

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL  
INSTITUTO DE INFORMÁTICA  
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MARILTON SANCHOTENE DE AGUIAR

**Um Modelo Categorizador Intervalar  
 $n$ -Dimensional com  $l$ -Camadas Baseado em  
Tesselações**

Tese apresentada como requisito parcial  
para a obtenção do grau de  
Doutor em Ciência da Computação

Prof. Dr. Antônio Carlos da Rocha Costa  
Orientador

Porto Alegre, dezembro de 2003

## CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Aguiar, Marilton Sanchotene de

Um Modelo Categorizador Intervalar  $n$ -Dimensional com  $l$ -Camadas Baseado em Tesselações / Marilton Sanchotene de Aguiar. – Porto Alegre: PPGC da UFRGS, 2003.

110 f.: il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2003. Orientador: Antônio Carlos da Rocha Costa.

1. Modelo baseado em tesselações. 2. Categorização topográfica. 3. Matemática intervalar. 4. Geoprocessamento. I. Costa, Antônio Carlos da Rocha. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitora: Prof<sup>a</sup>. Wrana Maria Panizzi

Pró-Reitor de Ensino: Prof. José Carlos Ferraz Hennemann

Pró-Reitora Adjunta de Pós-Graduação: Prof<sup>a</sup>. Jocélia Grazia

Diretor do Instituto de Informática: Prof. Philippe Olivier Alexandre Navaux

Coordenador do PPGC: Prof. Carlos Alberto Heuser

Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

*Dedico este trabalho em memória de minha mãe.*

*“Experience is a hard teacher  
because she gives the test first,  
the lesson afterwards.”*  
— VERNON SANDERS LAW

## AGRADECIMENTOS

Agradeço:

- ao amigo e orientador professor Dr. Antônio Carlos da Rocha Costa, pela amplidão de sua amizade e acolhida desde o início do curso, pela confiança, dedicação e ensinamentos indispensáveis que me fizeram perceber a importância deste trabalho;
- às professoras, amigas e mães, Graçaliz Pereira Dimuro e Renata Hax Sander Reiser, pelos anos de pesquisa juntos, pelo investimento em minha vida acadêmica que agora estamos colhendo juntos, pela presença nos diversos momentos que passei, fáceis e difíceis;
- aos amigos que me acompanharam em Porto Alegre: Cadinho, Rogério e Cláudia, Juliana, Du Bois, Mônica, Patrícia, Márcia e Wives por todo o carinho e alegria;
- aos amigos dos NAPI III e V que conviveram comigo na UCPel pelos momentos alegres, felizes e intensos, em especial ao meu incansável amigo e bolsista Rafael Krolow que participou ativamente nos últimos anos deste trabalho;
- aos meus irmãos Maurício e Marlise que sempre me incentivaram e compreenderam as minhas preocupações e ansiedades e, além disso, souberam entender as minhas necessidades durante todo o tempo da realização de mais esta tarefa;
- à minha amada Clarissa, fonte inspiradora e motivadora da minha vida acadêmica, que em todos os momentos soube entender e relevar minhas atitudes, prioridades e ausências; e,
- ao CNPq, pelo auxílio em forma de bolsa.

# SUMÁRIO

<b>LISTA DE FIGURAS</b> . . . . .	8
<b>LISTA DE TABELAS</b> . . . . .	10
<b>RESUMO</b> . . . . .	11
<b>ABSTRACT</b> . . . . .	12
<b>1 INTRODUÇÃO</b> . . . . .	13
1.1 Aspectos Gerais da Tese . . . . .	13
1.2 Estrutura do Texto . . . . .	14
<b>2 TÉCNICAS DE CATEGORIZAÇÃO</b> . . . . .	15
2.1 Introdução . . . . .	15
2.2 Componentes do Processo de Categorização . . . . .	16
2.3 Representação dos Padrões . . . . .	17
2.4 Medidas de Similaridade . . . . .	19
2.5 Algoritmos de Categorização . . . . .	21
2.5.1 Algoritmos de Categorização Hierárquicos . . . . .	21
2.5.2 Algoritmos de Categorização por Partições . . . . .	25
2.5.3 Categorização Baseada em Grafos . . . . .	27
2.5.4 Categorização Baseada no Vizinho Mais Próximo . . . . .	27
2.5.5 Categorização <i>Fuzzy</i> . . . . .	27
2.5.6 Representação de Categorias . . . . .	29
2.5.7 Categorizações utilizando Redes Neurais Artificiais . . . . .	31
2.5.8 Abordagens Evolucionárias para Categorização . . . . .	32
2.5.9 Abordagens Baseadas em Buscas . . . . .	34
2.6 Considerações sobre o Modelo <i>ICTM</i> como Técnica Categorizadora . . . . .	35
<b>3 O ESTADO DA ARTE DO PROCESSAMENTO DE DADOS GEOGRÁFICOS</b> . . . . .	37
3.1 Sistemas de Informações Geográficas . . . . .	37
3.1.1 Histórico . . . . .	37
3.1.2 Coleta de Dados . . . . .	40
3.1.3 Tipos de Dados . . . . .	40
3.1.4 Estruturas de Dados . . . . .	41
3.1.5 Análise de Dados . . . . .	42
3.2 Geomática . . . . .	42
3.3 Análise Digital de Terrenos . . . . .	45

3.3.1	Fontes e Estruturas dos Dados Digitais de Elevação . . . . .	46
3.3.2	Utilização e Estimação de Atributos Topográficos . . . . .	47
3.3.3	Tratamento de Erros e Incerteza . . . . .	48
<b>4</b>	<b>O MODELO <i>Topo-ICTM</i></b> . . . . .	<b>50</b>
<b>4.1</b>	<b>A Formalização do Modelo</b> . . . . .	<b>50</b>
4.1.1	Matriz Absoluta e Matriz Relativa . . . . .	51
4.1.2	Matrizes Intervalares . . . . .	52
4.1.3	Matriz de Estados e os Registradores de Declividade . . . . .	53
4.1.4	Matriz de Limites e as Sub-regiões de Declividade Constante . . . . .	55
<b>4.2</b>	<b>Considerações sobre o Modelo <i>Topo-ICTM</i></b> . . . . .	<b>56</b>
<b>4.3</b>	<b>A Ferramenta <i>Topo-ICTM</i></b> . . . . .	<b>57</b>
<b>5</b>	<b>O MODELO <i>ICTM</i></b> . . . . .	<b>59</b>
<b>5.1</b>	<b>Modelo <i>2d-lc-ICTM</i></b> . . . . .	<b>59</b>
<b>5.2</b>	<b>Modelo <i>nd-lc-ICTM</i></b> . . . . .	<b>62</b>
<b>5.3</b>	<b>Algoritmos para a Projeção das Camadas</b> . . . . .	<b>66</b>
<b>6</b>	<b>EXTRAÇÃO DE FATOS</b> . . . . .	<b>68</b>
<b>6.1</b>	<b>Extração de Fatos Representados por Sentenças Lógicas</b> . . . . .	<b>68</b>
6.1.1	Conhecimento Quantitativo Probabilístico . . . . .	74
<b>7</b>	<b>IMPLEMENTAÇÃO DO MODELO</b> . . . . .	<b>77</b>
<b>7.1</b>	<b>Étapas de funcionamento do <i>ICTM</i></b> . . . . .	<b>77</b>
7.1.1	Definição da Malha . . . . .	77
7.1.2	Caracterização Topográfica das Áreas Geográficas . . . . .	77
<b>7.2</b>	<b>Formação da Matriz Absoluta</b> . . . . .	<b>78</b>
7.2.1	Utilização de Técnicas Intervalares para Controle dos Erros . . . . .	78
7.2.2	Determinação dos Estados das Células do Modelo . . . . .	78
<b>7.3</b>	<b>Identificação das Regiões Topográficas</b> . . . . .	<b>79</b>
<b>7.4</b>	<b>Considerações Finais Sobre a Implementação</b> . . . . .	<b>80</b>
<b>8</b>	<b>RESULTADOS EXPERIMENTAIS</b> . . . . .	<b>82</b>
<b>9</b>	<b>CONCLUSÕES</b> . . . . .	<b>86</b>
	<b>REFERÊNCIAS</b> . . . . .	<b>89</b>
	<b>APÊNDICE A EXEMPLOS DE CATEGORIZAÇÕES</b> . . . . .	<b>95</b>

## LISTA DE FIGURAS

Figura 1.1:	Esquema do Processo Dinâmico . . . . .	14
Figura 2.1:	Categorização de dados . . . . .	15
Figura 2.2:	Categorização curvilínea cujos pontos estão equidistantes da origem .	18
Figura 2.3:	A e B são mais similares do que B e C . . . . .	20
Figura 2.4:	Após mudança de contexto, B e C são mais similares que B e A. . . .	20
Figura 2.5:	Categorização do tipo <i>Monothetic</i> . . . . .	21
Figura 2.6:	Pontos agrupados em três categorias . . . . .	22
Figura 2.7:	Dendrograma obtido pelo algoritmo hierárquico . . . . .	22
Figura 2.8:	Categorização de ligação-simples de um padrão contendo duas classes	23
Figura 2.9:	Categorização de ligação-completa de um padrão contendo duas classes	23
Figura 2.10:	Duas categorias concêntricas . . . . .	24
Figura 2.11:	O algoritmo de $k$ -médias é sensível à partição inicial . . . . .	26
Figura 2.12:	Árvore Geradora Mínima para identificar categorias . . . . .	27
Figura 2.13:	Categorias <i>Fuzzy</i> . . . . .	28
Figura 2.14:	Representação de uma categoria por pontos . . . . .	29
Figura 2.15:	Representação de categorias por árvore de classificação ou por proposições conjuntivas . . . . .	30
Figura 2.16:	Compressão de dados através de categorizações . . . . .	31
Figura 2.17:	Operação de Cruzamento . . . . .	33
Figura 2.18:	Algoritmos Genéticos executam busca globalizada . . . . .	33
Figura 3.1:	Processos biofísicos em camadas . . . . .	45
Figura 3.2:	Métodos de estruturação dos dados de elevação . . . . .	46
Figura 3.3:	Tarefas associadas a modelagem digital de terrenos . . . . .	49
Figura 4.1:	Etapas do Modelo <i>Topo-ICTM</i> . . . . .	51
Figura 4.2:	Esquema de todos os valores possíveis dos estados da célula . . . . .	55
Figura 4.3:	Esquemas das células limítrofes . . . . .	56
Figura 4.4:	A categorização da declividade de uma dada região $R$ . . . . .	56
Figura 4.5:	Mosaico das três órbitas-ponto do satélite LANDSAT TM7 . . . . .	58
Figura 4.6:	Mapa de localização das cenas do satélite ASTER . . . . .	58
Figura 5.1:	Visão do Modelo <i>2d-lc-ICTM</i> . . . . .	60
Figura 5.2:	Visão do <i>nd-lc-ICTM</i> , neste caso ( $n = 3$ ) . . . . .	63
Figura 5.3:	Notação do sentido dos registradores $a_i^\pm$ . . . . .	64
Figura 5.4:	Exemplo de projeção das camadas (tipo I) . . . . .	66
Figura 5.5:	Exemplo de projeção das camadas (tipo II) . . . . .	67

Figura 6.1:	Envelope de um polígono tridimensional . . . . .	68
Figura 6.2:	Erro encontrado na técnica do envelope . . . . .	69
Figura 6.3:	Região $\mathcal{SR}_i$ inscrita na região $\mathcal{SR}_j$ . . . . .	69
Figura 6.4:	Sub-regiões sem intersecção . . . . .	70
Figura 6.5:	Intersecção de sub-regiões . . . . .	70
Figura 6.6:	O fato $l_1\text{-}\mathcal{SR}_i \Rightarrow l_2\text{-}\mathcal{SR}_j$ extraído de uma malha $M$ . . . . .	70
Figura 6.7:	O fato $\neg(l_1\text{-}\mathcal{SR}_i \wedge l_2\text{-}\mathcal{SR}_j)$ extraído de uma malha $M$ . . . . .	71
Figura 6.8:	O fato $(l_3\text{-}\mathcal{SR}_k \wedge l_2\text{-}\mathcal{SR}_j) \Rightarrow l_1\text{-}\mathcal{SR}_i$ extraído de uma malha $M$ . . . . .	72
Figura 6.9:	Aproximação interior de uma sub-região através de intervalos . . . . .	73
Figura 6.10:	Problema de Ponto em Polígono . . . . .	74
Figura 6.11:	Casos de cruzamento e número de intersecções . . . . .	74
Figura 6.12:	Matriz de Ocorrências . . . . .	75
Figura 6.13:	Coleção de eventos mutuamente exclusivos . . . . .	76
Figura 7.1:	Vizinhança de Von Neumann . . . . .	78
Figura 7.2:	Interface do Protótipo . . . . .	80
Figura 8.1:	Imagem do DEM de resolução 1000m . . . . .	82
Figura 8.2:	Imagem do DEM de resolução 500m . . . . .	82
Figura 8.3:	Linhas de Contorno no DEM de resolução 1000m . . . . .	84
Figura 8.4:	Linhas de Contorno no DEM de resolução 500m . . . . .	84
Figura 8.5:	Graus de Declividade do DEM de resolução 1000m . . . . .	85

## LISTA DE TABELAS

Tabela 3.1:	Exemplos de atributos topográficos primários . . . . .	48
Tabela 4.1:	Condições das células $xy$ não limítrofes . . . . .	55
Tabela 4.2:	A matriz de limites $M_{xy}^{limit}$ associada a $R$ . . . . .	56
Tabela 5.1:	Condições das células $xy$ não limítrofes da camada $l$ . . . . .	62
Tabela 5.2:	Condições das células $a_1 \dots a_n$ não limítrofes da camada $l$ , para $i =$ $1 \dots n$ . . . . .	65
Tabela 8.1:	Resultados para o quadrante no DEM 1000m . . . . .	83
Tabela 8.2:	Resultados para o quadrante no DEM 500m . . . . .	83

## RESUMO

O *ICTM* (Interval Categorizer Tessellation Model), objeto da presente tese, é um modelo geral para análise de espaços de natureza geométrica, baseado em tesselações, que é capaz de produzir uma categorização confiável de conjunto de pontos de um dado espaço, de acordo com múltiplas características dos pontos, cada característica correspondendo a uma camada do modelo.

Por exemplo, na análise de terrenos geográficos, uma região geográfica pode ser analisada de acordo com a sua topografia, vegetação, demografia, dados econômicos etc, cada uma gerando uma subdivisão diferente da região.

O modelo geral baseado em tesselações não está restrito, porém, à análise de espaços bi-dimensionais. O conjunto dos pontos analisados pode pertencer a um espaço multi-dimensional, determinando a característica multi-dimensional de cada camada. Um procedimento de projeção das categorizações obtidas em cada camada sobre uma camada básica leva a uma categorização confiável mais significativa, que combina em uma só classificação as análises obtidas para cada característica. Isto permite muitas análises interessantes no que tange à dependência mútua das características.

A dimensão da tesselação pode ser arbitrária ou escolhida de acordo com algum critério específico estabelecido pela aplicação. Neste caso, a categorização obtida pode ser refinada, ou pela re-definição da dimensão da tesselação ou tomando cada sub-região resultante para ser analisada separadamente.

A formalização usando matrizes de registradores acarreta que a informação guardada nos registradores pode ser facilmente recuperada apenas pela indexação dos elementos das matrizes, em qualquer momento da execução. A implementação do modelo é naturalmente paralela, uma vez que a análise é feita basicamente por regras locais.

Como os dados de entrada numéricos são usualmente suscetíveis a erros, o modelo utiliza a aritmética intervalar para se ter um controle automático de erros. O modelo *ICTM* também suporta a extração de fatos sobre as regiões de modo qualitativo, por sentenças lógicas, ou quantitativamente, pela análise de probabilidade.

Este trabalho recebe apoio financeiro do CNPq/CTPETRO e FAPERGS.

**Palavras-chave:** Modelo baseado em tesselações, categorização topográfica, matemática intervalar, geoprocessamento.

## An Interval $n$ -dimensional $l$ -Layer Categorizer Tessellation Model

### ABSTRACT

The *ICTM* (Interval Categorizer Tessellation Model), object of the present thesis, is a general model for the analysis of spaces of geometric nature, based in tessellations, that are capable to produce a reliable categorization of point of a given data space, in accordance with multiple characteristics of the points, each characteristic corresponding to a layer of the model.

For example, in the geographic terrain analysis, a geographic region can be analyzed according to its topography, vegetation, demography, economic data, etc., each one generating a different subdivision of the region.

However, the tessellation-based general model is not restricted to bi-dimensional spaces analysis. The set of the analyzed points can belong to a multi-dimensional space, determining the multi-dimensional characteristic of each layer. A procedure of projection of the categorizations gotten in each layer on a basic layer leads to a meaningful reliable categorization, that combines in just one classification the analysis gotten for each characteristic. This allows many interesting analysis on the mutual dependence of these characteristics.

The dimension of the tessellation may be arbitrary or chosen according to specific criteria established by the application. In any case, the categorization obtained may be refined, either by dimension redefinition of the tessellation or taking each sub-region resultant to be analyzed separately.

The formalization using matrices of registers allows the information recorded in those registers to be easily recovered by the indexing elements of the matrices, at any time. The implementation of the model is naturally parallel since the analysis is performed on the basis of local rules.

As the input numerical data are usually susceptible to errors, the model uses interval arithmetic to have an automatic control of these errors. The model *ICTM* also supports the extraction of facts about the regions, both in a qualitative approach by logical sentences and in a quantitative approach by a probability analysis.

This work receives financial support from CNPq/CTPETRO and FAPERGS.

**Keywords:** tessellation-based model, topographic categorization, interval mathematics, geoprocessing.

# 1 INTRODUÇÃO

## 1.1 Aspectos Gerais da Tese

Esta tese contempla um modelo geral baseado em tesselações para ferramentas de categorização que são capazes de sub-dividir um espaço de natureza geométrica em sub-regiões apresentando características similares, isto é, contendo pontos com um mesmo conjunto de propriedades observáveis. O número de características que são analisadas determina o número de camadas do modelo. Em cada camada, uma análise diferente do espaço é obtida.

Uma projeção apropriada de todas as camadas na camada básica do modelo leva a uma sub-divisão mais significativa da região e a uma categorização das sub-regiões que considera a ocorrência simultânea de todas as características, de acordo com prioridades.

Além disso, o modelo provê mecanismos para a extração de fatos sobre o espaço analisado, a partir da correlação entre as características das sub-regiões tratadas pelo modelo. Esse conhecimento contempla aspectos relacionais (de características relativas entre regiões diferentes).

Para controlar os erros oriundos da discretização e resultantes das computações numéricas, são utilizadas técnicas intervalares (MOORE, 1979) para obter uma categorização mais confiável.

O modelo resultante dessa combinação de recursos técnicos e conceituais é chamado de *Interval Categorizer Tessellation-Based Model (ICTM)*.

O artigo de (COBLENTZ et al., 2003) propõe uma metodologia de sub-divisão confiável de áreas geológicas baseada na análise da monotonicidade da função que mapeia sua declividade. A área total é dividida em regiões perfiladas lado a lado e, por isso, a análise é efetuada unidimensionalmente percorrendo somente um sentido do modelo.

Inicialmente pretendia-se utilizar o conceito de autômatos celulares para embasar este tipo de análise categorizadora. Isso permitiria não apenas uma análise estática das características do espaço em um momento dado, mas também possibilitaria o avanço na direção de uma simulação dos aspectos dinâmicos desse espaço, em um intervalo de tempo. Entretanto, a possibilidade de extensão do modelo com mecanismos de descoberta de conhecimentos, levou à escolha da modelagem baseada em tesselações, que se limita a uma análise estática do espaço.

As tesselações podem ser consideradas como a estrutura estática subjacente aos autômatos celulares. Ou seja, assim como o autômato celular, as tesselações são malhas de células idênticas e discretas, onde cada uma destas células têm seu estado determinado localmente a partir dos estados da sua vizinhança de células.

Todavia, as evoluções destes estados, dadas pelas regras de transição de estados dos autômatos celulares (veja a Figura 1.1), não estão presentes nas tesselações, que se limi-

tam à geração de um único valor de estado fixo, para suas células.

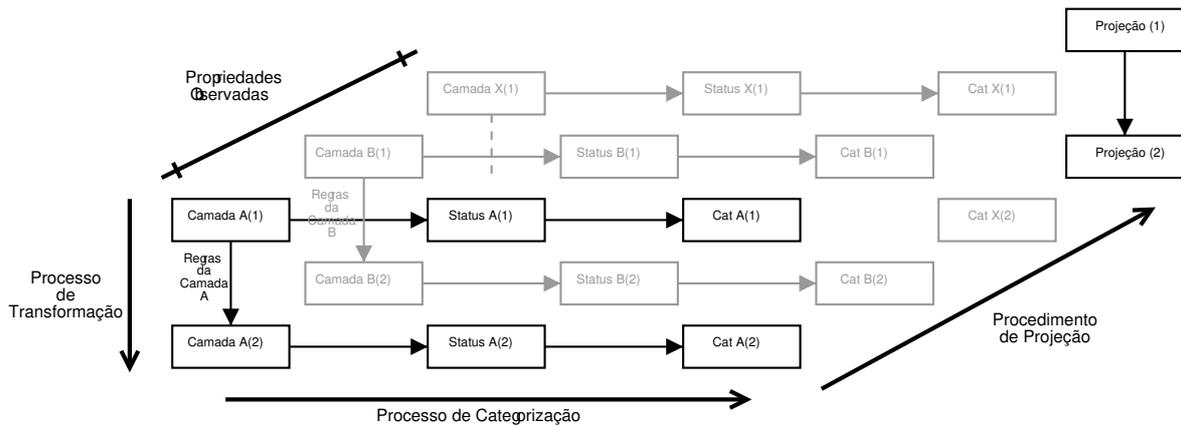


Figura 1.1: Esquema do Processo Dinâmico

Essa restrição conceitual da idéia possibilitou, no entanto, que fosse tentada a utilização imediata do modelo em uma área aplicada bem determinada, qual seja, a do processamento de imagens geográficas, mais especificamente, a análise digital de terrenos (WILSON; GALLANT, 2000).

Essa possibilidade de aplicação tornou-se real com a aprovação de um projeto no contexto do fundo setorial CT-PETRO, seguido de uma extensão do mesmo no contexto do fundo setorial CT-INFO. Isso motivou ainda mais a preocupação com a elaboração de meios para permitir análises realísticas de geoprocessamento a partir do modelo proposto, assim como o desenvolvimento de recursos de descoberta de conhecimento voltada para esses aspectos realísticos.

## 1.2 Estrutura do Texto

O capítulo 2 apresenta as componentes de um processo de categorização, as métricas utilizadas e os algoritmos de categorização tradicionais.

No capítulo 3, apresenta-se um apanhado geral do estado da arte dos temas Geoprocessamento, Geomática e Análise Digital de Terrenos, suas aplicações, as fontes e as estruturas dos dados, definição e utilização dos atributos topográficos e os tipos de erros.

No capítulo 4, apresenta-se a formalização do *Topo-ICTM*: matrizes absoluta e relativas, estados e registradores de declividade, a determinação dos limites e considerações sobre este modelo.

O modelo *ICTM*, que é a generalização do *Topo-ICTM*, é exposto no capítulo 5. A generalização se dá em duas frentes: a primeira, generalização do *ICTM* para multicamadas; e a segunda, a generalização nas dimensões de cada camada. Além disso, na seção 5.3, apresentam-se as definições dos algoritmos de projeção das camadas.

No capítulo 6, são apresentados os algoritmos para a extração de fatos no modelo *ICTM*.

Considerações sobre a implementação do modelo são apresentadas no capítulo 7, Além disso, os resultados experimentais são apresentados no capítulo 8. As conclusões e apontamentos para trabalhos futuros são apresentadas no capítulo 9.

## 2 TÉCNICAS DE CATEGORIZAÇÃO

### 2.1 Introdução

A análise de dados é fundamental para muitas aplicações, tanto na fase de desenvolvimento quanto em tempo de execução da aplicação. Procedimentos de análise de dados podem ser classificados como exploratórios ou confirmatórios, baseados no formato da disponibilidade dos dados. Segundo (BERKHIN, 2002), um elemento chave em ambos os tipos de procedimentos (para formação da hipótese ou para a tomada de decisão) é o agrupamento, ou seja, uma classificação baseada em medidas definidas pelo (i) próprio modelo, ou (ii) inferida a partir da análise do agrupamento (categorização) natural dos dados (*clustering*).

A análise de categorias é a organização de uma coleção de padrões – usualmente representados como um vetor ou um ponto em um espaço multi-dimensional – baseados na similaridade (JAIN; MURTY; FLYNN, 1999).

Intuitivamente, padrões em uma determinada categoria são mais similares entre si do que em relação a padrões de uma categoria diferente. Um exemplo de agrupamento pode ser observado na Figura 2.1. Os padrões de entrada são mostrados na Figura 2.1a e as categorias desejadas são mostradas na Figura 2.1b.

