ser também pesquisados outros domínios onde a aplicação da abordagem proposta seja viável, mesmo que sejam necessárias algumas adaptações.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Alpaydin, E. "Introduction to machine learning". Cambridge, USA: MIT Press, 2010.

[2] Barnes, C.; Fritz, H.; Yoo, J. "Hurricane disaster assessments with image-driven data mining in high-resolution satellite imagery". *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45-6, 2007, pp. 1631–1640.

[3] Barnes, C.; Fritz, H.; Yoo, J. "Image-Driven Data Mining for Image Content Segmentation, Classification, and Attribution". *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45-9, 2007, pp 2964–2978.

[4] Barros, R. C. "Evolutionary model tree induction". Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2009.

[5] Basgalupp, M. P.; Barros, R. C.; de Carvalho, A. C.; Freitas, A. A.; Ruiz, D. D. "Legal-tree: a lexicographic multi-objective genetic algorithm for decision tree induction". In: ACM Symposium on Applied Computing, 2009, pp. 1085–1090.

[6] Brasil, DNIT - Departamento Nacional de Infraestrutura de Transportes. "Batimetria". Capturado em: http://www.dnit.gov.br/hidrovias/manutencao-hidroviaria/barimetria, Dezembro 2011.

[7] Breiman, L. "Bagging predictors". In: Machine Learning, 1996, pp. 123–140.

[8] Breiman, L. "Random forests". *Machine Learning*, vol. 45-1, 2001, pp. 5–32.

[9] Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. "Classification and Regression Trees" New York, USA: Chapman & Hall, 1984.

[10] Davies, A. J.; Guinotte, J. M. "Global habitat suitability for framework-forming cold-water corals". *PLoS ONE*, vol. 6-4, 2011, e18483.doi:10.1371/journal.pone.0018483.

[11] Davies, A. J.; Wisshak, M.; Orr, J. C.; Roberts, J. M. "Predicting suitable habitat for the cold-water coral lophelia pertusa (scleractinia)". *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 55-8, 2008, pp . 1048–1062.

[12] Ding, G.; Wang, J.; Xu, N.; Zhang, L. "Automatic Image Annotations by Mining Web Image Data". In: 2009 IEEE International Conference on Data Mining Workshops, 2009, pp. 152–157.

[13] Fan, J.; Gao, Y.; Luo, H.; Jain, R. "Mining Multilevel Image Semantics via Hierarchical Classification". *IEEE Transactions on Multimedia*, vol. 10-2, Fevereiro 2008, pp. 167–187.

[14] Freund, Y.; Schapire, R. "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences*, vol. 55-1, 1997, pp. 23–37.

[15] Glover, A. G.; Smith, C. R. "The deep-sea floor ecosystem: current status and prospects of anthropogenic change by the year 2025". *Environmental Conservation*, vol. 30-3, 2003, pp. 219–241.

[16] González, R.; Woods, R. "Digital image processing". Pearson/Prentice Hall, 2008.

[17] Gueguen, L.; Datcu, M. "Image Time-Series Data Mining Based on the Information-Bottleneck Principle". *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45-4, Abril 2007, pp. 827–838.

[18] Han, J.; Kamber, M.; Pei, J. "Data mining: Concepts and techniques". San Francisco, USA: Morgan Kaufmann Publishers Inc., 2005.

[19] Kitahara, M. V.; Horn, N. O.; Abreu, J. G. N. "Utilização de registros de corais de profundidade (Cnidaria, Scleractinia) para prever a localização e mapear tipos de substratos na plataforma e talude continental do sul do Brasil". In: Papéis Avulsos de Zoologia, 2008, pp. 11–18.

[20] Kitamoto, A. "Spatio-temporal data mining for typhoon image collection". *Journal of Intelligent Information Systems*, vol. 19-1, 2002, pp. 25–41.

[21] Lu, K. C.; Yang, D. L. "Image Processing and Image Mining using Decision Trees". *Journal of Information Science and Engeneering*, vol. 1003, 2009, pp. 989–1003.

[22] Mallat, S. "A theory for multiresolution signal decomposition: the wavelet representation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11-7, Julho 1989, pp. 674–693.

[23] Mayer, L. "Frontiers in seafloor mapping and visualization". *Marine Geophysical Research*, vol. 27-1, 2006, pp. 7–17.

[24] Merriam Webster Dictionary. "bathymetry". Capturado em: http://www.merriam-webster.com/dictionary/bathymetry, Novembro 2011.

[25] de Moustier, C. "State of the art in swath bathymetry survey systems". *International Hydrographic Review*, vol. 65-2, 1988, pp. 25–54.

[26] Novak, C. L.; Shafer, S. A. "Anatomy of a Color Histogram". In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992, pp. 599–605.

[27] Olariu, S.; Zomaya, A. "Handbook of bioinspired algorithms and applications". Chapman & Hall/CRC, 2006.

[28] Pearl, J. "Probabilistic reasoning in intelligent systems: networks of plausible inference". Morgan Kaufmann Publishers, 1988.

[29] Quinlan, J. R. "C4.5: Programs for machine learning". San Francisco, USA: Morgan Kaufmann, 1992.

[30] Roberts, J. M.; Wheeler, A. J.; Freiwald, A. "Reefs of the deep: The biology and geology of cold-water coral ecosystems". *Science*, vol 312-5773, 2006, pp. 543–547.

[31] Robinson, A. H.; Morrison, J. L.; Muehrcke, P. C.; Kimerling, A. J.; Guptill, S. C. "Elements of cartography". Wiley, 1995.

[32] Rogers, A.; Hall-Spencer, J. "Cold-water corals: Version 2.0". UNEP World Conservation Monitoring Centre (UNEP-WCMC), 2005.

[33] Sherwood, O. A.; Risk, M. J. "Chapter twelve deep-sea corals: New insights to paleoceanography". In: Developments in Marine Geology, 2007, pp. 491–522.

[34] Tan, P.; Steinbach, M.; Kumar, V. "Introduction to data mining". Boston, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[35] Tittensor, D. P.; Baco, A. R.; Brewin, P. E.; Clark, M. R.; Consalvey, M.; Hall-Spencer, J.; Schlacher, A. A.; Stocks, K. I.; Rogers, A. D. "Predicting global habitat suitability for stony corals on seamounts". *Journal of Biogeography*, vol. 36-6, 2009, pp. 1111–1128.

[36] Wang, J. Z.; Wiederhold, G.; Firschein, O.; Wei, S. X. "Wavelet-based image indexing techniques with partial sketch retrieval capability". In: 4 Forum on Research and Technology Advances in Digital Libraries, 1997, pp. 13–24.

[37] Wright, J.; Rothery, D.; Open University. "The ocean basins: their structure and evolution". Butterworth-Heinemann, 1998.