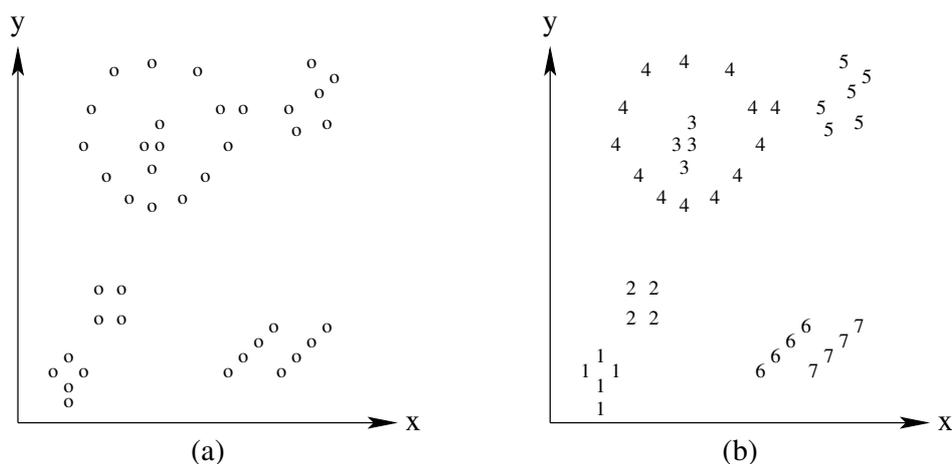


Figura 2.1: Categorização de dados

Aos pontos pertencentes à mesma categoria são dados o mesmo rótulo. Existe, uma variedade de técnicas para representar os dados, para medir a proximidade (similaridade) entre os elementos de dados e para agrupar elementos as quais serão mostradas nas próximas seções.

A categorização é utilizada em vários tipos de análises exploratórias de padrões, agrupamentos, tomadas de decisão, aprendizagem de máquina, mineração de dados, recuperação de documentos, segmentação de imagens e classificação de padrões (BERKHIN, 2002).

Entretanto, em muitos destes problemas, existe pouca informação conhecida *a priori*, isto é, modelos estatísticos dos dados, e é sobre estas restrições que a metodologia de categorização é apropriada, para a exploração de inter-relacionamentos entre os pontos de dados para fazer uma avaliação (preliminar) da sua estrutura.

## 2.2 Componentes do Processo de Categorização

Tipicamente, a atividade de categorização envolve os seguintes passos (JAIN; DUBES, 1988):

1. representação do padrão.
2. definição de uma medida para a proximidade (similaridade) do padrão adequada ao domínio de dados.
3. categorização ou agrupamento.
4. abstração dos dados (se necessário).
5. avaliação da saída (se necessário).

A representação do padrão refere-se ao número de classes, ao número de padrões disponíveis, e o número, tipo e escala das propriedades disponíveis para o algoritmo de categorização. Algumas destas informações podem não ser controladas pelo especialista.

A proximidade de padrões é usualmente medida por uma função distância definida sobre pares de padrões. Por exemplo, a medida de distância Euclidiana pode ser utilizada para refletir o grau de similaridade entre dois padrões (DIDAY; SIMON, 1976).

A etapa de agrupamento pode ser efetuado de vários modos: (i) a categorização resultante pode ser a partição dos dados em conjuntos, ou (ii) *fuzzy*, onde cada padrão tem um grau variável de pertinência em cada categorização resultante.

Os algoritmos de categorização hierárquicos produzem uma série de partições encaixadas que otimizam localmente um critério de categorização.

A abstração dos dados é o processo de extrair uma representação simples e concisa de um conjunto de dados. Normalmente, simplicidade refere-se à perspectiva de uma análise automática (efetuada pelo próprio algoritmo) ou à uma representação de fácil compreensão pelo analista. Uma abstração típica dos dados é uma descrição compacta de cada categoria, normalmente em termos dos centróides de cada categoria (DIDAY; SIMON, 1976).

A saída de um algoritmo de categorização pode ser avaliada. Todo algoritmo de categorização é apto a produzir categorias, obviamente, se os dados possuírem tais categorias. Na verdade, a avaliação dos resultados de um algoritmo de categorização é muito mais uma avaliação do domínio dos dados que do algoritmo propriamente.

O estudo da tendência de categorização, onde os dados de entrada são examinados para verificar o mérito de uma categorização antes de ser efetuada, ainda é uma área de pesquisa carente de estudos (JAIN; MURTY; FLYNN, 1999).

Por isso, a análise da validade da categorização é a avaliação da saída do procedimento de categorização normalmente utilizando um critério de otimalidade subjetivo.

Ao se utilizar uma abordagem estatística para a categorização a validação é efetuada através da aplicação de métodos estatísticos apropriados. Existem três tipos de estudos de validação (DUBES, 1987, 1993): (i) avaliação externa da validade, onde a categorização obtida é comparada com uma categorização escolhida *a priori*; (ii) avaliação interna da validade, que tenta determinar se a categorização é intrinsecamente apropriada para os dados; e, (iii) teste relativo, que compara duas categorizações e mede o seu mérito relativo.

**Definição 2.1.** *Um padrão  $\mathbf{x}$  (ou vetor de atributos) é o item de dado mais simples utilizado pelo algoritmo de categorização e, tipicamente, consiste de um vetor*

$$\mathbf{x} = (x_1, \dots, x_d)$$

onde  $d$  indica sua dimensionalidade e os  $x_i$  são os componentes escalares chamados de atributos.

**Definição 2.2.** *Um conjunto de padrões é denotado por  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , tal que o  $i$ -ésimo padrão em  $\mathcal{X}$  é denotado por*

$$\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d}).$$

Na maioria das vezes um conjunto de padrões a ser categorizado é visto como uma matriz de padrões  $n \times d$ .

Uma classe pode ser entendida como a origem de padrões cuja distribuição no espaço de atributos é governada por uma densidade de probabilidade particular à esta classe.

As técnicas de categorização tentam agrupar padrões de modo que as classes obtidas reflitam os diferentes processos de geração de padrões representados no conjunto de padrões.

As técnicas de categorização baseadas em conjuntos associam um rótulo  $l_i$  a cada classe de padrões  $\mathbf{x}_i$  para identificá-la.

**Definição 2.3.** *O conjunto de todos os rótulos para um conjunto de padrões  $\mathcal{X}$  é  $\mathcal{L} = \{l_1, \dots, l_n\}$ , com  $l_i \in \{1, \dots, k\}$ , onde  $k$  é o número de categorias.*

As técnicas de categorização *fuzzy* associam a cada padrão  $\mathbf{x}_i$  um grau de pertinência fracionário  $f_{ij}$  em cada categoria resultante  $j$ .

A medida de distância (uma especialização de uma medida de proximidade) é uma métrica sobre o espaço de atributos para quantificar a similaridade dos padrões.

## 2.3 Representação dos Padrões

Freqüentemente, assume-se que a representação dos padrões já está disponível antes da categorização, no sentido de que é uma tarefa bastante especializada e dificilmente o usuário iria sugerir a formatação apropriada dos padrões e atributos. O papel do usuário no processo limita-se à conjectura sobre os dados, efetuar alguma seleção ou extração de atributos.

Uma boa representação dos padrões pode assegurar um fácil e simples entendimento da categorização; por outro lado, uma representação deficiente pode resultar em uma categorização complexa, cuja estrutura dificulta ou impossibilita o discernimento das categorias.

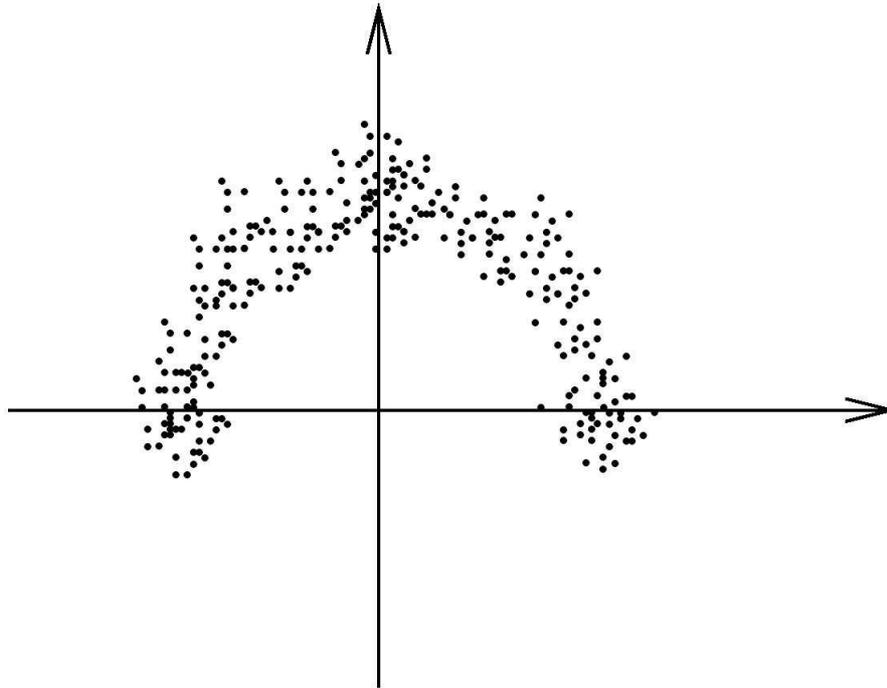


Figura 2.2: Categorização curvilínea cujos pontos estão equidistantes da origem

A Figura 2.2 mostra um exemplo desta característica comentada no parágrafo anterior. Os pontos neste espaço bidimensional estão arranjados em uma categoria curvilínea de distância aproximadamente constante a partir da origem.

Se forem escolhidas as coordenadas cartesianas para representar os padrões, muitos algoritmos de categorização fragmentariam a categoria em duas ou mais categorias. Por outro lado, se forem utilizadas coordenadas polares para a representação das categorias, a solução com uma categoria seria facilmente obtida.

Um padrão pode medir um objeto físico (por exemplo, uma cadeira) ou uma noção abstrata (como um estilo de escrita). Padrões são representados como vetores multidimensionais, onde cada dimensão é um atributo (DUDA; HART, 1974). Estes atributos pode ser quantitativos ou qualitativos.

Por exemplo, se peso e cor são dois atributos utilizados, então  $(20, \text{preto})$  é a representação de um objeto preto com 20 unidades de peso. Estes atributos podem ser sub-divididos nos seguintes tipos (GOWDA; DIDAY, 1992):

1. Atributos Quantitativos:

- (a) valores contínuos (peso);
- (b) valores discretos (número de computadores);
- (c) valores intervalares (duração de um evento).

2. Atributos Qualitativos:

- (a) nominal ou sem ordem (cor);
- (b) ordinal ou avaliações qualitativas de intensidade.

Pode-se ainda utilizar atributos estruturados (CULLINGFORD et al., 1982) que são representados como árvores, onde o nodo pai representa a generalização de seus nodos

filhos. Por exemplo, um nodo pai *veículo* pode ser a generalização dos nodos filhos *carros*, *ônibus*, *caminhões* e *motocicletas*.

## 2.4 Medidas de Similaridade

A similaridade é fundamental para a definição da categorização. Uma medida de similaridade entre dois padrões definidos a partir do mesmo espaço de atributo é essencial para os algoritmos de categorização.

Devido à variedade dos tipos de atributos e escalas, a medida de distância (ou medidas) deve ser escolhida criteriosamente. É possível também calcular a dissimilaridade (conceito dual da similaridade) entre dois padrões usando uma medida de distância definida no espaço de atributos.

A métrica mais utilizada para atributos contínuos é a distância Euclidiana

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left( \sum_{k=1}^d (x_{i,k} - x_{j,k})^2 \right)^{\frac{1}{2}} = \|\mathbf{x}_i - \mathbf{x}_j\|_2,$$

que é o caso especial ( $p = 2$ ) da métrica de Minkowski

$$d_k(\mathbf{x}_i, \mathbf{x}_j) = \left( \sum_{k=1}^d |x_{i,k} - x_{j,k}|^p \right)^{\frac{1}{p}} = \|\mathbf{x}_i - \mathbf{x}_j\|_k.$$

A distância Euclidiana normalmente é utilizada para avaliar a proximidade de objetos em espaços em duas ou três dimensões, quando o conjunto dos dados possui categorizações compactas ou isoladas (MAO; JAIN, 1994).

A correlação linear entre os atributos pode também distorcer as medidas de distância. Segundo (JAIN; MURTY; FLYNN, 1999), esta distorção pode ser abreviada aplicando-se uma transformação aos dados ou utilizando a distância de Mahalanobis

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j) \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)^T,$$

onde os padrões  $\mathbf{x}_i$  e  $\mathbf{x}_j$  são vetores linha e  $\Sigma$  é a matriz de covariância dos padrões;  $d_M$  associa diferentes pesos aos diferentes atributos baseados nas suas variâncias e na correlação linear entre pares de atributos.

Alguns algoritmos de categorização utilizam uma matriz de valores de proximidade ao invés do conjunto original de padrões. Em tais situações é comum pré-computar todos os  $\frac{n(n-1)}{2}$  valores de distância para os  $n$  padrões e armazená-los em uma matriz simétrica.

A computação de distâncias entre padrões com atributos não-contínuos é problemática, pois atributos de tipos diferentes não são comparáveis. Entretanto, pesquisadores, especialmente aqueles da área de aprendizagem de máquina, onde padrões de tipos variados são comuns, têm se voltado para o desenvolvimento de medidas de proximidade para padrões de tipo heterogêneos.

Padrões também podem ser representados por *strings* ou estruturas de árvores (KNUTH, 1973). *Strings* são utilizados em categorização sintática (FU; LU, 1977) e várias medidas de similaridade entre *strings* são apresentadas em (BAEZA-YATES, 1992). Um sumário das medidas de similaridade entre árvore é apresentado em (ZHANG, 1995) e uma comparação entre as abordagens estatísticas e sintáticas para reconhecimento de padrões pode ser encontrada em (TANAKA, 1995).

(CULLINGFORD et al., 1982) apresentam uma medida que leva em conta o efeito dos pontos vizinhos, chamada de contexto. A função  $f$  de similaridade entre os vizinhos dos pontos  $\mathbf{x}_i$  e  $\mathbf{x}_j$ , dado este contexto, é determinada por

$$s(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i, \mathbf{x}_j, \mathcal{E}),$$

onde  $\mathcal{E}$  é o contexto (conjunto de pontos vizinhos). Uma métrica definida usando este contexto é a distância mútua entre vizinhos (*Mutual Neighbor Distance – MND*), denotada por:

$$MND(\mathbf{x}_i, \mathbf{x}_j) = NN(\mathbf{x}_i, \mathbf{x}_j) + NN(\mathbf{x}_j, \mathbf{x}_i),$$

onde  $NN(\mathbf{x}_i, \mathbf{x}_j)$  é o número de vizinhos de  $\mathbf{x}_j$  com respeito a  $\mathbf{x}_i$ .

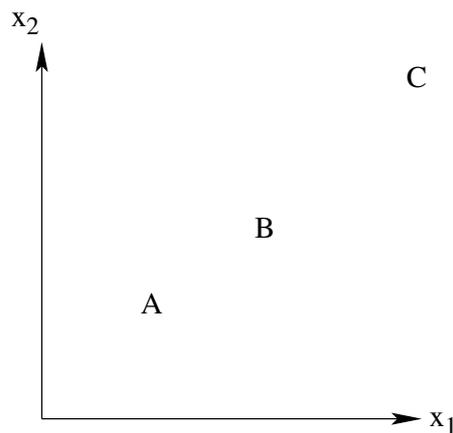


Figura 2.3: A e B são mais similares do que B e C

Na Figura 2.3, o vizinho mais próximo à A é B e o mais próximo à B é A. Então,  $NN(A, B) = NN(B, A) = 1$  e o  $MND$  entre A e B é 2. Entretanto,  $NN(B, C) = 1$  mas  $NN(C, B) = 2$ , e portanto  $MND(C, B) = 3$ .

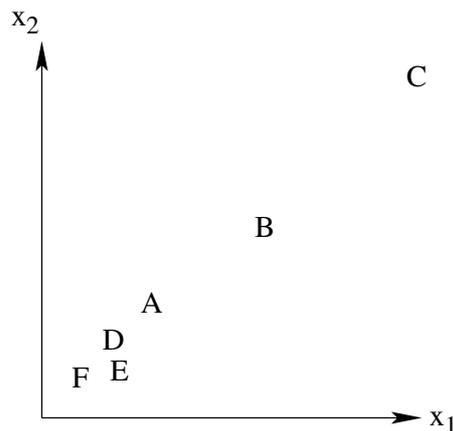


Figura 2.4: Após mudança de contexto, B e C são mais similares que B e A.

A Figura 2.4 foi obtida a partir da Figura 2.3 adicionando três novos pontos  $D$ ,  $E$  e  $F$ . Então,  $MND(B, C) = 3$ , mas  $MND(A, B) = 5$ . O  $MND$  entre A e B modificou-se devido à introdução de novos pontos, mesmo sem A e B terem se movidos.

No trabalho (ZHANG, 1995) afirma-se que  $MND$  não é uma métrica, entretanto,  $MND$  tem sido utilizada com sucesso em várias aplicações de categorização (GOWDA; DIDAY, 1992). Esta observação apóia o ponto de vista que a dissimilaridade não precisa ser uma métrica (JAIN; MURTY; FLYNN, 1999).

## 2.5 Algoritmos de Categorização

Nesta seção, serão apresentadas as técnicas de categorização existentes seguindo a taxonomia apresentada no trabalho de (JAIN; DUBES, 1988).

As técnicas de categorização podem ser classificadas quanto à sua estrutura algorítmica e funcionamento como sendo **aglomerativas** ou **divisoras**. Uma abordagem aglomerativa inicia o processo com cada padrão em uma categoria distinta, chamada de *singleton*, e sucessivamente agrega categorias vizinhas até satisfazer um critério de parada. Na abordagem divisora, o método parte com todos os padrões em uma mesma categoria e faz divisões (particionamentos) até satisfazer o critério de parada.

Quanto ao uso dos atributos, as técnicas podem ser dos tipos<sup>1</sup> *Monothetic* ou *Polythetic*. A maioria dos algoritmos são do tipo *Polythetic*, ou seja, todos os atributos fazem parte da computação das distâncias entre padrões, e decisões são baseadas nestas distâncias. Um algoritmo do tipo *Monothetic* considera os atributos seqüencialmente para dividir um dado conjunto de padrões.

A Figura 2.5 ilustra um algoritmo do tipo *Monothetic*. O conjunto de padrões é dividido em dois grupos utilizando o atributo  $x_1$  (a linha vertical  $V$  é a linha separadora destes grupos). Cada uma destas categorias é dividida mais uma vez utilizando o atributo  $x_2$ , conforme indicado nas linhas  $H_1$  e  $H_2$ . A grande deficiência deste tipo de abordagem é que o algoritmo gera  $2^d$  categorias, onde  $d$  é a dimensionalidade dos padrões. Para  $d > 100$ , situação bastante comum em aplicações de recuperação de informações, o número de categorias geradas pelo algoritmo é tão grande que o conjunto de dados é dividido em categorias muito fragmentadas e insignificantes.

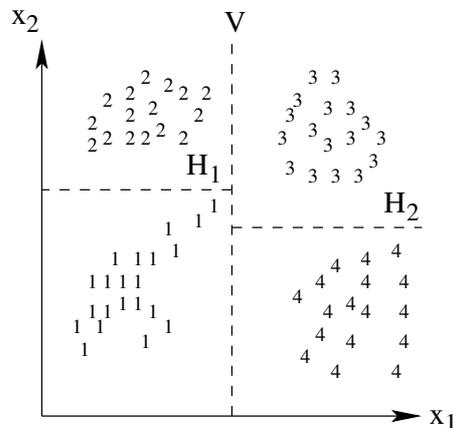


Figura 2.5: Categorização do tipo *Monothetic*

Quanto ao grau de pertinência de um ponto a uma categoria, um algoritmo de categorização pode ser chamado de *fuzzy* ou arbitrário. Uma categorização arbitrária aloca cada padrão a apenas uma categoria durante sua operação. O método de categorização *fuzzy* associa graus de pertinência de cada padrão de entrada a cada categoria.

### 2.5.1 Algoritmos de Categorização Hierárquicos

O funcionamento de um algoritmo de categorização hierárquico utilizando um conjunto de dados bidimensional pode ser observado na Figura 2.6.

Esta figura apresenta sete padrões  $A$ ,  $B$ ,  $C$ ,  $D$ ,  $E$ ,  $F$  e  $G$  e três categorias. Um algoritmo hierárquico constrói um *dendrograma* representando a categorização sucessiva

<sup>1</sup>Não foram encontrados termos equivalentes em português.

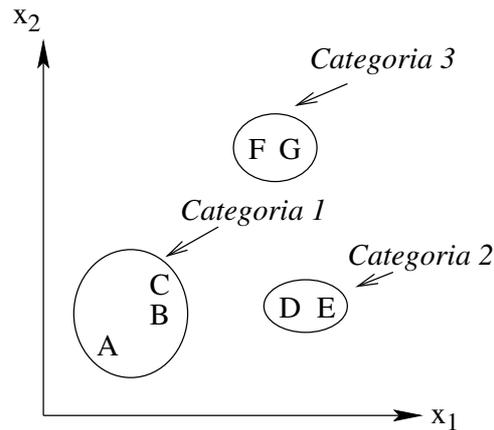


Figura 2.6: Pontos agrupados em três categorias

de padrões e os níveis de similaridade. O dendrograma dos pontos indicados na Figura 2.6 pode ser observado na Figura 2.7.

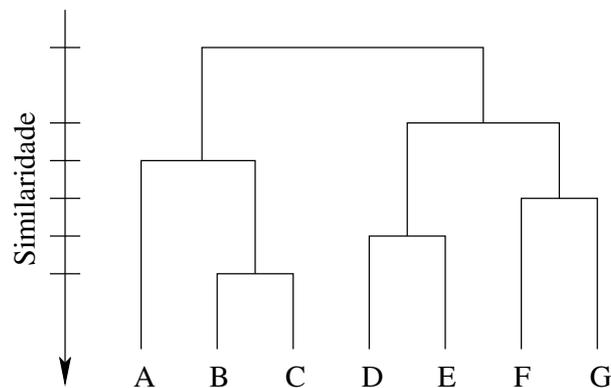


Figura 2.7: Dendrograma obtido pelo algoritmo hierárquico

O dendrograma pode ser partido em diferentes níveis para produzir diferentes categorizações dos dados. A maioria dos algoritmos de categorização hierárquicos são variantes dos algoritmos **ligação-simples** (SOKAL; SNEATH, 1963), **ligação-completa** (KING, 1967) e **variância-mínima** (WARD, 1963; MURTAGH, 1983).

Destes, os algoritmos de ligação-simples e ligação-completa são os mais importantes, pois diferem no modo que estes caracterizam a similaridade entre um par de categorias. No método de ligação-simples, a distância entre duas categorias é a mínima das distâncias entre todos os pares de padrões encontrados a partir das duas categorias (um padrão da primeira categoria, o outro da segunda).

No algoritmo de ligação-completa, a distância entre duas categorias é a máxima de todas as distâncias entre pares das duas categorias. Caso contrário, duas categorias são unidas para formar uma categoria maior baseado no critério de distância mínima. Este tipo de algoritmo produz categorias compactas (BAEZA-YATES, 1992). Entretanto, o algoritmo de ligação-simples tem a tendência de produzir categorias que são alongadas ou disformes (NAGY, 1968).

Nas Figuras 2.8 e 2.9, são apresentadas duas categorias separadas por uma *ponte* de padrões chamados de ruídos (preenchidos com \*). O algoritmo de categorização do tipo ligação-simples produz categorias como mostrado na Figura 2.8 enquanto que o algoritmo do tipo ligação-completa obtém categorias como mostrado na Figura 2.9.

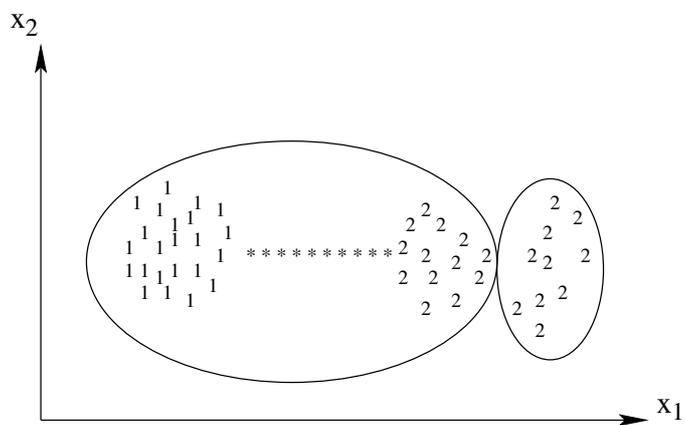


Figura 2.8: Categorização de ligação-simples de um padrão contendo duas classes

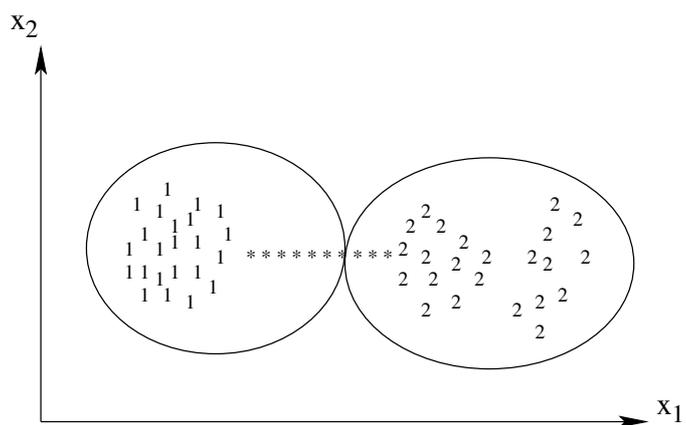


Figura 2.9: Categorização de ligação-completa de um padrão contendo duas classes

As categorias obtidas pelo algoritmo de ligação-completa são mais compactas que as obtidas pelo algoritmo de ligação-simples; a categoria 1 obtida usando o algoritmo de ligação-simples é alongada devido aos padrões de ruído. O algoritmo de ligação-simples é mais versátil do que o algoritmo de ligação-completa (JAIN; MURTY; FLYNN, 1999). Por exemplo, o algoritmo de ligação-simples poderia extrair as categorias concêntricas mostradas na Figura 2.10.

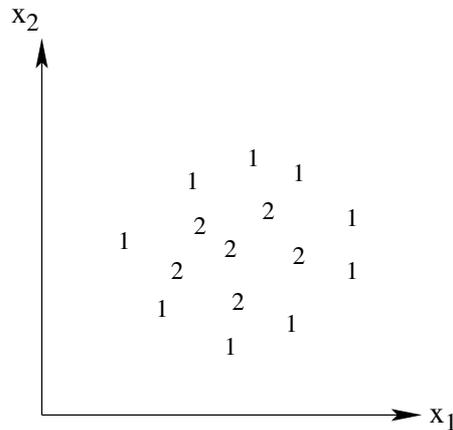


Figura 2.10: Duas categorias concêntricas

Entretanto, (JAIN; DUBES, 1988) observou que o algoritmo de ligação completa produz, em muitas aplicações, hierarquias mais significativas do que o algoritmo de ligação-simples.

O algoritmo de categorização aglomerativa de ligação-simples tem o seguinte funcionamento (JAIN; DUBES, 1988):

1. Colocar cada padrão em sua própria categoria. Construir uma lista de distâncias entre os padrões para todos os pares distintos de padrões e ordenar ascendentemente esta lista.
2. Percorrer a lista ordenada de distâncias formando para cada valor distinto de dissimilaridade  $d_k$  um grafo de padrões cujos pares de padrões mais próximos que  $d_k$  são conectados por uma aresta no grafo. Se todos os padrões são membros de um grafo conectado, então o procedimento pára. Caso contrário, repetir este passo.
3. A saída do algoritmo é uma hierarquia de grafos que podem ser cortados em um nível desejado de dissimilaridade formando uma partição (categoria) identificado pelos componentes conectados neste grafo.

(JAIN; DUBES, 1988) também apresentam o algoritmo para o procedimento de categorização aglomerativa de ligação-completa:

1. Colocar cada padrão em sua própria categoria. Construir uma lista das distâncias entre padrões para todos os pares de padrões distintos e ordenar esta lista ascendentemente.
2. Percorrer a lista ordenada das distâncias, formando para cada valor distinto de dissimilaridade  $d_k$  um grafo dos padrões cujos pares de padrões mais próximos do que  $d_k$  são conectados por uma aresta no grafo. Se todos os padrões são membros de um grafo completamente conectado, então o procedimento pára.

3. A saída do algoritmo é uma hierarquia de grafos que podem ser cortados em um nível de dissimilaridade desejado formando uma partição identificada pelos componentes completamente conectados neste grafo.

Adicionalmente, (JAIN; MURTY; FLYNN, 1999) apresentam de forma genérica o algoritmo de categorização hierárquico aglomerativo com os seguintes passos:

1. Computar a matriz de proximidade contendo a distância entre cada par de padrões.
2. Encontrar o par de categorias mais similar utilizando a matriz de proximidade. Unir estas duas categorias em uma categoria. Atualizar a matriz de proximidade para refletir esta operação de união.
3. Se todos os padrões pertencem a alguma categoria, então parar o procedimento. Caso contrário, fazer o passo 2.

## 2.5.2 Algoritmos de Categorização por Partições

Um algoritmo de categorização por partições obtém apenas uma categoria dos dados ao invés de uma estrutura de categorias, como o dendrograma produzido pela técnica hierárquica. Este tipo de método tem vantagens em aplicações que envolvem grandes conjuntos de dados para o qual a construção de um dendrograma é computacionalmente inviável (JAIN; DUBES, 1988).

Um problema neste tipo de técnica é a escolha do número desejado de categorias (DUBES, 1987). Normalmente, este tipo de técnica gera as categorias otimizando uma função definida localmente (sobre um subconjunto de padrões) ou globalmente (definida sobre todos os padrões). Na prática, o algoritmo é executado múltiplas vezes com diferentes estados de partida e a melhor configuração obtida a partir de todas as execuções é utilizada como a categorização resultante.

### 2.5.2.1 Algoritmos de Erro Quadrático

A função critério mais utilizada em técnicas de categorização por partições é o critério de erro quadrático indicada para categorias isoladas e compactas (JAIN; DUBES, 1988). O erro quadrático para uma categorização  $\mathcal{L}$  de um padrão  $\mathcal{X}$  (contendo  $K$  categorias) é

$$e^2(\mathcal{X}, \mathcal{L}) = \sum_{j=1}^K \sum_{i=1}^{n_j} \|\mathbf{x}_i^{(j)} - \mathbf{c}_j\|^2,$$

onde  $\mathbf{x}_i^{(j)}$  é o  $i$ -ésimo padrão pertencente à  $j$ -ésima categoria e  $\mathbf{c}_j$  é o centróide da  $j$ -ésima categoria.

O algoritmo que emprega o critério de erro quadrático mais utilizado é chamado de  $k$ -médias (McQueen, 1967). Este algoritmo parte com um particionamento inicial aleatório e prossegue associando os padrões às categorias baseado na similaridade entre o padrão e os centros de todas as categorias até satisfazer um critério de convergência (por exemplo, não há mais associações de algum padrão a uma categoria, ou o erro quadrático cessa de decrescer significantivamente após um número de iterações).

Este algoritmo é sensível à seleção da partição inicial e pode convergir para um mínimo local do valor da função critério se esta partição inicial não for apropriadamente escolhida (JAIN; MURTY; FLYNN, 1999).

A Figura 2.11 apresenta sete padrões bidimensionais. Se o algoritmo inicia com os padrões  $A$ ,  $B$  e  $C$  como médias iniciais, o algoritmo gera o particionamento  $\{A, B, C, D, E, F, G\}$  mostrado com elipses. Neste caso, o erro quadrático é muito maior do que para o particionamento ótimo  $\{A, B, C, D, E, F, G\}$ , destacado pelos retângulos, que indica o menor valor global da função critério de erro quadrático para uma categorização contendo três categorias. Esta solução poderia ser obtida pela escolha de  $A$ ,  $D$  e  $F$  como as médias iniciais.

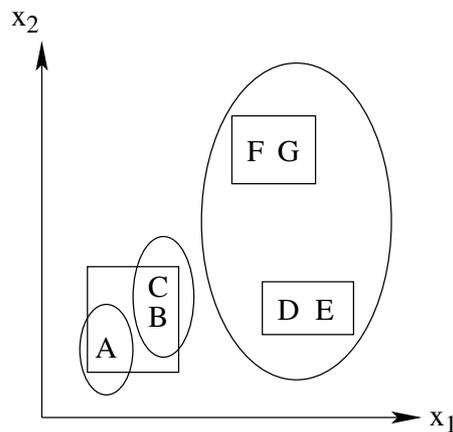


Figura 2.11: O algoritmo de  $k$ -médias é sensível à partição inicial

Em (McQueen, 1967) apresentou-se o algoritmo do método de categorização como erro quadrático, que pode ser resumido nas seguintes etapas:

1. Selecionar uma partição inicial de padrões com um número fixo de categorias e, portanto, de centros de categorias.
2. Associar cada padrão à categoria cujo centro está mais próximo deste e computar os novos centros das categorias. Repetir este passo até alcançar convergência.
3. Unir e partir categorias baseado em uma dada informação heurística e opcionalmente repetir o passo 2.

Em (JAIN; MURTY; FLYNN, 1999), apresentou-se o algoritmo de categorização  $k$ -médias, como segue:

1. Escolher  $k$  centros de categorias de forma a coincidir com os  $k$  padrões escolhidos aleatoriamente, ou  $k$  pontos definidos dentro do volume contendo o conjunto de padrões.
2. Associar cada padrão ao centro de categoria mais próximo.
3. Computar os centros de categorias utilizando os membros das categorias atuais.
4. Se um critério de convergência ainda não for satisfeito, retornar para o passo 2. Os critérios de convergência mais frequentemente encontrados são: inexistência de padrões para serem associados às categorias, ou inalteração do erro quadrático.

Outra variação deste algoritmo permite o particionamento e a união das categorias resultantes. Neste tipo de algoritmo, uma categoria é particionada quando sua variância está acima de um nível pré-determinado; e, duas categorias são unidas quando a distância entre os seus centróides está abaixo da medida pré-estabelecida.

### 2.5.3 Categorização Baseada em Grafos

Segundo (JAIN; MURTY; FLYNN, 1999), a maioria dos algoritmos de categorização baseados em grafos constrói uma árvore geradora mínima (AGM) dos dados (ZAHN, 1971), e então apaga as arestas de maiores comprimentos para gerar as categorias. A Figura 2.12 mostra a AGM obtida a partir de nove pontos bidimensionais.

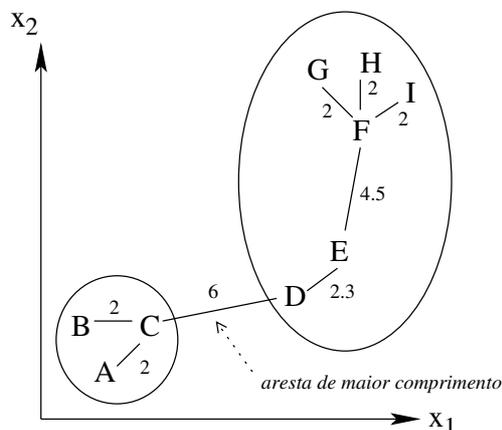


Figura 2.12: Árvore Geradora Mínima para identificar categorias

Ao se quebrar a ligação  $CD$  com 6 unidades de comprimento (a aresta com comprimento Euclidiano máximo) são obtidas duas categorias ( $\{A, B, C\}$  e  $\{D, E, F, G, H, I\}$ ). A segunda categoria, por conseguinte, pode ser dividida em duas categorias partindo da aresta  $EF$ , pois tem o maior comprimento dentre as arestas de sua categoria (4.5 unidades).

As abordagens hierárquicas também estão relacionadas com as categorizações baseadas em grafos. Os algoritmos de ligação-simples geram categorias que são sub-grafos da árvore geradora mínima dos dados (GOWER; ROSS, 1969). As categorias definidas por algoritmos de ligação-completa são sub-grafos máximos completos e estão relacionadas com a coloração dos nodos dos grafos (BACKER; HUBERT, 1976). Uma abordagem orientada a grafos para estruturas não hierárquicas e categorias com sobreposição foi apresentada em (OZAWA, 1985).

### 2.5.4 Categorização Baseada no Vizinho Mais Próximo

A proximidade desempenha papel importante na noção intuitiva de categoria, portanto as distâncias aos vizinhos mais próximos pode servir como base de procedimentos de categorização. Um procedimento interativo foi proposto em (LU; FU, 1978). Ele associa cada padrão não-rotulado à categoria do padrão vizinho rotulado mais próximo, satisfazendo um critério de distância àquele padrão. O processo continua até que todos os padrões sejam rotulados.

### 2.5.5 Categorização Fuzzy

As abordagens tradicionais de categorização geram partições, onde cada padrão pertence à apenas uma, e somente uma, categoria. Portanto, neste tipo de categorização, as categorias são disjuntas. A categorização *fuzzy* estende esta noção para associar cada padrão à toda categoria utilizando uma função (ZADEH, 1965).

A saída de tais algoritmos é uma categorização, mas não uma partição. O algoritmo de categorização *fuzzy* pode ser resumido nas seguintes etapas (JAIN; MURTY; FLYNN,

1999):

1. Selecionar uma partição *fuzzy* inicial de  $N$  objetos em  $K$  categorias formando uma matriz de pertinência  $U$  com  $N \times K$  elementos. Um elemento  $u_{ij}$  desta matriz representa o grau de pertinência do objeto  $\mathbf{x}_i$  na categoria  $\mathbf{c}_j$  e  $u_{ij} \in [0, 1]$ .
2. Encontrar em  $U$  o valor de uma função critério *fuzzy* (por exemplo, uma função de erro quadrático) associada com a categoria correspondente. Uma função critério *fuzzy* pode ser definida da seguinte maneira

$$E^2(\mathcal{X}, U) = \sum_{i=1}^N \sum_{k=1}^K u_{ik} \|\mathbf{x}_i - \mathbf{c}_k\|^2,$$

onde  $\mathbf{c}_k = \sum_{i=1}^N u_{ik} \mathbf{x}_i$  é o centro da  $k$ -ésima categoria *fuzzy*.

Reassociar os padrões às categorias para reduzir o valor da função critério e recomputar  $U$ .

3. Repetir o passo 2 até as entradas em  $U$  não sofrerem modificações significantes.

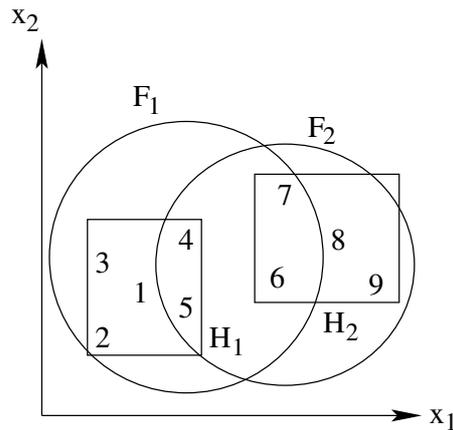


Figura 2.13: Categorias *Fuzzy*

Na categorização *fuzzy*, cada categoria é um conjunto *fuzzy* de todos os padrões. Esta idéia é ilustrada na Figura 2.13. Os retângulos determinam duas categorias convencionais (*hard clusters*) nos dados:  $H_1 = \{1, 2, 3, 4, 5\}$  e  $H_2 = \{6, 7, 8, 9\}$ . Um algoritmo de categorização *fuzzy* pode produzir as duas categorias *fuzzy*  $F_1$  e  $F_2$  indicada pelas elipses.

Os padrões terão valores de pertinência no intervalo  $[0, 1]$  para cada categoria. Por exemplo, a categoria *fuzzy*  $F_1$  poderia ser descrita como:

$$\{(1, 0.9), (2, 0.8), (3, 0.7), (4, 0.6), (5, 0.55), (6, 0.2), (7, 0.2), (8, 0.0), (9, 0.0)\}$$

e  $F_2$  poderia ser descrita como:

$$\{(1, 0.0), (2, 0.0), (3, 0.0), (4, 0.1), (5, 0.15), (6, 0.4), (7, 0.35), (8, 1.0), (9, 0.9)\}.$$

O par ordenado  $(i, \mu_i)$  em cada categoria representa o  $i$ -ésimo padrão e seu valor de pertinência à categoria  $\mu_i$ . Valores de pertinência maiores indicam alta confiança na

associação do padrão à categoria. Uma categorização convencional pode ser obtida a partir de uma partição *fuzzy* de acordo com os limiares (*thresholding*) dos valores de pertinência.

A teoria dos conjuntos *fuzzy* foi aplicada primeiramente à categorização por (RUSPINI, 1969). O algoritmo de categorização *fuzzy* mais utilizado é o algoritmo *fuzzy* de *c*-médias (JAIN; MURTY; FLYNN, 1999).

O problema mais importante na categorização *fuzzy* é o desenvolvimento das funções de pertinência que normalmente são baseadas na decomposição por similaridade e nos centróides das categorias.

### 2.5.6 Representação de Categorias

Em aplicações onde o número de classes ou categorias no conjunto dos dados deve ser descoberta, uma partição do conjunto dos dados é o produto resultante. Neste caso, uma partição dá idéia sobre o grau de separação dos pontos em categorias e sobre o grau de significância do emprego de um classificador supervisionado que assume um dado número de classes no conjunto de dados.

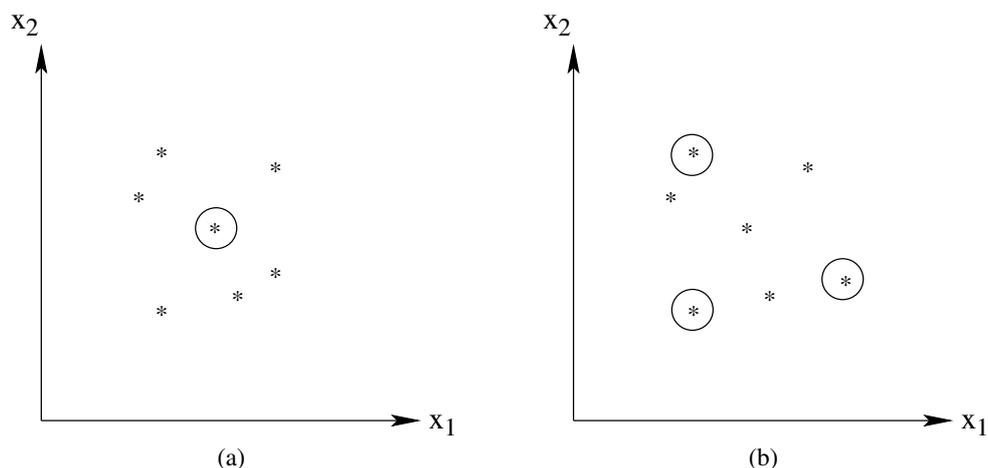


Figura 2.14: Representação de uma categoria por pontos

Entretanto, em muitas aplicações que envolvem tomada de decisões, as categorias resultantes tem que ser representadas ou descritas em forma compacta para denotar uma abstração dos dados. Entre outros trabalhos, (DIDAY; SIMON, 1976) estudaram e sugeriram os seguintes esquemas de representação para categorias:

1. Representar uma categoria pelo seu centróide (Figura 2.14a), ou por um conjunto de pontos distantes na categoria (Figura 2.14b).
2. Representar as categorias usando nodos em uma árvore de classificação, como ilustrado na Figura 2.15.
3. Representar as categorias utilizando expressões lógicas conjuntivas. Por exemplo, a expressão  $[x_1 > 3][x_2 < 2]$  na Figura 2.15 denota a afirmação lógica “ $x_1$  é maior do que 3” e “ $x_2$  é menor do que 2”.

O uso do centróide para representar uma categoria é o esquema mais utilizado e é mais indicado, segundo (JAIN; MURTY; FLYNN, 1999), quando as categorias são compactas.

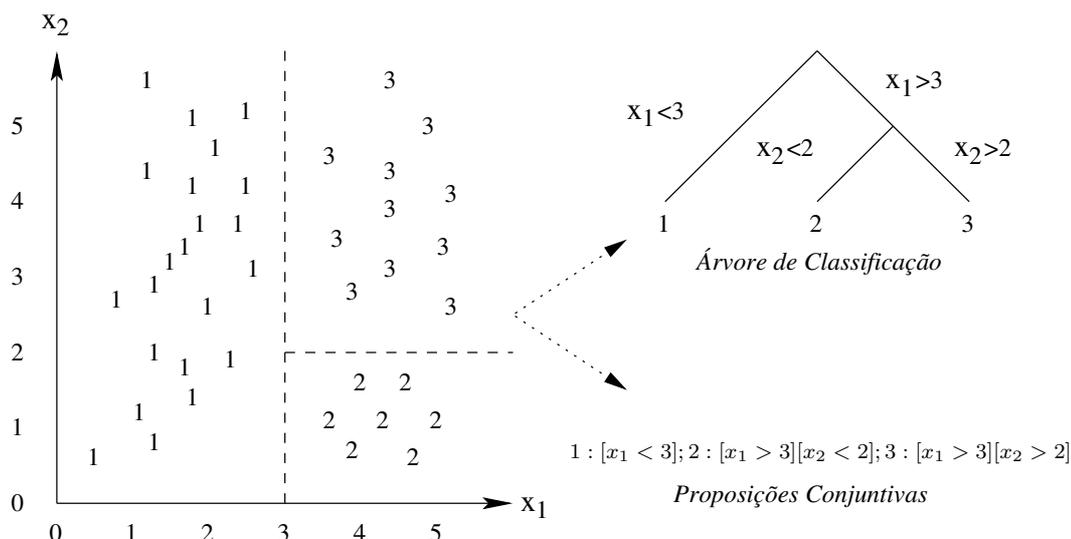


Figura 2.15: Representação de categorias por árvore de classificação ou por proposições conjuntivas

Entretanto, quando as categorias têm forma alongada este esquema não é capaz de representá-las apropriadamente e, nestes casos, utiliza-se uma coleção de pontos limítrofes para capturar a forma da categoria. O número de pontos utilizados para representar uma categoria aumenta com a complexidade da forma desta categoria.

As representações ilustradas na Figura 2.15 são equivalentes. Em um árvore de classificação, todo caminho a partir do nodo raiz até um nodo folha corresponde a uma afirmação conjuntiva. Uma limitação importante do uso típico deste tipo de representação é que esta pode descrever apenas categorias retangulares no espaço de atributos.

A abstração de dados é útil no processo de tomada de decisões pois:

1. Tem-se uma descrição simples e intuitiva das categorias, facilitando a compreensão para o analista. Na categorização conceitual (CULLINGFORD et al., 1982) e na categorização simbólica (GOWDA; DIDAY, 1992) esta representação é obtida trivialmente.
2. Auxilia na compressão dos dados que podem ser explorados por um computador (ANANTHANARAYANA; MURTY; SUBRAMANIAN, 2001). A Figura 2.16a apresenta amostras pertencentes à duas categorias encadeadas (cujos rótulos são 1 e 2).

Segundo (JAIN; MURTY; FLYNN, 1999), uma categorização por partições como o algoritmo de  $k$ -médias não consegue separar estas duas estruturas apropriadamente, por outro lado, o algoritmo de ligação-simples apresenta bons resultados, mas ainda inviáveis computacionalmente. Por isso, (JAIN; MURTY; FLYNN, 1999) apresentam uma abordagem híbrida, combinando as potencialidades destes algoritmos. O algoritmo de  $k$ -médias, computacionalmente mais eficiente, obtém 8 sub-categorias. Cada uma destas sub-categorias podem ser representadas pelos seus centróides como mostrado na Figura 2.16a. Na seqüência, o algoritmo de ligação-simples pode ser aplicado nestes centróides somente para categorizá-los em dois grupos. Os grupos resultantes são mostrados na Figura 2.16b. Neste caso, alcançou-se uma redução dos dados representando as sub-categorias pelos seus centróides.

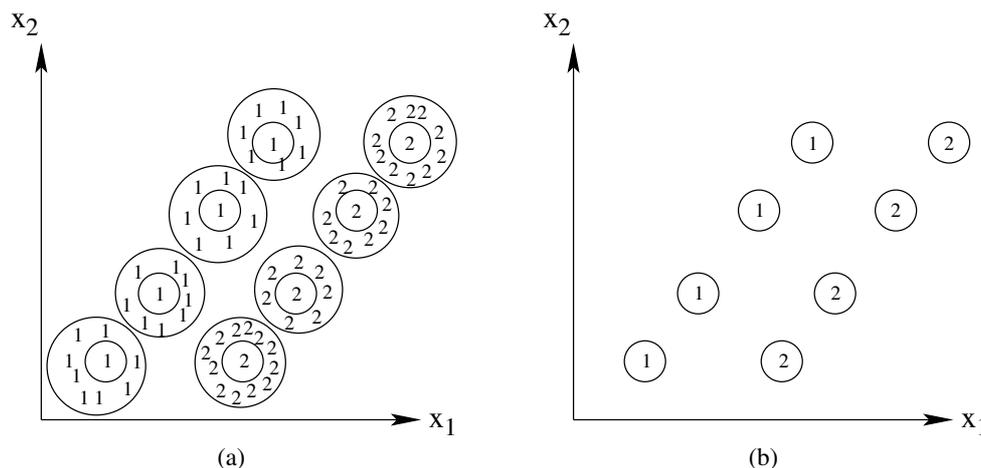


Figura 2.16: Compressão de dados através de categorizações

3. Aumenta a eficiência do processo de tomada de decisão. Em uma técnica de recuperação de documentos baseada em categorizações (SALTON, 1991), uma grande coleção de documentos é categorizado e cada uma destas categorias são representadas utilizando seus centróides.

No momento em que se recuperam os documentos relevantes a uma busca, os termos desta busca são comparados aos centróides das categorias ao invés de a todos os documentos que pertencem a categoria; em outros termos, a abstração dos dados assume o papel da indexação deste documentos, tornando o processo mais eficiente.

### 2.5.7 Categorizações utilizando Redes Neurais Artificiais

Segundo (MAO; JAIN, 1994), as características das redes neurais artificiais que são importantes para a categorização de padrões são:

1. Redes neurais processam vetores numéricos e então requerem que os padrões sejam representados apenas por atributos quantitativos.
2. Redes neurais são arquiteturas de processamento inerentemente paralelo e distribuído.
3. Redes neurais podem determinar seus pesos de interconexão adaptativamente. Mais especificamente, podem atuar como normalizadoras de padrões e seletoras de atributos pela seleção apropriada dos pesos.

Redes neurais competitivas (MAO; JAIN, 1994) são utilizadas freqüentemente para categorizar os dados de entrada. Na aprendizagem competitiva, os padrões similares são agrupados pela rede e representados por apenas uma unidade (neurônio). Este agrupamento é feito automaticamente baseado na correlação dos dados.

Dentre os exemplos mais conhecidos de redes neurais para categorização incluem-se os algoritmos **SOM** e **LVQ** ambos apresentados em (KOHONEN, 1989) e os modelos de ressonância adaptativas de (CARPENTER; GROSSBERG, 1990).

Os padrões são apresentados à entrada e são associados com os nodos de saída. Os pesos entre os nodos de entrada e os nodos de saída são alterados iterativamente (etapa de aprendizagem) até um critério de terminação ser alcançado. Encontra-se aprendizagem

competitiva em redes neurais biológicas. Entretanto, os procedimentos de aprendizagem ou de atualização dos pesos são similares aos presentes nas abordagens clássicas de categorização.

O algoritmo **SOM** gera um particionamento sub-ótimo se os pesos iniciais não são escolhidos apropriadamente. Além disso, sua convergência é controlada por vários parâmetros como a taxa de aprendizagem e a vizinhança do nodo onde a aprendizagem ocorre.

É possível que um padrão de entrada particular possa acarretar diferentes saídas em diferentes iterações, trazendo a idéia de estabilidade de sistemas de aprendizagens.

O sistema é dito ser estável se nenhum padrão nos dados de aprendizagem muda a sua categoria após um número finito de iterações de aprendizagem. Este problema está associado ao problema da plasticidade, que é a habilidade do algoritmo de se adaptar aos novos dados. Para atingir estabilidade, a taxa de aprendizagem deve tender a zero com o progresso das iterações e isto afeta a plasticidade.

### 2.5.8 Abordagens Evolucionárias para Categorização

As abordagens evolucionárias, motivadas pela evolução natural, fazem uso de operadores evolucionários e uma população de soluções para obter o particionamento ótimo global dos dados. As soluções candidatas ao problema de categorização são codificadas como cromossomos. Os operadores evolucionários mais utilizados são: seleção, recombinação e mutação.

Cada um destes operadores transformam um ou mais cromossomos de entrada em um ou mais cromossomos de saída. Uma função de adaptação é aplicada em um cromossomo e determina a probabilidade do cromossomo sobreviver na próxima geração. A seguir, apresenta-se uma descrição geral de um algoritmo evolucionário para categorização (JAIN; MURTY; FLYNN, 1999):

1. Escolher aleatoriamente uma população de soluções. Cada solução corresponde a um particionamento válido dos dados em  $k$  regiões. Associar um valor de adaptação a cada solução. Tipicamente, a adaptação é inversamente proporcional ao valor do erro quadrático. Uma solução com menor erro quadrático terá um valor maior de adaptação.
2. Utilizar os operadores evolucionários de seleção, recombinação e mutação para gerar a próxima população de soluções. Avaliar os valores de adaptação destas soluções.
3. Repetir o passo 2 até uma condição de término ser satisfeita.

As técnicas evolucionárias mais conhecidas são os Algoritmos Genéticos (AGs) (HOLLAND, 1975; GOLDBERG, 1989) e a Programação Evolutiva (PE) (FOGEL; OWENS; WALSH, 1966). As soluções são tipicamente representadas como *strings* binárias nos Algoritmos Genéticos. Nos algoritmos genéticos, um operador de seleção propaga soluções a partir da geração corrente para a próxima geração baseado em seu grau de adaptação.

A seleção emprega um esquema probabilístico tal que as soluções com maior grau de adaptação tem probabilidade maior de serem reproduzidas. Dentre uma grande variedade de operadores de recombinação, o operador de cruzamento *crossover* é o mais utilizado.

Este operador toma como entrada um par de cromossomos (chamados de pais) e produz um novo par de cromossomos (chamados de filhos) como ilustra a Figura 2.17.

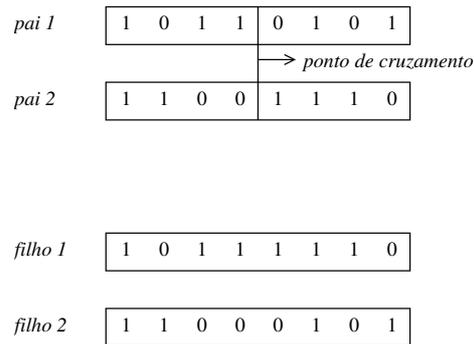


Figura 2.17: Operação de Cruzamento

Por exemplo, na Figura 2.17 os pais são as *strings* binárias ‘10110101’ e ‘11001110’. Os segmentos nos dois pais após o ponto de cruzamento (entre a quarta e a quinta posições) são trocadas para produzir os cromossomos dos novos filhos.

A mutação toma como entrada um cromossomo e produz um novo cromossomo apenas complementando o valor do *bit* em uma posição escolhida aleatoriamente no cromossomo de entrada. Por exemplo, a *string* ‘11111110’ é gerada pela aplicação do operador de mutação no segundo *bit* da *string* ‘10111110’. Ambos operadores são aplicados com probabilidades especificadas previamente.

Os algoritmos genéticos representam os pontos como *strings* binárias em um espaço de buscas e fazem a exploração deste espaço de busca através do operador de cruzamento. A mutação é utilizada neste tipo de algoritmo para se certificar que nenhuma parte do espaço de busca não foi explorado. Esta abordagem tem sido utilizada para solucionar problemas de categorização, os entendendo como uma minimização da função critério de erro quadrático. No trabalho de (FOGEL, 1994), encontram-se estudos sobre a convergência deste tipo de abordagem.

Os algoritmos genéticos fazem uma busca global por soluções em contraste com a maioria dos procedimentos de categorização que executam uma busca localizada. Neste último tipo de busca, a solução a ser obtida na próxima iteração do procedimento está na vizinhança da solução corrente. Neste sentido, o algoritmo de *k*-médias, os algoritmos de categorização *fuzzy* e redes neurais são todos técnicas de busca localizada.

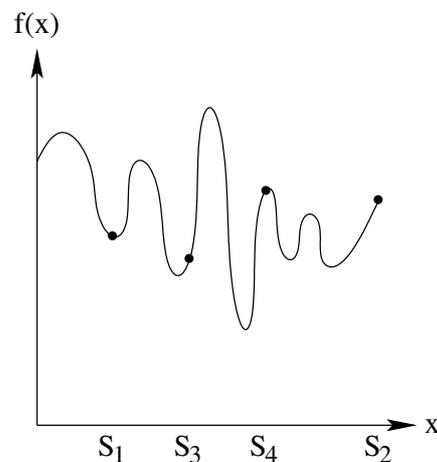


Figura 2.18: Algoritmos Genéticos executam busca globalizada

No caso dos algoritmos genéticos, os operadores de cruzamento e mutação pode pro-

duzir novas soluções que são completamente diferentes das soluções correntes (vide Figura 2.18).

Sejam um escalar  $x$  codificado com uma representação binária de 5 bits,  $S_1$  e  $S_2$  dois pontos em um espaço de busca unidimensional. Os valores decimais de  $S_1 = 8$  e  $S_2 = 31$  e suas representações binárias são  $S_1 = 01000$  e  $S_2 = 11111$ . Ao se aplicar o operador de cruzamento à estas *strings*, cujo ponto de cruzamento entre o segundo e terceiro bit mais significativo, um novo par de cromossomos  $S_3$  e  $S_4$  serão produzidos. Neste caso, os novos cromossomos terão a configuração  $S_3 = 01111$  e  $S_4 = 11000$ , cujos valores decimais são 15 e 24, respectivamente. Da mesma forma, a *string* binária 11111 pode ser produzida pela mutação do bit mais significativo da *string* binária 01111.

Estes saltos ou diferenças entre pontos em gerações subsequentes são muito maiores do que os produzidos por outras abordagens. Além disso, é possível ver o problema de categorização utilizando algoritmos genéticos diretamente como um problema de otimização que localiza os centróides ótimos das categorias do que encontrar um particionamento ótimo.

### 2.5.9 Abordagens Baseadas em Buscas

Neste tipo de abordagem são utilizadas técnicas de busca para obter o valor ótimo da função critério. Estas técnicas são divididas em técnicas de busca determinísticas e estocásticas.

As técnicas de busca determinísticas, oriundas de abordagens gulosas, garantem um particionamento ótimo pela execução de uma enumeração exaustiva. Por outro lado, as técnicas de busca estocásticas geram um particionamento próximo do ótimo razoavelmente rápido e garantem a convergência ao ótimo assintoticamente.

Dentre as técnicas apresentadas, a abordagem evolucionária é caracterizada como estocástica e as demais são consideradas determinísticas. Em (ROSE; GUREWITZ; FOX, 1991) foi apresentada uma abordagem determinística utilizando recozimento (*annealing*). Esta abordagem emprega uma técnica de recozimento cuja superfície de erro é suavizada, mas a convergência ao ótimo global não é garantida.

A abordagem do recozimento simulado (KIRKPARTICK; GELATT; VECCHI, 1983) é uma técnica de busca estocástica seqüencial, cujo procedimento é evitar soluções que correspondam a um ótimo local da função objetivo. Isto é efetuado aceitando para a próxima geração, com alguma probabilidade, uma nova solução de menor qualidade (medida pela função critério). A probabilidade de aceitação é governada por um parâmetro crítico chamado temperatura (fazendo analogia com o recozimento de metais), especificado em termos de valores inicial e final da temperatura.

Em (JAIN; MURTY; FLYNN, 1999), apresenta-se um algoritmo de alto nível de abstração da categorização baseada no recozimento simulado com os seguintes passos:

1. Selecionar aleatoriamente uma partição inicial  $P_0$  e computar o valor do erro quadrático,  $E_{P_0}$ . Selecionar valores para os parâmetros de controle, temperaturas inicial  $T_0$  e final  $T_f$ .
2. Selecionar um vizinho  $P_1$  de  $P_0$  e computar seu valor de erro quadrático,  $E_{P_1}$ . Se  $E_{P_1}$  é maior do que  $E_{P_0}$ , então associar  $P_1$  a  $P_0$  com uma probabilidade de dependência de temperatura. Caso contrário, associar  $P_1$  a  $P_0$ . Repetir este passo um número fixo de iterações.

3. Reduzir o valor de  $T_0$ . Se  $T_0$  é maior do que  $T_f$ , então o passo 2 deve ser refeito, caso contrário encerrar a execução do algoritmo.

Este algoritmo pode ser lento para alcançar a solução ótima pois os resultados ótimos requerem que a temperatura seja decrescida gradualmente, iteração à iteração.

Assim como o recozimento simulado, a busca Tabu (GLOVER, 1986) é um método designado para cruzar os limites da praticabilidade ou da otimalidade local e sistematicamente definir ou excluir restrições para permitir a exploração de regiões consideradas proibidas pelo algoritmo.

## 2.6 Considerações sobre o Modelo *ICTM* como Técnica Categorizadora

O modelo *ICTM*, que será apresentado no capítulo 4, não utiliza métricas como as definidas nas seções anteriores para determinar a categoria em que o ponto (célula) se encontra, mas sim a análise da monotonicidade da função que modela cada propriedade ou atributo presente.

Portanto, o *ICTM* não faz a categorização baseado na análise da similaridade (ou proximidade) entre categorias como as técnicas tradicionais, pois a análise da monotonicidade da função que modela a propriedade permite apenas três estados possíveis de declividade (sinal positivo, sinal negativo, ou nulo). Entretanto, apenas estes estados permitem a categorização apropriada, pois são suficientes para a determinação daqueles pontos que são limítrofes.

Quanto à estrutura e funcionamento algorítmico, o *ICTM* pode ser considerado como um método construtor (diferentemente dos conceitos de aglomerativo e divisor apresentados na seção 2.5). A cada análise do estado da célula em relação à sua vizinhança, a categorização vai sendo construída – o processo é paralelo e a determinação dos estados ocorrem em tempos diferentes conforme a demanda dos processadores – e só será uma categorização ao final de todo processo, por tanto, não existem sub-soluções no modelo *ICTM*, no sentido de aproximações da solução final, comuns nas técnicas de categorização.

No modelo *ICTM*, cada atributo (ou propriedade) – presente em uma camada – é analisado independentemente e a projeção destas categorizações individuais em uma camada base permite a identificação da dependência mútua entre estes atributos. Este tipo de abordagem pode ser considerada como uma hibridação das formas de manipulação de atributos chamados de *monothetic* e *polythetic* (seção 2.5).

Este procedimento de projeção, aliado ao recurso de ponderação das camadas (valores no intervalo  $[0, 1]$ ), permite efetuar análises como às apresentadas na abordagem *fuzzy*, resultando em graus de pertinência de um ponto à uma dada categoria. A abordagem tradicional, por conjuntos, é um caso particular neste procedimento de projeção, quando os pesos assumem apenas os valores 0 ou 1.

O *ICTM* trabalha sobre uma malha regular – onde cada posição nesta malha é referenciada no sistema de coordenadas cartesianas – e não apresenta as deficiências comentadas nas seções anteriores no que se refere às categorizações tanto curvilíneas (Figura 2.2) quanto concêntricas (Figura 2.10), pois a análise do estado da célula é feita em todas as direções, conforme a definição de vizinhança utilizada e do raio de células que formam esta vizinhança.

Uma categoria no *ICTM* é determinada por todos os pontos limites desta categoria –

ponto onde muda a configuração da declividade – e a categoria é representada por todos os pontos internos à esta região limitada. O *ICTM* não utiliza a noção de centróide para referência às categorias, mas esse conceito poderia ser agregado ao modelo através da noção de “centro de massa” das regiões categorizadas.

Quanto à compressão dos dados, de certo modo o *ICTM* permite a compactação das categorias, dando uma descrição abstrata dos dados presentes nas células quando representa uma categoria por uma coleção de intervalos. Esta aproximação interior de uma categoria (vide Definição 6.9), utilizada na extração de fatos, descreve intuitivamente categorias irregulares (disformes, alongadas ou compactas) ou mesmo concêntricas.

## 3 O ESTADO DA ARTE DO PROCESSAMENTO DE DADOS GEOGRÁFICOS

### 3.1 Sistemas de Informações Geográficas

#### 3.1.1 Histórico

Os sistemas de informações geográficas (SIG) surgiram na transição das décadas de 60 e 70 e têm se tornado ferramentas bastante úteis nas diversas áreas de conhecimento (MENEGUETTE, 2000). É interessante observar que o emprego do conceito de computação para o processamento de dados geográficos remonta ao século passado, quando o Censo Americano, utilizou cartões perfurados e uma máquina tabuladora para agilizar as atividades relativas ao censo de 1890, tendo finalizado após três anos. Isto resultou em grande avanço comparado ao censo anterior, de 1880, que demorou oito anos para ser efetuado convencionalmente.

Entretanto, a falta de ferramentas matemáticas adequadas para descrever quantitativamente a variação espacial era ressentida pela comunidade científica. A seguir, serão apresentadas as principais etapas históricas dos SIG's são as seguintes, conforme (BROUGH, 1989).

Os primeiros modelos matemáticos apareceram nas décadas de 30 e 40, juntamente com os métodos estatísticos para análise de séries temporais.

Apenas com a disponibilidade do computador digital (década de 60) ocorreu o desencadeamento da utilização de ferramentas computacionais adequadas para o mapeamento temático quantitativo e análise espacial.

O primeiro SIG surgiu em 1964 no Canadá (*Canada Geographic Information System*) por iniciativa do Dr. Roger Tomlinson, que embora tenha construído os módulos básicos de *software*, impulsionando o desenvolvimento de *hardware* e elaborado uma complexa base de dados, só publicou seus trabalhos uma década depois.

Somente ao final da década de 70 é que o desenvolvimento dos SIG's se estabeleceu solidamente, favorecendo o surgimento da versão comercial dos primeiros sistemas no início da década de 80.

Os governos americano, canadense e alguns europeus apoiavam financeiramente iniciativas voltadas tanto à Cartografia Assistida por Computador (CAC). E neste período passou-se a tornar disponíveis ao público bases de dados digitais, tais como os modelos digitais de elevação (DEM's – *Digital Elevation Models*)(MENEGUETTE, 2000).

Atualmente, percebe-se um crescimento acentuado das aplicações de SIG's, devido à disseminação do PC, além da introdução de tecnologia de relativo baixo custo e alta capacidade de performance das *Workstations*.

Além do serviço de venda de mapas analógicos aos usuários, surgiu uma outra al-

ternativa: o arrendamento de dados através do estabelecimento de contratos, definindo inclusive a frequência das suas atualizações.

Com o surgimento dos Sistemas de Informação, associou-se à informação o conceito de valor agregado, que é obtido ao se reunir de forma ordenada conjuntos de dados que previamente estavam não relacionados, e cuja combinação pode ser utilizada para a realização de tarefas adicionais.

Segundo (RODRIGUES, 1988), *Geoprocessamento* é a tecnologia de coleta e tratamento de informações espaciais e de desenvolvimento de sistemas que as utilizam. Ainda, (RODRIGUES, 1988, 1990) apresenta uma classificação dos Sistemas de Geoprocessamento como:

**sistemas aplicativos:** são conjuntos de programas que realizam operações associadas a atividades de projeto, análise, avaliação, planejamento etc, em áreas tais como transportes, mineração, hidrologia, urbanismo; geralmente, são voltados à representação espacial e à realização de operações sobre estas representações; são sistemas voltados à entrada de dados, saída de dados e realização de tarefas em projeto assistido por computador e mapeamento automatizado;

**sistemas de informações:** são *softwares* que desempenham a coleta, tratamento e apresentação de informações espaciais. De maneira mais geral, o SIG é o conjunto de *software*, *hardware*, procedimentos de entrada e saída dos dados, fluxos de dados do sistema, normas de codificação de dados, normas de operação, pessoal técnico etc que desempenham as funções de coleta, tratamento e apresentação de informações.

**sistemas especialistas:** sistemas computacionais que empregam o conhecimento na solução de problemas que normalmente demandariam a inteligência humana; emulam o desempenho de um especialista atuando em uma dada área do conhecimento.

Define-se informação geográfica como o conjunto de dados ou valores que podem ser apresentados em forma gráfica, numérica ou alfanumérica, e cujo significado contém associações ou relações de natureza espacial.

Entretanto, na literatura especializada, não há concordância no estabelecimento da classificação dos sistemas de geoprocessamento, e na maioria das vezes, apresentam-se múltiplas características com predominância de um conjunto particular de funções (RODRIGUES, 1990). Por isso, faz-se necessário apresentar a diferenciação feita por (KORTE, 1994) entre CADD-CAM, AM-FM e GIS:

**CADD:** *Computer Aided Design and Drafting*, ou Projeto Assistido por Computador, é uma tecnologia normalmente empregada pelo CAM (*Computer Assisted Mapping*), ou Mapeamento Assistido por Computador, para a produção de mapas em substituição ao processo cartográfico tradicional.

Os dados são organizados em camadas (*layers*), empregados para organizar as feições do mapa por temas (*themes*). A utilização do CAM reduz o tempo de produção de mapas e possibilita a economia de recursos financeiros quando comparado aos processos cartográficos tradicionais. Assim, as atualizações se tornam mais simples e rápidas, uma vez que se modifica somente o elemento selecionado sem causar alteração nos demais.

Entretanto, CAM não é um sistema muito adequado para realizar análises; pois, as relações espaciais não são definidas na estrutura de dados, requerendo processamentos adicionais (mais demorados) para a inspeção de tais relações.

**AM-FM:** Mapeamento Automatizado (*Automated Mapping*) e Gerenciamento de Serviços de Utilidade Pública (*Facility Management*) baseiam-se também em tecnologia CADD.

Entretanto, a apresentação gráfica geralmente não é tão precisa e detalhada quanto em sistemas CAM porque sua ênfase está centrada no armazenamento, na análise e na emissão de relatórios.

As relações entre os componentes do sistema de utilidade pública são definidas como redes (*networks*) que são associadas à atributos. Todavia, relações espaciais não são definidas nestes sistemas.

**GIS:** Sistema de Informação Geográfica (*Geographic Information System*) é recomendado para a análise de dados geográficos.

Este tipo de sistema difere dos dois sistemas anteriormente apresentados porque define as relações espaciais entre todos os elementos dos dados. Esta relação espacial é conhecida como **topologia dos dados** e pode indicar um conjunto muito grande de informações, além da descrição da localização e geometria das feições cartográficas.

A Topologia também descreve como as feições lineares estão conectadas, como as áreas são limitadas e quais áreas são contíguas. Para definir a topologia do mapa, o SIG usa uma estrutura de dados espacial, empregando nodos (*nodes*), arcos (*lines*) e áreas (*polygons*).

O SIG também contém dados em atributos e dados geométricos espaciais que são associados com os elementos topológicos para representar informações descritivas.

Além disso, o SIG permite o acesso aos dados (espaciais e atributos) ao mesmo tempo e, por isso, o dado atributo pode ser relacionado com o dado espacial e vice-versa.

Outros autores apresentam definições que ajudam a compreender a complexidade funcional e estrutural de um SIG:

- Sistema Geográfico de Informação (SGI) constitui o tipo de estrutura mais importante em termos de viabilização do Geoprocessamento. O SGI é um conjunto de procedimentos computacionais que operando sobre bases de dados geocodificados ou, mais evolutivamente, sobre bancos de dados geográficos executa análise, reformulações e sínteses sobre os dados ambientais disponíveis (SILVA; SOUZA, 1987).
- Sistemas de Informações Geográficas são modelos do mundo real úteis a um certo propósito; subsidiam o processo de observação (atividades de definição, mensuração e classificação), a atuação (atividades de operação, manutenção, gerenciamento, construção etc) e a análise do mundo real (RODRIGUES; QUINTANILHA, 1991).
- SIG's são constituídos por uma série de processos de análise para focalizar o relacionamento de determinado fenômeno da realidade com sua localização espacial. Utilizam uma base de dados computadorizada que contém informação espacial, sobre a qual atuam uma série de operadores espaciais; baseia-se numa tecnologia de armazenamento, análise e tratamento de dados espaciais, não-espaciais e temporais e na geração de informações correlatas (TEIXEIRA; MORETTI; CHRISTOFOLLETTI, 1992).

- SIG's integram numa única base de dados: informações espaciais provenientes de dados cartográficos, dados de censo e de cadastro urbano e rural, imagens de satélite, redes e modelos numéricos de terreno. Além disso, combinam as várias informações, através de algoritmos de manipulação, para gerar mapeamentos derivados; consultar, recuperar, visualizar e plotar o conteúdo da base de dados geocodificados (CAMARA, 1993).

A seguir, comentam-se as principais atividades envolvidas na utilização de um SIG.

### 3.1.2 Coleta de Dados

Um SIG permite a integração de dados que foram recolhidos em diferentes tempos, escalas, utilizando diferentes métodos de coleta de dados (RODRIGUES, 1988). São fontes de dados: mapas (papel ou transparências), dados escritos, arquivos digitais contendo informações planialtimétricas e temáticas e informações armazenadas na memória do especialista. A integração destes dados de diferentes formatos, de tempos variados e em diferentes escalas, seria inviável economicamente e temporalmente sem a utilização do SIG.

Os dados podem ser importados para o SIG através da digitação dos dados textuais, ou da digitalização através de mesa *scanner* dos mapas existentes. Entretanto, estes métodos são limitados porque os mapas originais frequentemente são antiquados, possuem erros de transcrição e podem não ter a escala apropriada.

Hoje, a coleta está cada vez mais sofisticada em função da maior diversidade do conhecimento humano e de tecnologias e equipamentos mais precisos, como: sensores remotos, fotogrametria e levantamentos a campo executados com GPS (*Global Positioning System* – Sistema de Posicionamento Global) e estações TOTAL.

A entrada de dados imprecisos causa interpretações errôneas ou sem sentido da informação derivada de um SIG, por isso, recorrentemente se afirma que “um SIG é tão bom quanto as informações que contém”.

O GPS provê dados exatos e atualizados instantaneamente, a um custo relativamente baixo. Ao se utilizar GPS, pode-se definir um dicionário de dados e recolher os atributos no campo ao mesmo tempo que se recolhem dados de posição. Desta forma, eliminam-se os erros de transição (entrada de dados no sistema) e garante-se a atualização das informações da base de dados.

A fotogrametria obtém informações confiáveis sobre objetos e sobre o meio ambiente com o uso de processos de registro, medições e interpretações das imagens fotográficas e padrões de energia eletromagnética.

A necessidade contínua de se obter novos dados espaciais representa um dos maiores custos na utilização de um SIG. Com uma ferramenta de coleta de dados, o SIG simplifica a coleta de dados iniciais e também garante que a informação esteja sempre atualizada.

### 3.1.3 Tipos de Dados

Dados **cartográficos** e dados **não-gráficos** são os dois tipos de dados mais recorrentes em um SIG (MENEGUETTE, 2000).

Os dados cartográficos são as informações geográficas oriundas de mapas armazenadas digitalmente. Cada entidade destes mapas é classificada como pontos, linhas e polígonos, estes últimos também são chamados de áreas ou regiões.

Um ponto representa uma característica que necessita somente uma localidade geográfica (por exemplo, latitude-longitude) para referenciá-la. Ou seja, um ponto pode

representar a posição de um estações meteorológicas, de poços e postes.

Uma linha (arco) é formada por uma série de pontos conectados unidimensionalmente, possuindo apenas comprimento (ausência de largura). Por exemplo, riachos, estradas e rastros de animais podem ser características representadas por linhas no SIG.

Um polígono é uma área cercada por linhas e a área compreendida pelo polígono possui comprimento e largura (bidimensional). Por exemplo, áreas com mesmo tipo de solo, regiões para plantação de arroz e banhados são representadas por polígonos.

Os dados não-gráficos consistem de informações descritivas sobre as características (pontos, linhas e áreas) armazenadas numa base de dados e referenciada em um mapa.

Estas informações descritivas são nomeadas de atributos. Um atributo comum à todas as características é a situação geográfica, ao qual pode dar-se o nome de atributo SITUAÇÃO. Outros atributos dependem do tipo de característica e de que características são importantes para um propósito ou aplicação em particular. Por exemplo:

- uma parcela de terreno possui um proprietário, um tamanho e um uso;
- um poço de petróleo é de um determinado tipo e possui um índice de fluxo diário;
- uma estrada possui um nome, um tipo de superfície e pode possuir uma rota ou número de designação.

Cada uma destas características pode ser identificada especificamente num SIG ao dar-lhe um nome de atributo tal como DONO, USO-TERRENO, ou NOME-ESTRADA. O conjunto de valores assumido por cada atributo possui é chamado de **domínio**. Por exemplo, o domínio para o atributo NOME-ESTRADA são todos os nomes de estrada na área de interesse.

### 3.1.4 Estruturas de Dados

As estruturas de dados existentes em SIG's são a topologia e as camadas (KORTE, 1994). A estrutura de topologia refere-se à conexão das características das relações espaciais fundamentais. A topologia fornece a lógica que conecta pontos, linhas e polígonos. As camadas indicam apenas o modo que o SIG estratifica seus dados.

A informação topológica descreve a relação espacial entre as características e geralmente não é modificada pelo profissional que opera o SIG.

Ao fazer a descrição da posição de algum objeto, usualmente diz-se que está à esquerda, ao lado de, ou a determinada distância de um dado objeto. Esta definição não é precisa o suficiente para um SIG. Para se realizar uma análise espacial são requeridas as definições precisas fornecidas pela topologia.

A topologia define a relação posicional das características de acordo com as suas propriedades, por exemplo: as informações sobre que rótulos estão vinculados a cada característica, como os pontos estão ligados uns aos outros e, que pontos e linhas formam um polígono em particular.

Esta informação topológica, armazenada no SIG, permite que sejam efetuadas relações espaciais, tais como: a sobreposição de polígonos, o isolamento de polígonos, determinar se uma linha está dentro de um polígono e determinar a proximidade entre características. Um sistema SIG que realiza manipulações e análises não topológicas (como ocorre em sistemas CAD) são limitados.

Geralmente, um SIG permite a separação das informações de um mapa em categorias lógicas chamadas de camadas, temas, níveis de informação ou planos de informação (PI).

Os planos de informação geralmente contém informações sobre um tipo de característica (como áreas de tipo de solo) ou sobre um pequeno grupo de características relacionadas (por exemplo, recursos de utilidade pública como telefone, gás e linhas de transmissão de energia).

Os dados de um mapa são separados logicamente em camadas para que assim possam ser manipulados e analisados espacialmente, isoladamente ou em combinação com outras camadas. Para se obter resultados analíticos significativos, os planos de informação devem estar referenciados geograficamente entre si por um sistema de coordenadas comum.

Estes planos podem ser combinados a fim de criar mapas compostos a partir da sobreposição destas, de maneira análoga à sobreposição de transparências em um retro-projetor. Durante esta análise criam-se novas sobreposições com a combinação matemática de sobreposições já existentes. Pode-se utilizar estas combinações para a criação de cenários alternativos.

### 3.1.5 Análise de Dados

A análise dos dados permite derivar informação a partir do conteúdo da base de dados do sistema. Este análise de dados compreende (TEIXEIRA; MORETTI; CHRISTOFOLETTI, 1992):

- Sobreposição espacial das características;
- Questões à base de dados;
- Reclassificação, combinação e eliminação de características;
- Cálculo de proximidade de características.

Por exemplo, o operador do SIG necessita saber que tipos de casas são construídas em um dado tipo de solo. Para responder a este tipo de pergunta o sistema utiliza dois conjuntos de dados: os tipos de solo e a localização geográfica das casas construídas na área solicitada.

Estes dados são fornecidos pelas duas camadas de informações e o SIG as combina para formar uma nova fonte de informações. A ligação entre as duas camadas de dados é a localização geográfica (latitude-longitude) de cada entidade.

Este exemplo é bastante simples e favorece ao operador visualizar a seleção das casas que estão construídas na área indicada. No entanto, na maioria dos casos manipulam-se áreas muito grandes com dados bastante detalhados e, por isso, a tarefa pode ser extremamente difícil.

A visualização de dados compreende a criação de imagens e mapas, a visualização dos dados das características espaciais, a criação de produtos cartográficos e a combinação de todos estes elementos para a visualização na tela, ou para impressão das tabelas ou mapas, ou em arquivos para a utilização em outros programas.

## 3.2 Geomática

Segundo (GAGNON; BÉDARD, 1996), durante a última década, a área de mapeamento tem sido incluída em um novo paradigma. No Canadá, como na maioria dos países, tem-se adotado o nome Geomática (*Geomatics*) para identificar este campo de

atividade de mapeamento. A evolução do contexto de produção, otimização e gerenciamento de dados e informações espaciais tem sido a base da redefinição das atividades numa perspectiva global de planejamento em diferentes áreas.

Dessa forma, a Geomática representa a evolução do campo de atividades de levantamentos e mapeamento, congregando as disciplinas mais tradicionais (cartografia, geodésia, topografia) com as novas tecnologias (sensoriamento remoto, processamento digital de imagens, SIG, GPS e fotogrametria digital) e os novos campos de aplicação surgidos (entre eles a mineração, agricultura, meio ambiente, transportes, informática, turismo, saúde, telecomunicações etc).

A geomática engloba pelo menos quatro categorias técnicas relacionadas ao levantamento, mapeamento e planejamento (GAGNON; BÉDARD, 1996):

- Técnicas para coleta de informação espacial: cartografia, sensoriamento remoto, GPS, topografia convencional, fotogrametria, levantamentos de dados geográficos;
- Técnicas para armazenamento de informação espacial: banco de dados (orientados a objetos, relacional, hierárquico etc);
- Técnicas para tratamento e análise de informação espacial: modelagem de dados, geoestatística, aritmética, lógica, funções topológicas, redes;
- Técnicas para uso integrado de informação espacial: SIG's, LIS (Sistemas de Informação de Terrenos – *Land Information Systems*), AM/FM e CADD.

Para ressaltar algumas das principais aplicações na área de Geomática, assim como para o geoprocessamento, podemos citar os projetos de SIG's para (COSTA, 2003):

- elaboração de mapas urbanos básicos
- elaboração/atualização de mapas de arruamentos
- cadastro, mapeamento de cadastro urbano e rural
- fonte de apoio para trabalhos com GPS
- uso e ocupação do solo
- regularização dos limites de propriedades, demarcação de pequenas glebas
- previsão de safras, controle de pragas e agricultura de precisão
- estimativa de potencial econômico, projetos de desenvolvimento sustentável

A criação e a difusão do uso da *internet*, aliado ao avanço no desenvolvimento de redes de computadores cada vez mais rápidas, indicaram a possibilidade da aplicação do processamento distribuído aos SIG's.

As potenciais aplicações do processamento distribuído no geoprocessamento estão ligadas, na maioria das vezes, a sistemas de decisão on-line, por exemplo (SILVA, 2001; BURROUGH; MCDONNELL, 1998; MACKAY, 1996; GRAYSON; BLOSCHL; MORE, 1994):

1. Na logística:

- a escolha da melhor rota a ser seguida pelos caminhões na distribuição de produtos.
- mapas turísticos e de localização para viajantes, como o serviço prestado pelo *site* <http://www.apontador.com/>, que possibilita pesquisar o mapa de uma cidade (capitais) para localizar ruas, endereços, serviços e estabelecimentos mais próximos (como bancos, farmácias ou supermercados), para encontrar a melhor rota entre dois endereços.
- a implantação de serviços WAP (*Wireless Application Protocol*), para usuários de Internet por celular, para localização de endereços e roteirização. Se o celular for munido de um receptor GPS, as opções tornam-se ainda maiores: o usuário identifica exatamente o ponto onde se encontra e ainda usufruir de serviços de roteamento para descobrir a melhor maneira de chegar ao local desejado.

## 2. Na agricultura:

- para o processamento de informações climáticas, para o suporte do trabalho de plantação e colheita.
- para previsão de veranicos, geadas, secas prolongadas, excessos de água ou condições climáticas que favoreçam o surgimento de pragas e doenças. Reduzindo a possibilidade de prejuízos e perdas na produção de culturas climaticamente regionalizadas. O agricultor, armado de informações básicas atualizadas, tem melhores condições e maiores chances de sucesso no combate às adversidades climáticas.
- para o manejo da água e do solo, se mal executado pela falta de informação ou se executado fora de época, pode ser danoso e impedir uma maior produtividade. O agricultor pode escolher as épocas adequadas para plantio e colheita, o momento e a quantidade certa de irrigar, o momento correto de tratar o solo e a forma de se proteger das pragas.

## 3. Na distribuição de energia:

- como ferramenta de comunicação interna, no que tange aos sistemas interligados e mapas georreferenciados, permitindo a disseminação de informação de forma rápida, precisa e segura, com uma visão facilitada dos pontos críticos da rede.
- para a integração de mapas com relatórios detalhados.
- para a descrição do aplicativo em nível gerencial e operacional, o uso de ferramentas web e a integração aos demais sistemas da empresa sob a mesma plataforma.
- para a atualização do cadastro da rede elétrica, utilizando sistema construído a partir de tecnologias de geoprocessamento.
- para descrição do sistema de transmissão e geração, que é estruturado pela junção das plantas cartográficas, levantamentos aerofotogramétricos, dados das propriedades envolvidas e imagens digitalizadas dos processos.
- para precificação de postes, localização de consumidores inadimplentes e localização de agentes arrecadadores.

### 3.3 Análise Digital de Terrenos

O desenvolvimento e a aplicação de SIG's em análise de terrenos para as ciências ambientais, tema de aplicação desta tese, foram motivadas pela visão do mundo em camadas (MACKEY, 1996), onde os processos biofísicos estão posicionados hierarquicamente como apresentado na Figura 3.1.

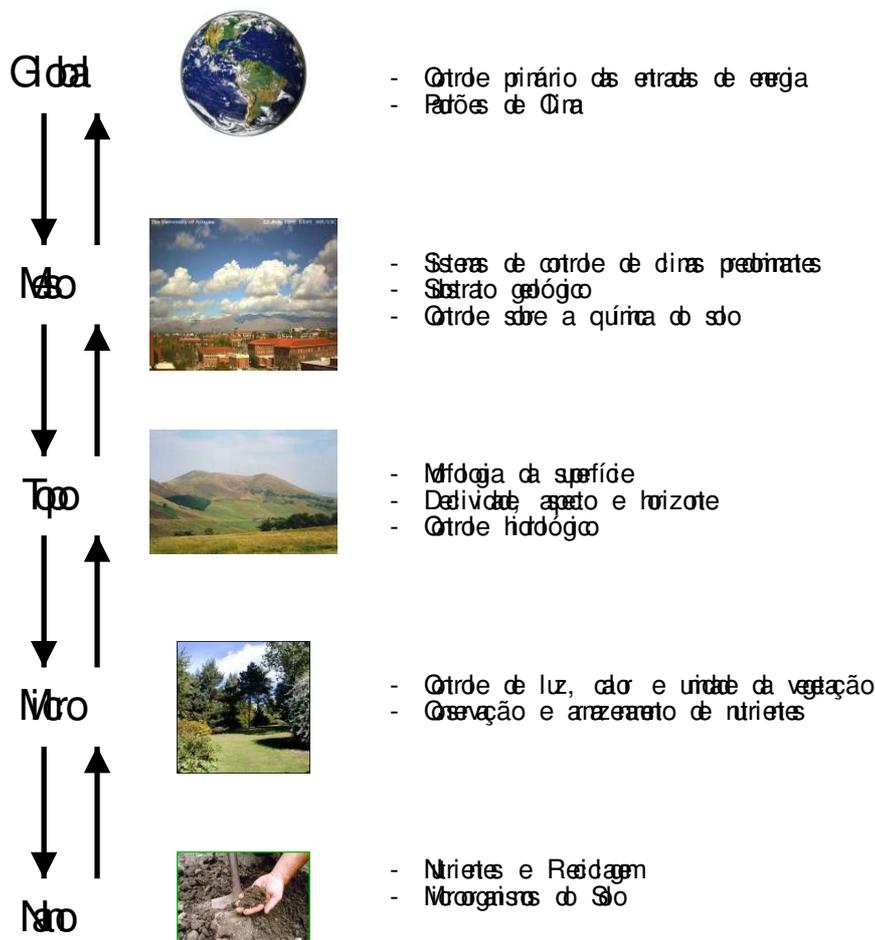


Figura 3.1: Processos biofísicos em camadas

Este tipo de divisão é bastante útil pois demonstra a complexidade dos processos individuais além de algumas dificuldades que são encontradas em delinear apropriadamente as escalas espaciais e temporais.

Muitos dos processos biofísicos mais importantes que ocorrem na superfície da terra (ou próximos) são influenciados pelo controle das interações e pelos níveis dos eventos passados e contemporâneos. Estes inter-relacionamentos são complexos e podem ser melhores entendidos usando a abordagem de modelagem de sistemas dinâmicos (KIRKBY et al., 1996).

O trabalho de (PHILLIPS, 1986) apresenta exemplos de sistemas de populações interagindo em uma comunidade ecológica; e (SCHAFFER, 1981) trata a geomorfologia fluvial; ambos demonstram que a operação de processos-chaves sobre diferentes escalas podem ser considerados independentemente uma em relação à outra.

Entretanto, não é aconselhável a utilização de modelos para uma dada escala originalmente desenvolvidos em outra escala.

A maioria das pesquisas hidrológicas, geomorfológicas e ecológicas são conduzidas na escalas global, nano e micro (MACKEY, 1996). As meso e topo-escalas recebem pouca atenção apesar de serem realmente importantes pois muitas das soluções para problemas ambientais, como erosão acelerada do solo e poluição, requerem o gerenciamento de estratégias nestas escalas (MOORE; HUTCHINSON, 1991).

A influência da morfologia da superfície sobre a canalização na hidrologia e o impacto da declividade e do aspecto sobre a exposição ao sol representam os controles mais importantes operando na topo-escala.

Além disso, diversos estudos mostram como a forma da superfície do terreno pode afetar a migração lateral e o acúmulo da água, sedimentos e outros constituintes (MACKEY, 1996).

A crescente popularidade do trabalho nestas duas escalas intermediárias tem influenciado a disponibilidade de dados de elevação digitais contínuos e de alta-resolução e o desenvolvimento de novas ferramentas computadorizadas de análise de terrenos (WILSON, 1996; BURROUGH; MCDONNELL, 1998; WILSON; BURROUGH, 1999).

O modelo *ICTM*, introduzido nesta tese, representa uma contribuição para a análise de dados geográficos nessas escalas intermediárias.

### 3.3.1 Fontes e Estruturas dos Dados Digitais de Elevação

A maioria dos conjuntos de dados digitais de elevação disponíveis atualmente são resultantes da captura de dados fotogramétricos (BOLSTAD; STOWE, 1994). Estas fontes dependem da interpretação estereoscópica de fotografias aéreas ou imagens de satélite usando *stereoplotters* manuais ou automáticos.

Outros conjuntos de dados podem ser gerados a partir da digitalização de linhas de contorno dos mapas topográficos e pela pesquisa de campo.

Estes dados de elevação estão usualmente organizados em uma das seguintes estruturas de dados (JAIN; MURTY; FLYNN, 1999) (ver Figura 3.2): (i) malhas regulares, (ii) redes irregulares triangulares e (iii) contornos – dependendo da fonte e/ou do método de análise utilizado.

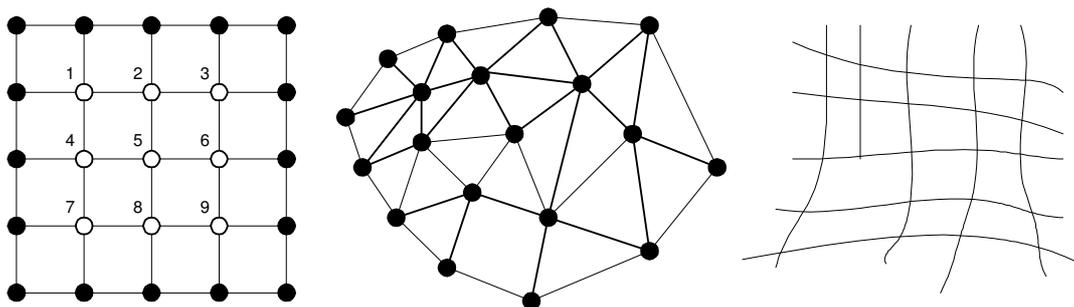


Figura 3.2: Métodos de estruturação dos dados de elevação

Modelos digitais de terrenos (MDT's) com malhas quadradas emergiram como a estrutura de dados mais utilizada por causa de sua simplicidade e da fácil implementação computacional. Por exemplo, matrizes de elevações que registram implicitamente as relações topológicas entre o dados do ponto com os dados dos pontos vizinhos.

O tamanho do grão da malha, que afeta diretamente os requisitos de armazenagem, eficiência e qualidade dos resultados e a dificuldade da malha quadrada representar fielmente mudanças abruptas na elevação são alguns obstáculos que foram ultrapassados nos últimos anos.

O aparecimento de várias novas técnicas de compressão reduziram os requisitos de armazenagem e melhoraram a eficiência computacional (SMITH; LEWIS, 1994). Além disso, encontram-se disponíveis modelos de elevação com malhas de 500m, 100m, 30m, 10m e 1m de resolução para diferentes partes do globo (HUTCHINSON et al., 1996).

Malhas irregulares são utilizadas amplamente e são baseadas em elementos triangulares com vértices nos pontos de amostra. Estes elementos consistem de planos juntando os três pontos adjacentes na rede e são usualmente construídos usando triangularização de Delaunay (PREPARATA; SHAMOS, 1985).

As redes triangulares incorporam descontinuidades facilmente e podem constituir estruturas de dados eficientes porque a densidade dos triângulos pode variar conforme a rugosidade do terreno.

A estrutura de contornos divide a área em pequenos polígonos de forma irregular baseados nas linhas de contorno e suas ortogonais. Esta estrutura é usada freqüentemente em aplicações hidrológicas porque podem reduzir as complexas equações de fluxo tridimensional em uma série de equações unidimensionais em áreas de relevo complexas (GRAYSON; BLOSCHL; MOORE, 1994).

### **3.3.2 Utilização e Estimação de Atributos Topográficos**

Muitos dos atributos topográficos mais importantes como declividade, canalização, aspecto e curvaturas, podem ser derivados a partir de três tipos de dados de elevação, em cada elemento da malha, em função dos atributos de sua vizinhança (MOORE; GRAYSON; LADSON, 1991).

As ferramentas de análise de terrenos têm suas classificações baseadas nas características dos atributos computados e/ou em sua extensão espacial. Alguns autores distinguem ferramentas que executam operações sobre a vizinhança local daquelas que operam sobre vizinhanças estendidas, para cálculos de drenagem, vertentes, etc (KORTE, 1994).

Usualmente, distinguem-se atributos primários (veja Tabela 3.1), que são computados diretamente dos modelos de elevação, de atributos secundários ou compostos, que envolvem combinações dos atributos primários e constituem índices derivados empiricamente ou fisicamente. Estes últimos podem caracterizar a variabilidade espacial de processos específicos ocorrendo no ambiente. Uma lista completa de atributos primários e seus derivados, atributos secundários, pode ser encontrado no trabalho de (MOORE; GRAYSON; LADSON, 1991).

A maioria destes atributos primários são calculados a partir de derivadas direcionais de uma superfície topográfica e são computados diretamente com um esquema de diferenças finitas de segunda ordem, ou encontrando uma função de interpolação bivariada ao modelo de elevação juntamente com as derivadas desta função (MOORE; LEWIS; GALLANT, 1993).

Os atributos secundários são computados a partir dois ou mais atributos primários e são importantes porque oferecem uma alternativa de descrever padrões em função do processo. Por exemplo, atributos que quantificam a função da topografia na redistribuição de água na superfície tem importante influência hidrológica, geomorfológica e ecológica. Atributos deste tipo representam as características do solo, distribuição e abundância da água no solo, suscetibilidade da superfície à erosão pela água, ou a dispersão da flora e da fauna em uma dada região.

Tabela 3.1: Exemplos de atributos topográficos primários

Atributo	Definição	Significância
Altitude	Elevação	<i>Clima, vegetação, energia potencial</i>
Altura do aclave	Altura média da área do aclave	<i>Energia potencial</i>
Aspecto	Azimute da inclinação	<i>Incidência solar, evaporação e transpiração, distribuição da flora e fauna</i>
Inclinação	Gradiente	<i>Velocidade de fluxo abaixo e acima da superfície, precipitação, vegetação, geomorfologia, conteúdo de água no solo</i>
Inclinação do Aclave	Média da inclinação do aclave	<i>Velocidade de dispersão</i>
Inclinação da Dispersão	Inclinação média da área de dispersão	<i>Taxa de drenagem do solo</i>
Comprimento do caminho de fluxo	Distância máxima do fluxo de água a um ponto da canalização	<i>Taxas de erosão, campos de sedimentos, tempo de concentração</i>
Curvatura de perfil	Inclinação da curvatura de perfil	<i>Aceleração de fluxo, taxa de erosão e deposição, geomorfologia</i>
Percentil de elevação	Proporção de células em um círculo mais baixo que a célula central	<i>Posição relativa, distribuição de fauna e flora</i>

### 3.3.3 Tratamento de Erros e Incerteza

Erros sistemáticos e não-sistemáticos nos modelos de elevação podem confundir os relacionamentos esperados entre os atributos do terreno e as condições do terreno controladas localmente. Estes problemas podem ser amplificados no momento que são calculadas as derivadas de primeira e segunda ordem (BOLSTAD; STOWE, 1994).

Os problemas mais sérios são encontrados usualmente quando os atributos secundários são derivados, por exemplo: os índices de capacidade de transporte de sedimentos e de umidade são muito sensíveis à presença de erros na fonte de dados em áreas planas e, também, à escolha do algoritmo de roteamento de fluxo.

Muitos estudos tem examinado as causas, detecção, visualização e correção dos erros nos modelos digitais de elevação e vários métodos foram propostos para a estimativa da magnitude destes erros (POLIDORI; CHOROWICZ; GUILLANDI, 1991; KRAUS, 1994).

Alguns pesquisadores sugerem que a precisão dos atributos topográficos primários e secundários não pode ser determinado pela comparação dos valores calculados com os de referência porque a superfície não é matematicamente suave.

É natural representar o terreno na forma de modelos de elevação digital. A Figura 3.3 mostra o MDT no centro das interações entre a fonte de captura de dados e as aplicações (MOORE; GRAYSON; LADSON, 1991; WILSON; GALLANT, 2000). Estas interações são suportadas pelos métodos de geração e por uma variedade crescente de técnicas para a interpretação e visualização de MDT's.

Usualmente, a determinação da resolução apropriada de um modelo digital de elevação interpolado ou filtrado é um compromisso entre alcançar fidelidade à superfície real e respeitar os limites práticos relacionados com a densidade e precisão das fontes de

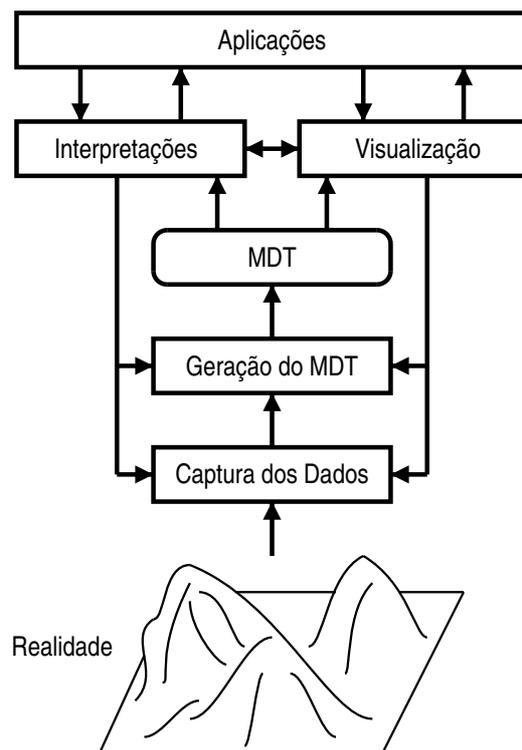


Figura 3.3: Tarefas associadas a modelagem digital de terrenos

dados. Esta característica influencia na varredura eficiente dos dados, pois os requisitos de armazenamento do modelo de elevação são bastante sensíveis à resolução.

Os métodos de adequação da resolução de MDT's à fonte de informação são apropriados quando os dados espaciais são obtidos uniformemente, como por exemplo, contornos de elevação a partir de mapas topográficos em escala fixa, ou a partir de dados de elevação em malhas de sensoriamento remoto.

## 4 O MODELO *Topo-ICTM*

Chama-se de *ICTM* (*Interval Categorizer Tessellation-Based Model*) um modelo geral baseado em tesselações para ferramentas de categorização que são capazes de sub-dividir um espaço de natureza geométrica em sub-regiões apresentando características similares.

Neste capítulo, apresenta-se o *Topo-ICTM* (AGUIAR; COSTA, 2001), que é um *ICTM* topográfico, bidimensional, de uma camada (*2d-1c-ICTM*), definido para permitir a análise da variação do sinal de declividade da função que mapeia a topografia de uma dada área geográfica, subdividindo esta área em regiões que apresentam o mesmo comportamento com respeito à declividade do relevo.

Cada uma dessas regiões é dita pertencer a uma dada categoria de declividade de acordo com o sinal (positivo, negativo, nulo) da declividade da função relevo<sup>1</sup>. Uma aplicação imediata é na Geofísica, onde um sub-divisão apropriada de áreas geográficas em segmentos apresentando características similares é freqüentemente conveniente (COBLENTZ et al., 2003, 2000). Outra aplicação é na análise do estado de erosão de terrenos (COSTA, 2003).

Esta versão do *ICTM* surgiu diretamente a partir do estudo do relatório técnico de (COBLENTZ et al., 2000), recentemente publicado em (COBLENTZ et al., 2003), que apresentou um método baseado em uma análise unidimensional para subdividir áreas geofísicas em sub-regiões de monotonicidade, considerando apenas uma direção.

Diferentemente, o modelo baseado em tesselações apresentado aqui faz uma análise bidimensional da declividade, usando regras locais para a criação e categorização das sub-regiões, dada a situação relativa de cada sub-região com respeito à área total, a partir dos estados assumidos pelas células da tesselação.

Além disso, a análise pode ser facilmente refinada, pela repetição do procedimento para focar uma sub-região de uma certa categoria de declividade, ou mudando os parâmetros de entrada (número de células da tesselação etc), ou considerando um raio de vizinhança maior, por exemplo.

### 4.1 A Formalização do Modelo

Esta seção introduz o modelo categorizador intervalar baseado em tesselações para a categorização de terrenos em função da declividade topográfica, chamado de *Topo-ICTM* (ou *2d-1c-ICTM*), formalizado em termos de operações sobre matrizes. A Figura 4.1 representa esquematicamente as etapas de funcionamento do modelo.

---

<sup>1</sup>Neste trabalho, utiliza-se o termo **declividade** para significar o **sinal da função de declividade**.

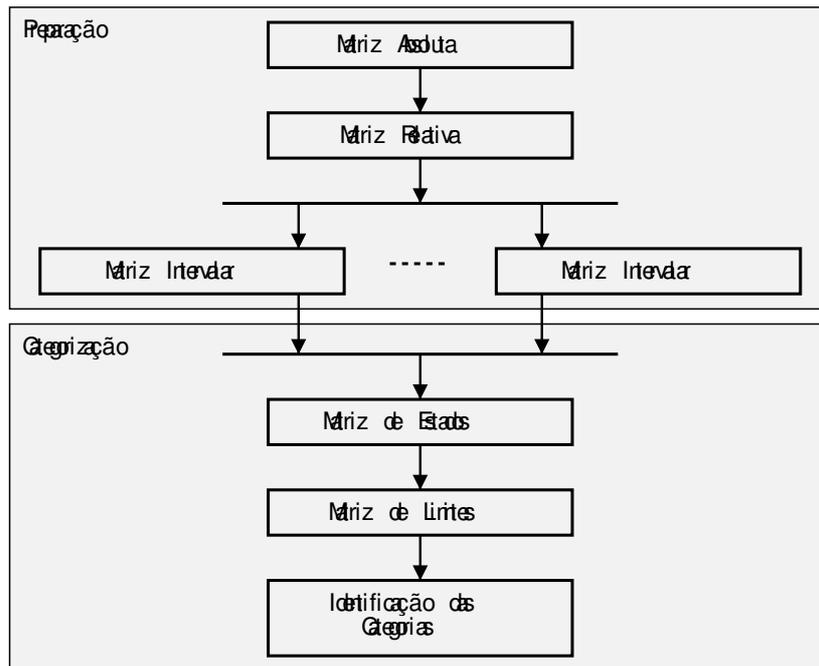


Figura 4.1: Etapas do Modelo *Topo-ICTM*

#### 4.1.1 Matriz Absoluta e Matriz Relativa

Os dados de entrada do modelo são extraídos a partir de um modelo digital de elevação da região geográfica a ser analisada, onde as alturas são dadas nos pontos referenciados pelas coordenadas de latitude e longitude. Esta região geográfica é então representada por uma tesselação regular que é determinada pela subdivisão da área total em subáreas retangulares suficientemente pequenas, cada uma representada por uma célula da tesselação. Esta subdivisão é feita de acordo com o tamanho da célula estabelecida pelo analista e está diretamente associado ao grau de refinamento da tesselação.

**Definição 4.1.** Uma tesselação é uma matriz  $M$  com  $n_r$  linhas e  $n_c$  colunas. A entrada na  $x$ -ésima linha e na  $y$ -ésima coluna é chamada de célula  $xy$  da tesselação  $M$ .

Ao relacionar as células de uma tesselação com os pontos do modelo digital de elevação da área, cada célula corresponderá a vários pontos do modelo. Então, toma-se, como valor de altura de cada célula, o valor médio das alturas nos pontos correspondentes no modelo de elevação. Esses valores médios são as entradas da matriz da tesselação  $M$ :

**Definição 4.2.** A matriz absoluta de uma tesselação  $M$  é a matriz  $n_r \times n_c$   $M^{abs} = [m_{xy}^{abs}]$ , onde a entrada  $m_{xy}^{abs}$  é o valor absoluto da altura média dos pontos representados pela célula  $xy$  da tesselação  $M$ .

A fim de simplificar os dados da matriz  $M^{abs}$ , os valores absolutos de suas entradas são normalizados, dividindo-os pelo maior de seus valores  $m_{max}$ .

**Definição 4.3.** A matriz relativa  $M^{rel}$  é definida como a matriz  $n_r \times n_c$  dada por

$$M^{rel} = \frac{M^{abs}}{m_{max}}.$$

### 4.1.2 Matrizes Intervalares

As alturas são medidas exatas em um modelo digital de elevação, então os únicos erros nos valores  $m_{xy}$  vêm da discretização da área em termos do conjunto discreto das células da tesselação.

É necessário conhecer os valores da função que mapeia o relevo  $h_{\xi v}$  no modelo digital de elevação, para todo  $\xi$  e  $v$ , mas somente os valores  $h_{xy} \equiv m_{xy}^{rel} = \frac{m_{xy}^{abs}}{m_{max}}$  para  $(1, 1), \dots, (1, n_r), \dots, (n_c, 1), \dots, (n_c, n_r)$ , determinados pela divisão da área em  $n_r n_c$  células, são utilizados efetivamente nos cálculos.

Seguindo a abordagem (COBLENTZ et al., 2000) baseada na Matemática Intervalar, utilizam-se intervalos para controlar os erros associados aos valores das células. Para cada  $\xi v$ , que é diferente de  $xy$ , é razoável estimar  $h_{\xi v}$  como o valor  $m_{xy}^{rel}$  no ponto  $xy$  que está mais próximo a  $\xi v$ , significando que  $\xi v$  pertence à mesma célula que  $xy$ .

Para um  $y$  fixo, quando  $\xi > x$ , o ponto  $xy$  é ainda o mais próximo até que seja alcançado o ponto médio  $x_{mid}y = \frac{(x+(x+1))}{2}y$  entre  $xy$  e  $(x+1)y$ .

O maior erro de aproximação possível  $|m_{xy}^{rel} - h_{\xi y}|$  para tais pontos é obtido quando a distância entre  $xy$  e  $\xi y$  é a maior, ou seja, quando  $\xi y = x_{mid}y$ . Neste caso, o erro de aproximação é igual a  $|h_{x_{mid}y} - m_{xy}^{rel}|$ .

As proposições a seguir generalizam as proposições estabelecidas em (COBLENTZ et al., 2003).

**Lema 4.1.** *Para um  $y$  fixo, se  $\xi > x$ , então o erro de aproximação  $\epsilon$  é limitado por  $0.50 \times |m_{(x+1)y}^{rel} - m_{xy}^{rel}|$ .*

*Prova.* Se os pontos  $xy$  e  $(x+1)y$  pertencem ao mesmo segmento de área, então a dependência de  $n_{\xi y}$  sobre  $\xi y$  seria razoavelmente suave para  $\xi \in [x, (x+1)]$ . Portanto, em um intervalo pequeno  $[x, (x+1)]$ , pode-se, com razoável exatidão, ignorar os termos quadráticos e de mais alta ordem na expansão de  $h_{(\xi+\Delta\xi)y}$  e assim, aproximar  $h_{\xi y}$  por uma função linear. Para uma função linear  $\xi \mapsto h_{\xi y}$ , a diferença  $h_{x_{mid}y} - m_{xy}^{rel}$  é igual à metade da diferença  $m_{(x+1)y}^{rel} - m_{xy}^{rel}$ . Por outro lado, se os pontos  $xy$  e  $(x+1)y$  pertencem a células diferentes, então a dependência  $h_{\xi y}$  deveria apresentar alguma não suavidade, e é razoável esperar que a diferença  $|m_{(x+1)y}^{rel} - m_{xy}^{rel}|$  é muito maior que o erro de aproximação. Em ambos casos, o erro de aproximação  $\epsilon$  é limitado por  $0.50 \times |m_{(x+1)y}^{rel} - m_{xy}^{rel}|$ .  $\square$

**Lema 4.2.** *Para um  $y$  fixo, se  $\xi < x$ , então o erro de aproximação  $\epsilon$  é limitado por  $0.50 \times |m_{xy}^{rel} - m_{(x-1)y}^{rel}|$ .*

**Proposição 4.1.** *Para o erro de aproximação  $\epsilon_x$ ,*

$$\epsilon_x \leq \Delta_x = 0.5 \times \min \{ |m_{xy}^{rel} - m_{(x-1)y}^{rel}|, |m_{(x+1)y}^{rel} - m_{xy}^{rel}| \}.$$

*Prova.* Segue dos Lemas 4.1 e 4.2.  $\square$

Como um resultado, considerado um dado  $y$ , juntamente com os valores centrais  $m_{xy}^{rel}$ , para cada  $x$ , têm-se os intervalos  $m_{xy}^{x[1]}$  contendo todos os valores possíveis de  $h_{\xi y}$ , para  $x - \frac{1}{2} \leq \xi \leq x + \frac{1}{2}$ .

**Corolário 4.1.** *Considerando um  $y$  fixo, para cada  $x$ , se  $x - \frac{1}{2} \leq \xi \leq x + \frac{1}{2}$ , então  $h_{\xi y} \in m_{xy}^{x[1]} = [m_{xy}^{x-}, m_{xy}^{x+}]$ , onde  $m_{xy}^{x-} = m_{xy}^{rel} - \Delta_x$  e  $m_{xy}^{x+} = m_{xy}^{rel} + \Delta_x$ .*

Usando uma argumentação análoga, considerando um  $x$  fixo, é possível concluir que:

**Proposição 4.2.** Para o erro de aproximação  $\epsilon_y$ ,

$$\epsilon_y \leq \Delta_y = 0.5 \times \min \left\{ |m_{xy}^{rel} - m_{x(y-1)}^{rel}|, |m_{x(y+1)}^{rel} - m_{xy}^{rel}| \right\}.$$

**Corolário 4.2.** Considerando um  $x$  fixo, para cada  $y$ , se  $y - \frac{1}{2} \leq v \leq y + \frac{1}{2}$ ,  $h_{xv} \in m_{xy}^{y[1]} = [m_{xy}^{y-}, m_{xy}^{y+}]$ , onde  $m_{xy}^{y-} = m_{xy}^{rel} - \Delta_y$  e  $m_{xy}^{y+} = m_{xy}^{rel} + \Delta_y$ .

As matrizes intervalares podem então ser definidas do seguinte modo:

**Definição 4.4.** Se  $m_{xy}^{x\pm} = m_{xy}^{rel} \pm \Delta_i$  e  $m_{xy}^{y\pm} = m_{xy}^{rel} \pm \Delta_j$ , então as matrizes intervalares  $M^{x[1]}$  e  $M^{y[1]}$ , associadas à matriz relativa  $M^{rel}$ , são definidas pelas matrizes  $n_r \times n_c$  intervalares

$$M^{x[1]} = [m_{xy}^{x[1]}] = \left[ [m_{xy}^{x-}, m_{xy}^{x+}] \right], M^{y[1]} = [m_{xy}^{y[1]}] = \left[ [m_{xy}^{y-}, m_{xy}^{y+}] \right].$$

### 4.1.3 Matriz de Estados e os Registradores de Declividade

Faz-se agora a categorização das células conforme a sua declividade, de acordo com o trabalho de (COBLENTZ et al., 2000, 2003). Assume-se inicialmente que as funções de aproximação do relevo introduzidas pelo modelo baseado em tesselações são funções lineares por partes. Modela-se todo processo como um tipo de problema de satisfação de restrições, onde o modelo baseado em tesselação está a cargo de encontrar uma função de aproximação linear, também por partes, do relevo (e o conjunto de pontos limites entre as regiões resultantes) que satisfaz as restrições impostas pela matriz intervalar.

Para simplificar o espaço de soluções, toma-se uma abordagem qualitativa das funções de aproximação do relevo, agrupando-as em classes de equivalência de acordo com o sinal de sua declividade (positiva, negativa, nula). Assim o modelo constrói uma solução qualitativa para o problema de satisfação de restrições, isto é, a classe das funções de aproximação compatíveis com as restrições da matriz intervalar. Proceda-se da seguinte forma:

**Proposição 4.3.** Sejam  $M^{x[1]}$  e  $M^{y[1]}$  as matrizes intervalares. Para um dado  $xy$ , se:

- (i)  $m_{xy}^{x+} \geq m_{(x+1)y}^{x-}$ , então existe uma função de aproximação do relevo não crescente entre  $xy$  e  $(x+1)y$  (direção oeste-leste).
- (ii)  $m_{(x-1)y}^{x-} \leq m_{xy}^{x+}$ , então existe uma função de aproximação do relevo não decrescente entre  $(x-1)y$  e  $xy$  (direção oeste-leste).
- (iii)  $m_{xy}^{y+} \geq m_{x(y+1)}^{y-}$ , então existe uma função de aproximação do relevo não crescente entre  $xy$  e  $x(y+1)$  (direção norte-sul).
- (iv)  $m_{x(y-1)}^{y-} \leq m_{xy}^{y+}$ , então existe uma função de aproximação do relevo não decrescente entre  $x(y-1)$  e  $xy$  (direção norte-sul).

*Prova.* Apresenta-se um esquema da prova. Em (i), toma-se, por exemplo,  $\mu_{xy} = m_{xy}^{x+}$ ,  $\mu_{(x+1)y} = m_{(x+1)y}^{x-}$  e se usa uma interpolação linear para definir os valores  $\mu_{ky}$  para  $x < k < x+1$ . As provas de (ii)-(iv) são similares.  $\square$

Para cada célula, são definidos quatro registradores direcionados de declividade –  $reg.n$  (norte),  $reg.s$  (sul),  $reg.e$  (leste) e  $reg.w$  (oeste), indicando os sinais de declividade admissíveis da função que aproxima a função do relevo em uma destas direções, levando em consideração os valores das suas células vizinhas. A análise da declividade é feita de acordo com a Proposição 4.3.

**Definição 4.5.** Um registrador de declividade da célula  $xy$  é uma tupla  $reg = (reg.n, reg.s, reg.e, reg.w)$ , cujos valores dos registradores direcionados de declividade são dados por:

(i) Para células que não estão na borda da tesselação:

$$reg.n = \begin{cases} 0 & \text{se a condição do item (iv) da Proposição 4.3 é verdadeira;} \\ 1 & \text{caso contrário.} \end{cases}$$

$$reg.s = \begin{cases} 0 & \text{se a condição do item (iii) da Proposição 4.3 é verdadeira;} \\ 1 & \text{caso contrário.} \end{cases}$$

$$reg.e = \begin{cases} 0 & \text{se a condição do item (i) da Proposição 4.3 é verdadeira;} \\ 1 & \text{caso contrário.} \end{cases}$$

$$reg.w = \begin{cases} 0 & \text{se a condição do item (ii) da Proposição 4.3 é verdadeira;} \\ 1 & \text{caso contrário.} \end{cases}$$

(ii) Para as células das bordas norte, sul, leste e oeste:  $reg.n = 0$ ,  $reg.s = 0$  e  $reg.e = 0$  e  $reg.w = 0$ , respectivamente<sup>2</sup>. Os demais registradores direcionados de declividade são determinados usando (i).

Observa-se que, para cada célula, existe uma, e somente uma, possibilidade de atribuição de valores aos registradores de declividade.

**Definição 4.6.** A matriz dos registradores de declividade é definida como uma matriz  $n_r \times n_c$  denotada por  $M^{reg} = [m_{xy}^{reg}]$ , onde a entrada na  $x$ -ésima linha e na  $y$ -ésima coluna é o valor do registrador de declividade da célula correspondente.

**Corolário 4.3.** Considerando a direção oeste-leste, uma função de aproximação do relevo  $m_{xy}$  é:

(i) estritamente crescente entre  $xy$  e  $(x+1)y$  se  $m_{xy}^{reg.e} = 1$  (neste caso,  $m_{(x+1)y}^{reg.w} = 0$ );

(ii) estritamente decrescente entre  $xy$  e  $(x+1)y$  se  $m_{(x+1)y}^{reg.w} = 1$  (neste caso,  $m_{xy}^{reg.e} = 0$ );

(iii) constante entre  $xy$  e  $(x+1)y$  se  $m_{xy}^{reg.e} = 0$  e  $m_{(x+1)y}^{reg.w} = 0$ .

Resultados similares do Corolário 4.3 valem também para a direção norte-sul.

**Definição 4.7.** Sejam  $w_{reg.e} = 1$ ,  $w_{reg.s} = 2$ ,  $w_{reg.w} = 4$  e  $w_{reg.n} = 8$  os pesos a serem associados aos registradores de declividade. A matriz de estados é definida como uma matriz  $n_r \times n_c$  dada por  $M^{state} = [m_{xy}^{state}]$ , cuja entrada na  $x$ -ésima linha e na  $y$ -ésima coluna é o valor correspondente ao do estado da célula, calculado como o valor da codificação binária dos registradores de declividade correspondentes, da seguinte forma:

$$m_{xy}^{state} = w_{reg.e} \times m_{xy}^{reg.e} + w_{reg.s} \times m_{xy}^{reg.s} + w_{reg.w} \times m_{xy}^{reg.w} + w_{reg.n} \times m_{xy}^{reg.n}$$

Assim, para um dado  $xy$ , a célula correspondente pode assumir um, e apenas um, dos estados apresentados na Figura 4.2, representado por um valor  $m_{xy}^{state} = 0..15$ .

<sup>2</sup>Isto é consistente com a função relevo sendo constante nas células da borda.

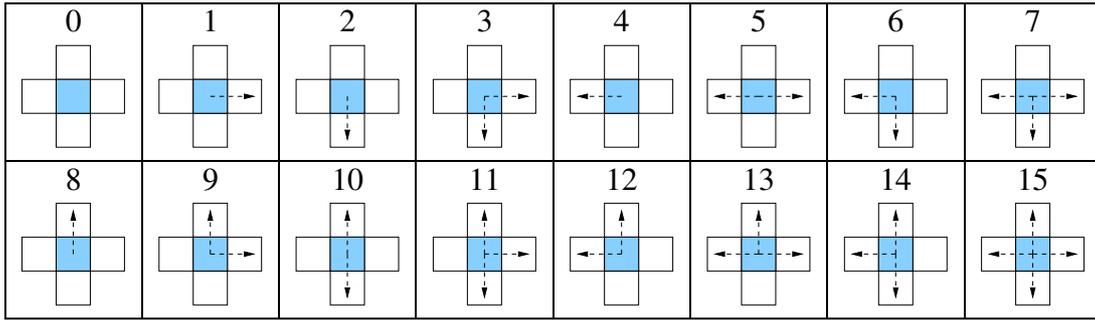


Figura 4.2: Esquema de todos os valores possíveis dos estados da célula

Tabela 4.1: Condições das células  $xy$  não limítrofes

Id	Condições
1	$m_{(x-1)y}^{reg.e} = m_{xy}^{reg.e} = 1$
2	$m_{xy}^{reg.w} = m_{(x+1)y}^{reg.w} = 1$
3	$m_{(x-1)y}^{reg.e} = m_{xy}^{reg.e} = m_{xy}^{reg.w} = m_{(x+1)y}^{reg.w} = 0$
4	$m_{x(y-1)}^{reg.s} = m_{xy}^{reg.s} = 1$
5	$m_{xy}^{reg.n} = m_{x(y+1)}^{reg.n} = 1$
6	$m_{x(y-1)}^{reg.s} = m_{xy}^{reg.s} = m_{xy}^{reg.n} = m_{x(y+1)}^{reg.n} = 0$

#### 4.1.4 Matriz de Limites e as Sub-regiões de Declividade Constante

Uma célula é definida como limítrofe quando a função do relevo muda sua declividade, apresentando pontos críticos (máximo, mínimo ou pontos de inflexão). Para se identificar tais células limítrofes, utiliza-se um registrador de limite associado àquela célula. As células da borda de toda região são consideradas células limítrofes.

**Definição 4.8.** A matriz de limites é definida como a matriz  $n_r \times n_c$  denotada por  $M^{limit} = [m_{xy}^{limit}]$ , onde a entrada na  $x$ -ésima linha e na  $y$ -ésima coluna é determinada como

$$m_{xy}^{limit} = \begin{cases} 0, & \text{se uma das condições listadas na Tabela 4.1 é verificada;} \\ 1, & \text{caso contrário.} \end{cases}$$

Analisando a matriz de limites é fácil detectar a existência de configurações de relevo conhecidas (por exemplo, os esquemas mostrados na Figura 4.3). A presença de células limítrofes permitem a subdivisão da área total em categorias de declividade (regiões de mesma declividade).

**Definição 4.9.** A sub-região de declividade constante associada à célula não limítrofe  $xy$ , denotada por  $\mathcal{SR}_{xy}$ , é definida indutivamente como segue:

- (i)  $xy \in \mathcal{SR}_{xy}$ ;
- (ii) Se  $x'y' \in \mathcal{SR}_{xy}$ , então todas suas células vizinhas não limítrofes também pertencem a  $\mathcal{SR}_{xy}$ .

Observe que  $\mathcal{SR}_{xy} = \mathcal{SR}_{x'y'}$  se, e somente se,  $x'y' \in \mathcal{SR}_{xy}$  (respectivamente,  $xy \in \mathcal{SR}_{x'y'}$ ). Como resultado final, toda a área torna-se dividida em categorias (regiões) de declividade bem definidas.

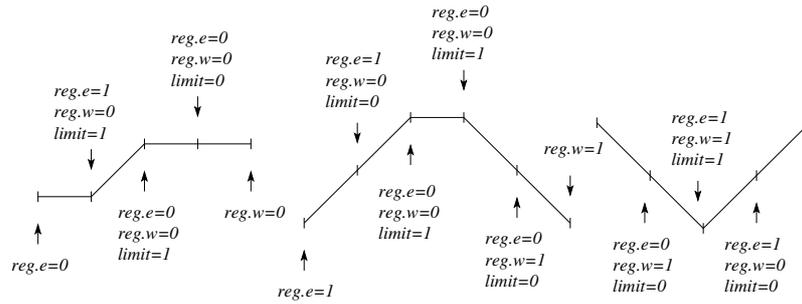


Figura 4.3: Esquemas das células limítrofes

A Definição 4.9 induz a um algoritmo recursivo similar aos comumente utilizados em Computação Gráfica para preenchimento de polígonos (FOLEY et al., 1990). Considere-se, por exemplo, a matriz de limites produzida pelo modelo no processo de categorização de uma dada região *R*, mostrada na Tabela 4.2. A categorização de declividade correspondente é apresentada na Figura 4.4<sup>3</sup>.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0	1
2	1	0	0	0	0	0	0	1	1	0	0	1	1	1	0	1
3	1	0	0	0	0	0	0	1	1	0	0	1	1	1	0	1
4	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0	1
5	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0	1
6	1	0	0	0	0	1	1	0	0	0	0	1	1	1	0	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Tabela 4.2: A matriz de limites  $M_{xy}^{limit}$  associada a *R*

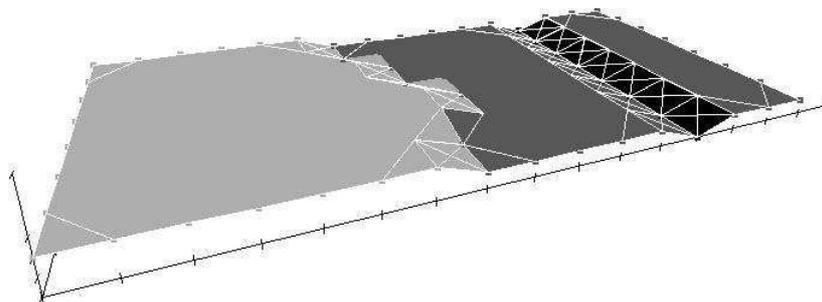


Figura 4.4: A categorização da declividade de uma dada região *R*

## 4.2 Considerações sobre o Modelo *Topo-ICTM*

O *Topo-ICTM* analisa apenas uma característica da área geográfica (por isso, possui apenas uma camada), neste caso, a declividade da função que mapeia o relevo da área

<sup>3</sup>Os resultados numéricos e gráficos foram produzidos pelo sistema *Topo-ICTM* (ver capítulo 7)

considerada, fazendo uma análise bidimensional, conforme a latitude e a longitude.

O tamanho  $n_r \times n_c$  da tesselação pode ser arbitrária ou escolhida de acordo com um critério específico estabelecido pela aplicação. De qualquer forma, a categorização obtida pode ser refinada definindo-se outro tamanho para a tesselação ou tomando cada região resultante como uma nova área a ser analisada separadamente.

A análise pode ser feita até ser alcançado um número conveniente de sub-divisões, caracterizando, assim, o dinamismo do modelo. A formalização usando matrizes de registradores permite que a informação armazenada nestes seja facilmente acessada a qualquer momento do processamento, devido a indexação dos elementos nas matrizes.

A implementação do modelo pode ser naturalmente paralelizada, pois a análise é feita localmente pelas regras. Como os dados de entrada são numéricos e, portanto, suscetíveis a erros, aplica-se a aritmética intervalar.

### 4.3 A Ferramenta *Topo-ICTM*

A geração de uma superfície topográfica é o fundamento para qualquer trabalho subsequente em Geoprocessamento.

Como princípio metodológico deve-se obter uma superfície topográfica na escala adequada aos propósitos visados.

A ferramenta *Topo-ICTM*, que implementa o modelo (ver capítulo 7) possibilita uma definição e visualização precisa da geometria de superfícies topográficas, adequadas aos objetivos de futuras aplicações em escalas variadas.

Esta ferramenta está sendo aplicada para a categorização da superfície topográfica da porção leste do Estado Rio Grande do Sul, abrangendo uma área que compreende imagens do Satélite LANDSAT TM7 ETM+, órbita-ponto 221-080/221-081/221-082 (veja o mosaico na Figura 4.5) e o detalhamento compreendendo uma região de 60x120 Km da serra gaúcha, a partir de dados do Modelo de Elevação Digital gerados pelo Satélite ASTER. O mapa de localização das cenas do satélite ASTER pode ser visualizado na Figura 4.6.

Encontram-se nesta região quatro unidades geomorfológicas distintas: a Planície Costeira, Escudo Sul-riograndense, Depressão Central e o Planalto. Dependendo da região em estudo, da sua declividade e variação altimétrica será necessário partir de dados com resolução espacial mais adequada à região. Esta base de dados topográficos em escalas distintas e a aplicação em áreas geomorfológicamente distintas permitirá analisar os resultados com maior veracidade.

Os dados gerados pelo sistema *Topo-ICTM* poderão ser comparados com os dados de uma base de dados geográfica da superfície topográfica, produzidas a partir das cartas topográficas na escala 1:250.000, permitindo gerar Modelos de Elevação Digital com resoluções espacial de 125m.

Estas informações geradas poderão ser comparadas e os resultados do *Topo-ICTM* extrapolados em escalas maiores a partir dos dados das cartas topográficas na escala 1:50.000, proporcionando, assim, gerar o modelo de elevação digital com resolução espacial de 25m, e dos dados da Imagem ASTER com uma resolução espacial de 10m.

As imagens acima referidas encontram-se com referências geográficas, corrigidas geometricamente e orto-retificadas de acordo com a base cartográfica em escala coerente com a resolução espacial da Imagem LANDSAT TM7.

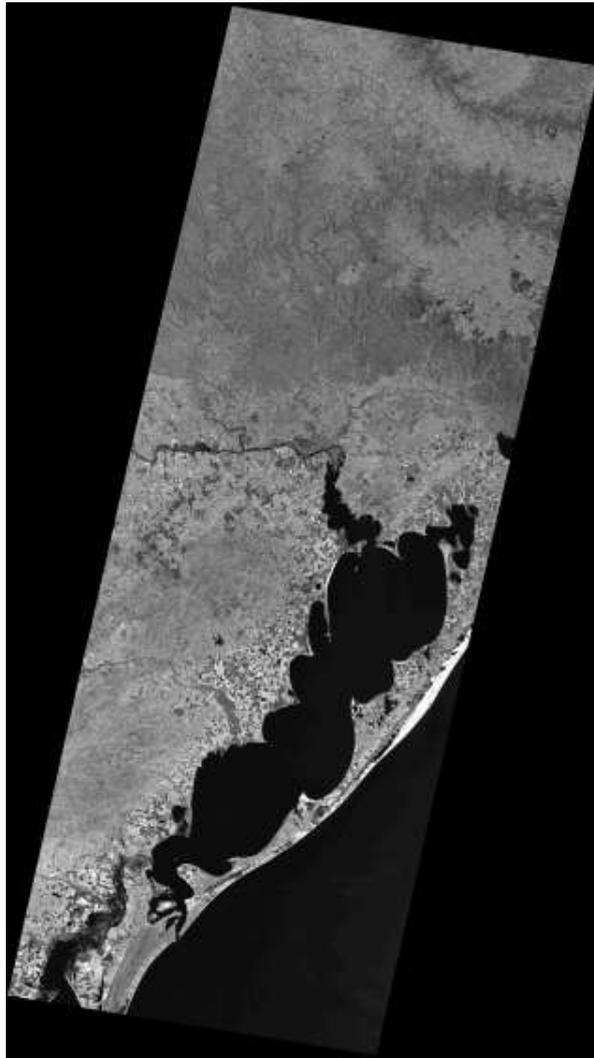


Figura 4.5: Mosaico das três órbitas-ponto do satélite LANDSAT TM7

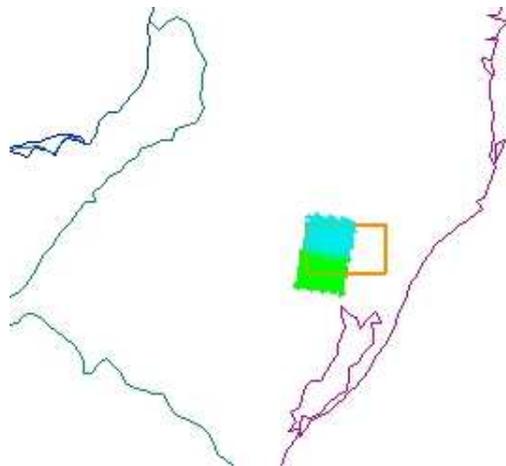


Figura 4.6: Mapa de localização das cenas do satélite ASTER

## 5 O MODELO *ICTM*

Neste capítulo, apresenta-se a definição do *ICTM* em sua forma mais geral, chamado *nd-lc-ICTM*, que é um modelo geral, baseado em tesselações, capaz de produzir uma categorização confiável de regiões de um espaço de características geométricas com um número  $n$  de dimensões (por exemplo, como apresentado em (AGUIAR et al., 2003)), podendo analisar múltiplas características conhecidas em suficientemente muitos pontos, e suportando descoberta de conhecimento relacional entre as sub-regiões identificadas.

A categorização do espaço determinada por cada característica é efetuada em uma camada do modelo, gerando diferentes sub-divisões do espaço analisado. Por exemplo, uma área geográfica pode ser analisada conforme sua topografia, vegetação, demografia, dados econômicos etc.

Uma categorização global pode ser alcançada, a partir da categorização de cada camada, mediante um procedimento de projeção. Esta categorização global determinará uma sub-divisão mais confiável e com maior significância, combinando as análises efetuadas para cada característica. Este tipo de projeção permite análises sobre a dependência mútua destas características.

Nas seções seguintes, apresenta-se a formalização do modelo *ICTM* geral. Inicialmente, considera-se a generalização do número de camadas, conservando-se o modelo bidimensional. Depois, generaliza-se o número de dimensões.

### 5.1 Modelo *2d-lc-ICTM*

Cada característica do espaço está representada em uma camada do modelo *ICTM*. Assim, devido à independência das análises, as sub-divisões em cada camada também ocorrem de forma independente.

Na Figura 5.1, cada camada bidimensional é representada por um rótulo alfanumérico.

O primeiro passo em direção à generalização do modelo é a definição do conceito de tesselações com multi-camadas.

**Definição 5.1.** *Seja uma tesselação  $M$ . Uma camada  $l$  desta tesselação é a matriz absoluta  $n_r \times n_c$   $l$ - $M^{abs} = [l-m_{xy}^{abs}]$ , onde a entrada  $l$ - $m_{xy}^{abs}$  é o valor absoluto da altura média dos pontos representados pela célula  $xy$  na camada  $l$  da tesselação  $M$ .*

**Definição 5.2.** *Sejam  $M$  uma tesselação  $n_c \times n_r$  e  $l \in \mathbb{N}$ , uma tesselação multi-camada  $\mathcal{L}$ - $\mathcal{M}$  é a estrutura*

$$\mathcal{L}\text{-}\mathcal{M} = (1\text{-}M^{abs}, \dots, l\text{-}M^{abs})$$

*cujas entradas na  $l$ -ésima camada,  $x$ -ésima linha e na  $y$ -ésima coluna é denotada por  $l$ - $m_{xy}$ .*

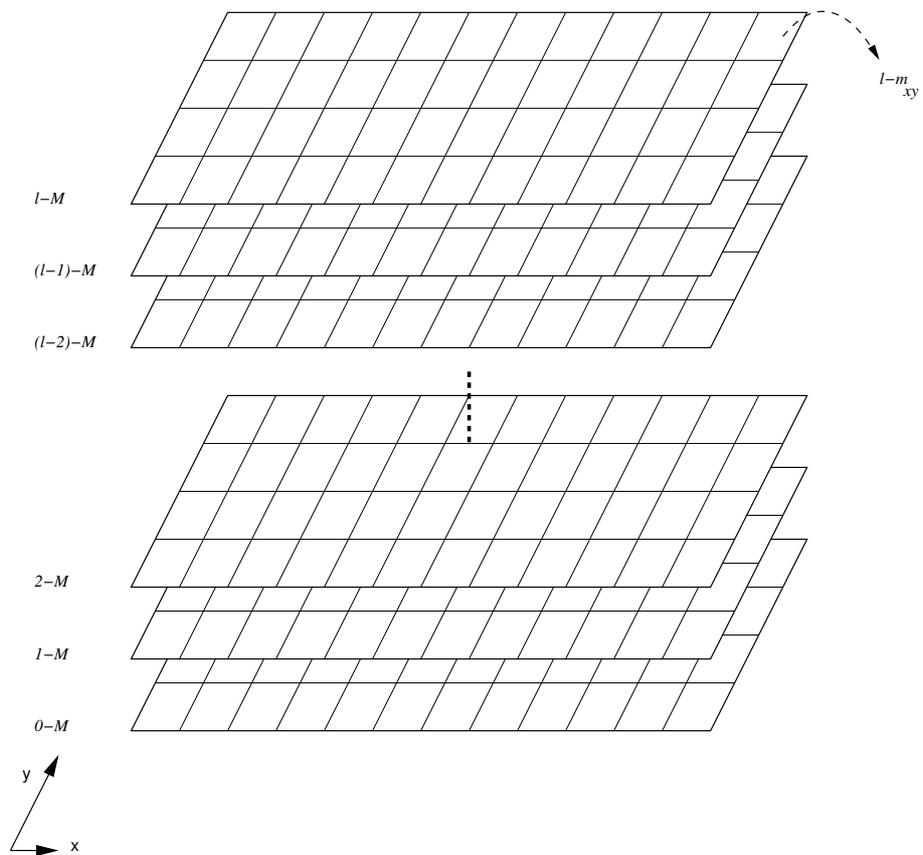


Figura 5.1: Visão do Modelo  $2d-lc-ICTM$

**Definição 5.3 (Tesselação Multi-camada Rotulada).** *Sejam  $M$  uma tesselação  $n_c \times n_r$  e  $l \in \Sigma^*$ , uma tesselação multi-camada  $\mathcal{L}\text{-}\mathcal{M}$  é a estrutura*

$$\mathcal{L}\text{-}\mathcal{M} = (\text{camada\_um-}M^{abs}, \dots, \text{camada\_l-}M^{abs})$$

*cuja entrada na camada chamada camada  $l$ ,  $x$ -ésima linha e na  $y$ -ésima coluna é denotada por  $\text{camada\_l-}m_{xy}$ .*

O rótulo permite uma identificação da característica que está sendo analisada. Além disso, a utilização de rótulos pode descaracterizar a idéia de camadas (uma camada sobre a outra) e passar a apresentar cada camada como um quadro de visualização distinto do outro.

A camada de número 0 de uma tesselação  $M$ , denotada por  $0\text{-}M$  ou rotulada como  $\pi\text{-}M$ , é considerada a camada base da projeção das demais camadas<sup>1</sup>.

Assim, todas as células da tesselação ganham um novo indexador, que é a referência à camada desejada.

As definições 4.3, 4.4, 4.6, 4.7, 4.8 e a Tabela 4.1, apresentadas na seção 4.1, podem ser então reescritas com o identificador da camada a qual pertencem e são apresentadas a seguir.

**Definição 5.4.** *A matriz relativa da camada  $l$  denotada por  $l\text{-}M^{rel}$  é definida como a matriz  $n_r \times n_c$  dada por*

$$l\text{-}M^{rel} = \frac{l\text{-}M^{abs}}{l\text{-}m_{max}}$$

**Definição 5.5.** *Se  $l\text{-}m_{xy}^{x\pm} = l\text{-}m_{xy}^{rel} \pm \Delta_i$  e  $l\text{-}m_{xy}^{y\pm} = l\text{-}m_{xy}^{rel} \pm \Delta_j$ , então as matrizes intervalares da camada  $l$ ,  $l\text{-}M^{x[1]}$  e  $l\text{-}M^{y[1]}$ , associadas à matriz relativa  $l\text{-}M^{rel}$ , são definidas pelas matrizes  $n_r \times n_c$  intervalares*

$$l\text{-}M^{x[1]} = [l\text{-}m_{xy}^{x[1]}] = \left[ [l\text{-}m_{xy}^{x-}, l\text{-}m_{xy}^{x+}] \right], l\text{-}M^{y[1]} = [l\text{-}m_{xy}^{y[1]}] = \left[ [l\text{-}m_{xy}^{y-}, l\text{-}m_{xy}^{y+}] \right].$$

**Definição 5.6.** *A matriz dos registradores de declividade da camada  $l$  é definida como uma matriz  $n_r \times n_c$  denotada por  $l\text{-}M^{reg} = [l\text{-}m_{xy}^{reg}]$ , onde a entrada na  $x$ -ésima linha e na  $y$ -ésima coluna da camada  $l$  é o valor do registrador de declividade da célula correspondente.*

**Definição 5.7.** *A matriz de estados da camada  $l$  é definida como uma matriz  $n_r \times n_c$  dada por  $l\text{-}M^{state} = [l\text{-}m_{xy}^{state}]$ , cuja entrada na  $x$ -ésima linha e na  $y$ -ésima coluna da camada  $l$  é o valor correspondente ao do estado da célula, calculado como o valor da codificação binária dos registradores de declividade correspondentes, da seguinte forma*

$$l\text{-}m_{xy}^{state} = w_{reg.e} \times l\text{-}m_{xy}^{reg.e} + w_{reg.s} \times l\text{-}m_{xy}^{reg.s} + w_{reg.w} \times l\text{-}m_{xy}^{reg.w} + w_{reg.n} \times l\text{-}m_{xy}^{reg.n}$$

**Definição 5.8.** *A matriz de limites da camada  $l$  é definida como a matriz  $n_r \times n_c$  denotada por  $l\text{-}M^{limit} = [l\text{-}m_{xy}^{limit}]$ , onde a entrada na  $x$ -ésima linha e na  $y$ -ésima coluna da camada  $l$  é determinada como*

$$l\text{-}m_{xy}^{limit} = \begin{cases} 0, & \text{se uma das condições listadas na Tabela 5.1 é verificada;} \\ 1, & \text{caso contrário.} \end{cases}$$

Na próxima seção, apresenta-se a generalização do modelo *ICTM* para a multi-dimensionalidade da tesselação.

<sup>1</sup>A proposta dos algoritmos para a projeção das camadas será apresentada na seção 5.3.

Tabela 5.1: Condições das células  $xy$  não limítrofes da camada  $l$ 

Id	Condições
1	$l-m_{(x-1)y}^{reg.e} = l-m_{xy}^{reg.e} = 1$
2	$l-m_{xy}^{reg.w} = l-m_{(x+1)y}^{reg.w} = 1$
3	$l-m_{(x-1)y}^{reg.e} = l-m_{xy}^{reg.e} = l-m_{xy}^{reg.w} = l-m_{(x+1)y}^{reg.w} = 0$
4	$l-m_{x(y-1)}^{reg.s} = l-m_{xy}^{reg.s} = 1$
5	$l-m_{xy}^{reg.n} = l-m_{x(y+1)}^{reg.n} = 1$
6	$l-m_{x(y-1)}^{reg.s} = l-m_{xy}^{reg.s} = l-m_{xy}^{reg.n} = l-m_{x(y+1)}^{reg.n} = 0$

## 5.2 Modelo $nd$ - $lc$ - $ICTM$

O modelo  $2d$ - $lc$ - $ICTM$ , mesmo manipulando apenas tesselações bidimensionais, tem sua aplicação em uma gama considerável de situações práticas (ver capítulo 7). Nesta seção, faz-se a generalização do número de dimensões da tesselação. Assim, possibilita-se a análise de espaços que necessitem mais de duas coordenadas para serem definidos ou referenciados.

A Figura 5.2 apresenta as várias camadas de um espaço de três dimensões.

**Definição 5.9.** Uma tesselação de  $n$ -dimensional é uma matriz  $M$  de  $n$  dimensões, indicadas por  $a_1 \times a_2 \times a_3 \times \dots \times a_n$ .

**Definição 5.10.** Sejam  $M$  uma tesselação  $n$ -dimensional e  $l \in \mathbb{N}$ , uma tesselação multi-camada  $\mathcal{L}$ - $\mathcal{M}$  é a estrutura

$$\mathcal{L}\text{-}\mathcal{M} = (1\text{-}M^{abs}, \dots, l\text{-}M^{abs})$$

cujas entradas na  $l$ -ésima camada, é denotada por  $l\text{-}m_{a_1 \dots a_n}$ .

Uma célula é referida pela sua coordenada no sistema multi-dimensional. Portanto, a célula  $l\text{-}m_{2,3,4,5}$  está posicionada: a duas unidades da origem no eixo (dimensão)  $a_1$ , a três unidades da origem no eixo  $a_2$ , a quatro unidades da origem no eixo  $a_3$  e a cinco unidades da origem no eixo  $a_4$ .

**Definição 5.11 (Tesselação Rotulada).** Sejam  $M$  uma tesselação  $n$ -dimensional e  $l \in \Sigma^*$ , uma tesselação multi-camada rotulada  $\mathcal{L}$ - $\mathcal{M}$  é a estrutura

$$\mathcal{L}\text{-}\mathcal{M} = (\text{camada\_um}\text{-}M^{abs}, \dots, \text{camada\_l}\text{-}M^{abs})$$

cujas entradas na camada chamada  $\text{camada\_l}$  é denotada por  $\text{camada\_l}\text{-}m_{a_1 \dots a_n}$ .

**Definição 5.12.** A matriz absoluta da camada  $l$  de uma tesselação  $M$   $n$ -dimensional é a matriz  $a_1 \times \dots \times a_n$  denotada por  $l\text{-}M^{abs} = [l\text{-}m_{a_1 \dots a_n}^{abs}]$ , onde a entrada  $l\text{-}m_{a_1 \dots a_n}^{abs}$  é o valor absoluto da altura média dos pontos representados pela célula  $a_1 \dots a_n$  na camada  $l$  da tesselação  $M$ .

**Definição 5.13.** A matriz relativa da camada  $l$  denotada por  $l\text{-}M^{rel}$  é definida como a matriz  $a_1 \times \dots \times a_n$  dada por

$$l\text{-}M^{rel} = \frac{l\text{-}M^{abs}}{l\text{-}m_{max}}$$

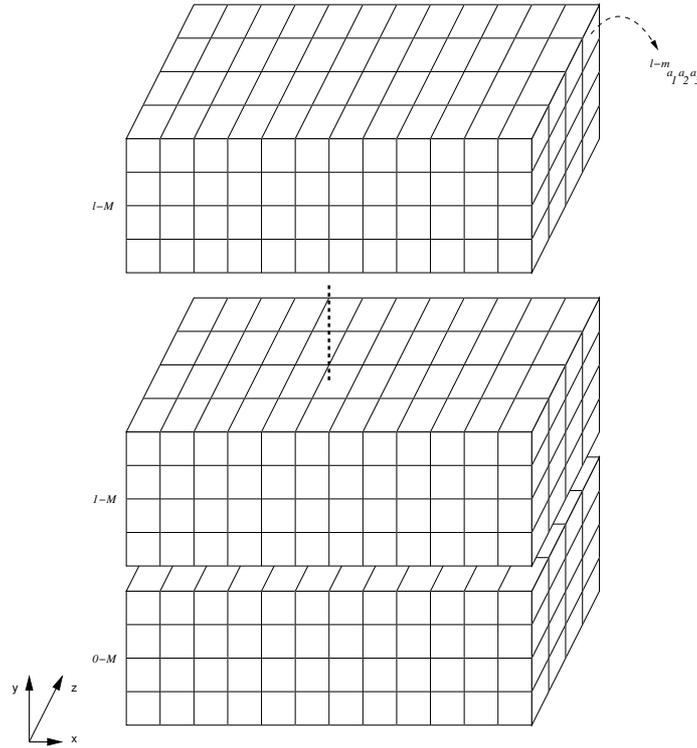


Figura 5.2: Visão do  $nd$ -lc-ICTM, neste caso ( $n = 3$ )

As Proposições 4.1 e 4.2 são reescritas na seguinte Proposição:

**Proposição 5.1.** *O erro de aproximação  $\epsilon_{a_i}$ , em uma tesselação multi-dimensional*

$$\epsilon_{a_i} \leq \Delta_{a_i} = 0.5 \times \min \left\{ \left| l-m_{a_1 \dots a_i \dots a_n}^{rel} - l-m_{a_1 \dots a_{i-1} \dots a_n}^{rel} \right|, \left| l-m_{a_1 \dots a_{i+1} \dots a_n}^{rel} - l-m_{a_1 \dots a_i \dots a_n}^{rel} \right| \right\},$$

para  $i=1 \dots n$ .

**Definição 5.14.** *Se  $l-m_{a_1 \dots a_n}^{a_i^\pm} = l-m_{a_1 \dots a_n}^{rel} \pm \Delta_{a_i}$ , então as matrizes intervalares da camada  $l$ ,  $l-M_{a_1}^{a_1^{[1]}}$ ,  $\dots$ ,  $l-M_{a_n}^{a_n^{[1]}}$ , associadas à matriz relativa  $l-M^{rel}$ , são definidas pelas matrizes  $a_1 \times \dots \times a_n$  intervalares*

$$l-M_{a_i}^{a_i^{[1]}} = \left[ l-m_{a_1 \dots a_n}^{a_i^{[1]}} \right] = \left[ \left[ l-m_{a_1 \dots a_n}^{a_i^-}, l-m_{a_1 \dots a_n}^{a_i^+} \right] \right], \text{ para } i = 1 \dots n.$$

No caso multi-dimensional existe um  $\epsilon$  (erro de aproximação) para cada eixo do sistema de coordenadas, logo tem-se uma matriz intervalar para cada eixo. Ou seja, se a camada tem  $n$ -dimensões, então existirão  $n$  matrizes intervalares indicando os erros de aproximação para cada camada do modelo.

A seguir, apresenta-se a re-leitura da Proposição 4.3.

**Proposição 5.2.** *Sejam  $M_{a_i}^{a_i^{[1]}}$  (para  $i = 1 \dots n$ ) as matrizes intervalares. Para um dado  $a_1 \dots a_n$ , se:*

- (i)  $l-m_{a_1 \dots a_i \dots a_n}^{a_i^+} \geq l-m_{a_1 \dots a_{i+1} \dots a_n}^{a_i^-}$ , então existe uma função de aproximação do relevo não crescente entre  $a_1 \dots a_i \dots a_n$  e  $a_1 \dots a_{i+1} \dots a_n$  (direção  $a_i$ - $a_{i+1}$ ).

- (ii)  $l-m_{a_1 \dots a_{i-1} \dots a_n}^{a_i^-} \leq l-m_{a_1 \dots a_i \dots a_n}^{a_i^+}$ , então existe uma função de aproximação do relevo não decrescente entre  $a_1 \dots a_{i-1} \dots a_n$  e  $a_1 \dots a_i \dots a_n$  (direção  $a_{i-1}-a_i$ ).

Os registradores de declividade sofrem uma pequena modificação em sua Definição 4.5.

**Definição 5.15.** Um registrador de declividade da célula  $a_1 \dots a_n$  é uma tupla  $reg = (reg.a_1^-, reg.a_1^+, reg.a_2^-, reg.a_2^+, \dots, reg.a_n^-, reg.a_n^+)$ , cujos valores dos registradores direcionados de declividade são dados por:

- (i) Para células que não estão na borda da tesselação:

$$reg.a_i^- = \begin{cases} 0 & \text{se a condição do item (i) da Proposição 5.2 é verdadeira;} \\ 1 & \text{caso contrário.} \end{cases}$$

$$reg.a_i^+ = \begin{cases} 0 & \text{se a condição do item (ii) da Proposição 5.2 é verdadeira;} \\ 1 & \text{caso contrário.} \end{cases}$$

- (ii) Para as células cuja(s) face(s)  $a_i^\pm$  está(ão) na borda, então  $reg.a_i^\pm = 0^2$ . Os demais registradores direcionados de declividade são determinados usando (i).

Observa-se que os índices  $a_i^\pm$  são sensíveis ao contexto onde são empregados, ou seja, quando eles aparecem na forma  $reg.a_i^\pm$  (Definição 5.15), eles indicam o registrador de declividade no eixo da coordenada  $a_i$  na direção que cresce (+) ou decresce (-) o seu valor. A Figura 5.3 apresenta a distribuição dos  $a_i$  para o caso tridimensional.

Quando os índices se apresentarem na forma  $l-m_{a_1 \dots a_n}^{a_i^\pm}$ , estes indicam os limites inferior ( $a_i^-$ ) e superior ( $a_i^+$ ) do intervalo do erro de aproximação da célula  $a_1 \dots a_n$ , na coordenada  $a_i$ .

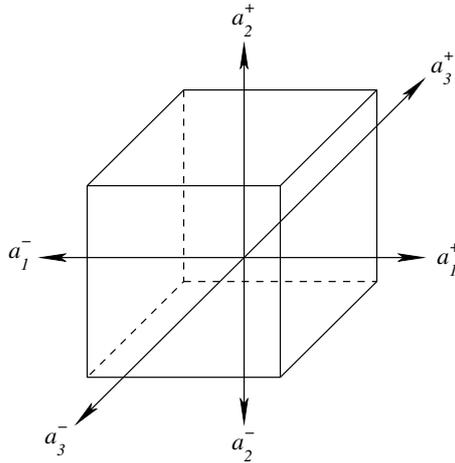


Figura 5.3: Notação do sentido dos registradores  $a_i^\pm$

**Definição 5.16.** A matriz dos registradores de declividade da camada  $l$  é definida como uma matriz  $n$ -dimensional  $a_1 \times \dots \times a_n$  denotada por  $l-M^{reg} = [l-m_{a_1 \dots a_n}^{reg}]$ .

No caso multi-dimensional, o corolário 4.3 é reformulado como segue:

<sup>2</sup>Isto é consistente com a função relevo sendo constante nas células da borda.

Tabela 5.2: Condições das células  $a_1 \dots a_n$  não limítrofes da camada  $l$ , para  $i = 1 \dots n$ 

Id	Condições
$(3 \times i) - 2$	$l-m_{a_1 \dots a_{i-1} \dots a_n}^{a_i^+} = l-m_{a_1 \dots a_i \dots a_n}^{a_i^+} = 1$
$(3 \times i) - 1$	$l-m_{a_1 \dots a_i \dots a_n}^{a_i^-} = l-m_{a_1 \dots a_{i+1} \dots a_n}^{a_i^-} = 1$
$3 \times i$	$l-m_{a_1 \dots a_{i-1} \dots a_n}^{a_i^+} = l-m_{a_1 \dots a_i \dots a_n}^{a_i^+} = l-m_{a_1 \dots a_i \dots a_n}^{a_i^-} = l-m_{a_1 \dots a_{i+1} \dots a_n}^{a_i^-} = 0$

**Corolário 5.1.** Considerando a direção  $a_i - a_{i+1}$ , uma função de aproximação do relevo  $l-m_{a_1 \dots a_n}$  é:

- (i) estritamente crescente entre  $a_1 \dots a_i \dots a_n$  e  $a_1 \dots a_{i+1} \dots a_n$  se  $l-m_{a_1 \dots a_i \dots a_n}^{reg.a_i^+} = 1$  (neste caso,  $l-m_{a_1 \dots a_{i+1} \dots a_n}^{reg.a_i^-} = 0$ );
- (ii) estritamente decrescente entre  $a_1 \dots a_i \dots a_n$  e  $a_1 \dots a_{i+1} \dots a_n$  se  $l-m_{a_1 \dots a_{i+1} \dots a_n}^{reg.a_i^-} = 1$  (neste caso,  $l-m_{a_1 \dots a_i \dots a_n}^{reg.a_i^+} = 0$ );
- (iii) constante entre  $a_1 \dots a_i \dots a_n$  e  $a_1 \dots a_{i+1} \dots a_n$  se  $l-m_{a_1 \dots a_i \dots a_n}^{reg.a_i^+} = 0$  e  $l-m_{a_1 \dots a_{i+1} \dots a_n}^{reg.a_i^-} = 0$ .

O peso de cada registrador deve ser determinado de modo a não configurarem valores conflitantes para o registrador de estado da célula. Ou seja, cada estado da célula deve ser alcançado somente a partir de uma combinação de pesos dos registradores de declividade. A Definição 5.17 caracteriza algoritmicamente a determinação dos pesos dos registradores de declividade.

**Definição 5.17.** Sejam os pesos  $w_{reg.a_i^-} = 2^{(2 \times i) - 2}$  e  $w_{reg.a_i^+} = 2^{(2 \times i) - 1}$  (para  $i = 1 \dots n$ ) que serão associados aos registradores de declividade direcionados. A matriz  $n$ -dimensional de estados da camada  $l$  é definida como uma matriz  $a_1 \times \dots \times a_n$  dada por  $l-M^{state} = [l-m_{a_1 \dots a_n}^{state}]$ , cuja entrada na coordenada  $a_1 \dots a_n$  da camada  $l$  é o valor correspondente ao do estado da célula, calculado como o valor da codificação binária dos registradores de declividade correspondentes, da seguinte forma

$$l-m_{a_1 \dots a_n}^{state} = \sum_{i=1}^n w_{reg.a_i^-} \times l-m_{a_1 \dots a_n}^{reg.a_i^-} + w_{reg.a_i^+} \times l-m_{a_1 \dots a_n}^{reg.a_i^+}.$$

**Definição 5.18.** A matriz de limites da camada  $l$  é definida como a matriz  $a_1 \times \dots \times a_n$  denotada por  $l-M^{limit} = [l-m_{a_1 \dots a_n}^{limit}]$ , onde a entrada  $a_1 \dots a_n$  da camada  $l$  é determinada como

$$l-m_{a_1 \dots a_n}^{limit} = \begin{cases} 0, & \text{se uma das condições listadas na Tabela 5.2 é verificada;} \\ 1, & \text{caso contrário.} \end{cases}$$

A definição de sub-região 4.9 deve ser reformulada da seguinte maneira:

**Definição 5.19.** A sub-região de declividade constante associada à célula não limítrofe  $a_1 \dots a_n$ , denotada por  $l-\mathcal{SR}_{a_1 \dots a_n}$ , é definida indutivamente como segue:

- (i)  $a_1 \dots a_n \in l-\mathcal{SR}_{a_1 \dots a_n}$ ;
- (ii) Se  $a'_1 \dots a'_n \in l-\mathcal{SR}_{a_1 \dots a_n}$ , então todas suas células vizinhas não limítrofes também pertencem a  $l-\mathcal{SR}_{a_1 \dots a_n}$ .

### 5.3 Algoritmos para a Projeção das Camadas

A camada  $\pi-M_{a_1 \dots a_n}^{limit}$  ou  $0-M_{a_1 \dots a_n}^{limit}$  do modelo *nd-lc-ICTM* foi reservada para receber a projeção das células limites de todas as camadas ou de algumas camadas selecionadas. Esta projeção tem o intuito de identificar informações para o analista, como:

- as células que são limites em todas as camadas;
- a projeção de todas as sub-regiões;
- o grau de certeza de uma determinada célula ser limite, etc.

Para esse tipo de análise inicial (pois é a mais imediata) pôde-se prever dois tipos de projeções.

No primeiro tipo de projeção (tipo I), basta a célula ser limítrofe em apenas uma das camadas e ela será projetada como limítrofe na camada  $\pi-M_{a_1 \dots a_n}^{limit}$ . Assim, o método trivial de projeção pode (vide Figura 5.4) obter, na camada base, todas as regiões geradas individualmente em cada camada.

**Definição 5.20.** Cada camada  $l$  de uma tesselação  $n$ -dimensional  $M$  tem um peso associado  $0 \leq w_i \leq 1$ , para  $i = 1, \dots, l$ .

Para este tipo de projeção,  $w_i = 1$  ( $w_i = 0$ ) indica que a camada  $i$ - $M$  (não) foi selecionada para a projeção das células limites.

**Definição 5.21.** Seja uma tesselação  $M$  com  $l$ -camadas e  $n$ -dimensional. A projeção  $\pi-M_{a_1 \dots a_n}^{limit}$  das camadas selecionadas sobre a camada base, indicando todas as células limites, é dada por

$$\pi-M_{a_1 \dots a_n}^{limit} = 0-M_{a_1 \dots a_n}^{limit} = \bigvee_{i=1}^l i-M_{a_1 \dots a_n}^{limit} \times w_i, \forall a_1 \dots a_n \in M.$$

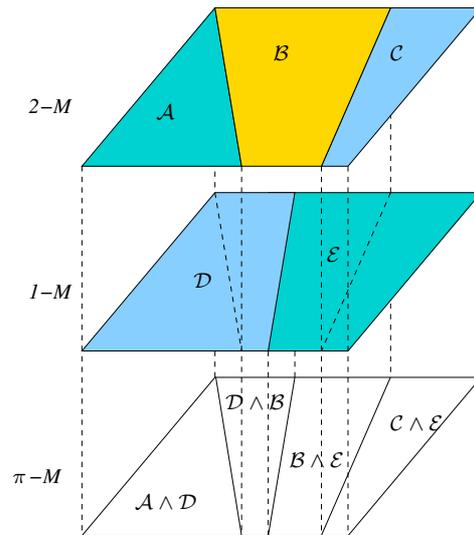


Figura 5.4: Exemplo de projeção das camadas (tipo I)

No segundo tipo de projeção (tipo II), cada camada pode apresentar participações diferentes na determinação da projeção das regiões limites. Assim, cada camada tem seu grau de participação na projeção final, normalmente definido pelo especialista.

Ainda, as camadas podem possuir pesos que não sejam complementares<sup>3</sup>, ou seja, a sua soma pode não ser 1, ou 100%. Portanto, os pesos das camadas devem ser normalizados.

**Definição 5.22.** *Sejam  $0 \leq w_i \leq 1$ , os pesos associados às camadas da tesselação  $M$ . A normalização destes pesos, denotada por  $\bar{w}_i$  é dada por:*

$$\bar{w}_i = \frac{w_i}{\sum_{j=1}^l w_j}, \text{ para } i = 1, \dots, l.$$

**Definição 5.23.** *Seja uma tesselação  $M$  com  $l$ -camadas e  $n$ -dimensional e os pesos normalizados  $\bar{w}_i$  associados a cada camada. A projeção  $\pi-M_{a_1 \dots a_n}^{limit}$  das  $l$  camadas sobre a camada base, indicando todas as células limites, é dada por*

$$\pi-M_{a_1 \dots a_n}^{limit} = 0-M_{a_1 \dots a_n}^{limit} = \sum_{i=1}^l i-M_{a_1 \dots a_n}^{limit} \times \bar{w}_i, \forall a_1 \dots a_n \in M.$$

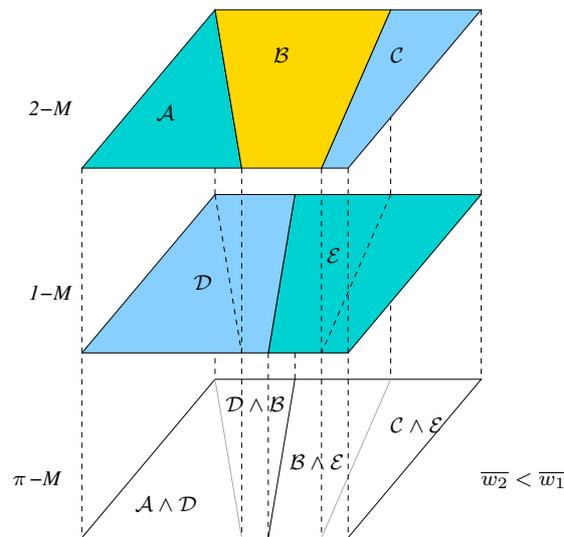


Figura 5.5: Exemplo de projeção das camadas (tipo II)

Neste segundo tipo de projeção (veja Figura 5.5), as células marcadas como limite na camada  $\pi-M$  também indicarão o grau de certeza desta marcação. Os valores do registrador de limite  $\pi-M_{a_1 \dots a_n}^{limit}$  assumirão valores entre 0 e 1. A camada com peso  $\bar{w}_i = 0$  indica que esta camada não é selecionada para a projeção, como ocorre na projeção tipo I.

Fazendo uma analogia com a representação gráfica de uma célula limite, se  $\pi-M_{a_1 \dots a_n}^{limit} = 1$  então ela poderia ser representada pela cor preta. Se  $\pi-M_{a_1 \dots a_n}^{limit} = 0$ , ela seria branca e os demais valores seriam representados como tons de cinza.

<sup>3</sup>Quando o analista tem a clara percepção da importância de cada camada (em relação as demais camadas) os pesos se tornam complementares naturalmente. Caso não seja possível esta correlação, a não complementaridade dos pesos pode ser uma ferramenta bastante útil.

## 6 EXTRAÇÃO DE FATOS

Neste capítulo, apresenta-se a etapa de extração dos fatos a partir das camadas do modelo *ICTM*. Pretende-se extrair os fatos presentes nas características das regiões do espaço identificadas pelo modelo.

Nas próximas definições, a fim de simplificar a notação, adota-se  $l\text{-}\mathcal{SR}_i$  como a representação de uma dada região  $i$  na camada  $l$  da tesselação  $M$  em substituição ao  $l\text{-}\mathcal{SR}_{a_1\dots a_n}$  da Definição 5.19. Assim, faz-se a referência ao conjunto de pontos que formam a região  $\mathcal{SR}_i$  e não aos pontos individualmente.

Apresentam-se duas abordagens para a extração de fatos. Uma abordagem qualitativa, em que os fatos são representados por sentenças lógicas, uma abordagem quantitativa, em que o conhecimento é representado por expressões de probabilidades condicionais, relativas às características das regiões.

### 6.1 Extração de Fatos Representados por Sentenças Lógicas

A estratégia desta abordagem consiste na detecção de pontos comuns em diferentes regiões de diferentes camadas. Pode-se também detectar se uma dada região de uma camada está dentro de outra região, numa camada distinta, ou regiões que não possuem pontos em comum.

Para este tipo de abordagem, a noção de conjunto parece bastante interessante. Para a detecção de áreas comuns, propõe-se utilizar a técnica do envelope, bastante difundida em Computação Gráfica (FOLEY et al., 1990).

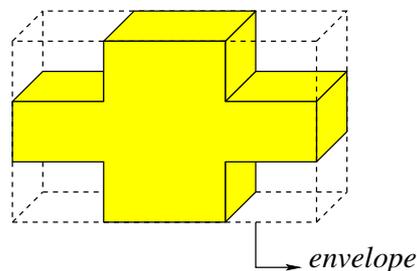


Figura 6.1: Envelope de um polígono tridimensional

A Figura 6.1 apresenta um exemplo de envelope de um polígono em três-dimensões.

Este tipo de técnica deve ser utilizada no estágio inicial da extração de fatos, visto que pode apresentar erro (conforme Figura 6.2) na determinação das intersecções, mais especificamente falsos-positivo. No entanto, consegue-se excluir, com certeza, aqueles

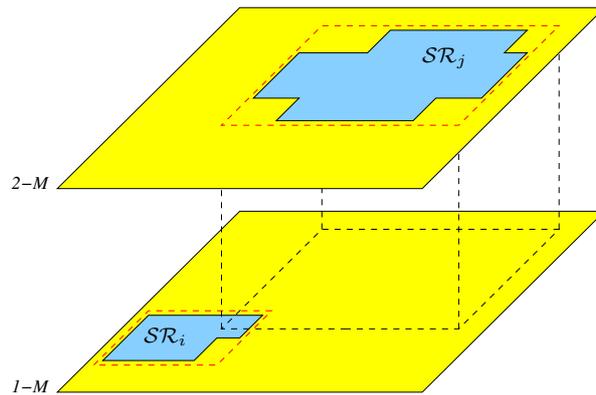


Figura 6.2: Erro encontrado na técnica do envelope

casos onde não ocorrem intersecções.

Por exemplo, se  $SR_i \subseteq SR_j$  (Figura 6.3), todos os vértices do envelope de  $SR_i$  estão dentro do envelope de  $SR_j$ .

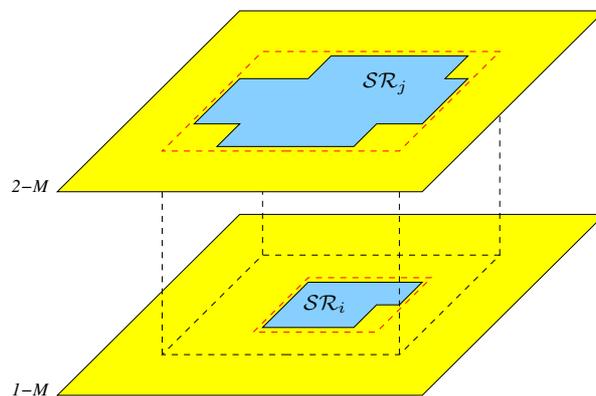


Figura 6.3: Região  $SR_i$  inscrita na região  $SR_j$

Porém, se  $SR_i \cap SR_j = \emptyset$  (Figura 6.4), então nenhum dos vértices do envelope de  $SR_i$  está dentro do envelope de  $SR_j$ .

Além disso, se  $SR_i \cap SR_j \neq \emptyset$  (Figura 6.5), então pelo menos um dos vértices do envelope de  $SR_i$  está dentro do envelope de  $SR_j$ .

A partir do estabelecimento das relações entre as regiões pode-se extrair sentenças lógicas, como as apresentadas nas seguintes definições.

**Definição 6.1.** *Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional ( $l > 2$ ),  $l_1$ - $M$  e  $l_2$ - $M$  duas camadas distintas da tesselação  $M$ . Se o conjunto de pontos formadores de uma determinada região  $l_1$ - $SR_i$  está contido no conjunto  $l_2$ - $SR_j$  ( $l_1$ - $SR_i \subseteq l_2$ - $SR_j$ , veja Figura 6.6), então  $l_1$ - $SR_i \Rightarrow l_2$ - $SR_j$ .*

**Definição 6.2.** *Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional ( $l > 2$ ),  $l_1$ - $M$  e  $l_2$ - $M$  duas camadas distintas da tesselação  $M$ . Se nenhum ponto do conjunto de pontos formadores de uma dada região  $l_1$ - $SR_i$  está contido em outro conjunto  $l_2$ - $SR_j$  ( $l_1$ - $SR_i \cap l_2$ - $SR_j = \emptyset$ , veja Figura 6.7), então  $\neg(l_1$ - $SR_i \wedge l_2$ - $SR_j)$ .*

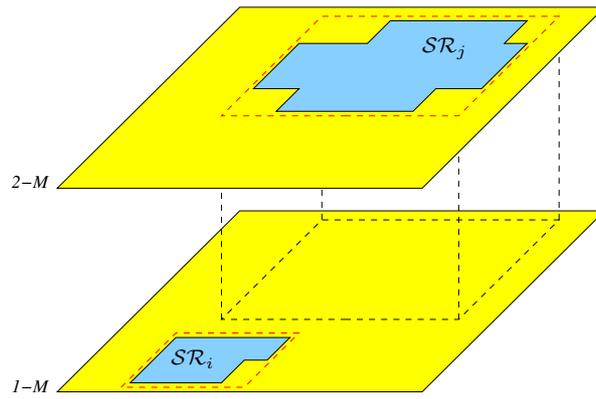


Figura 6.4: Sub-regiões sem intersecção

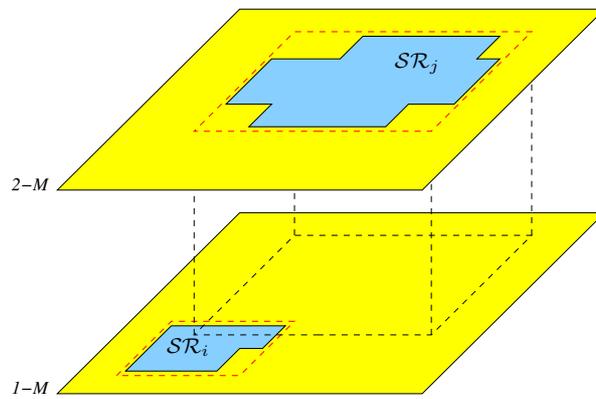


Figura 6.5: Intersecção de sub-regiões

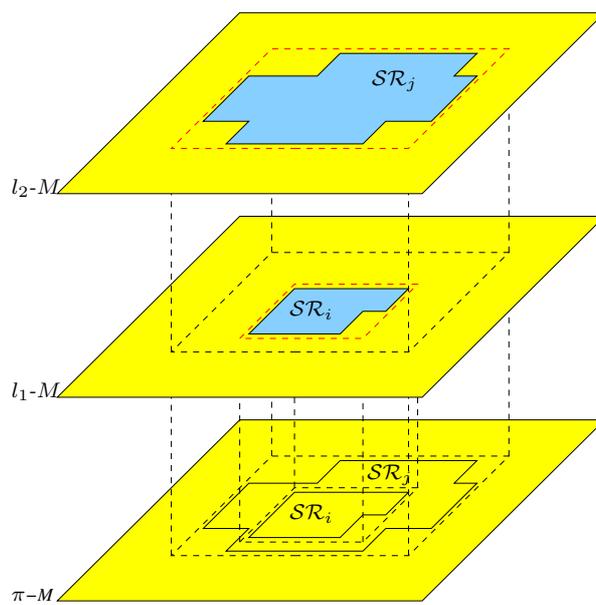


Figura 6.6: O fato  $l_1-SR_i \Rightarrow l_2-SR_j$  extraído de uma malha  $M$

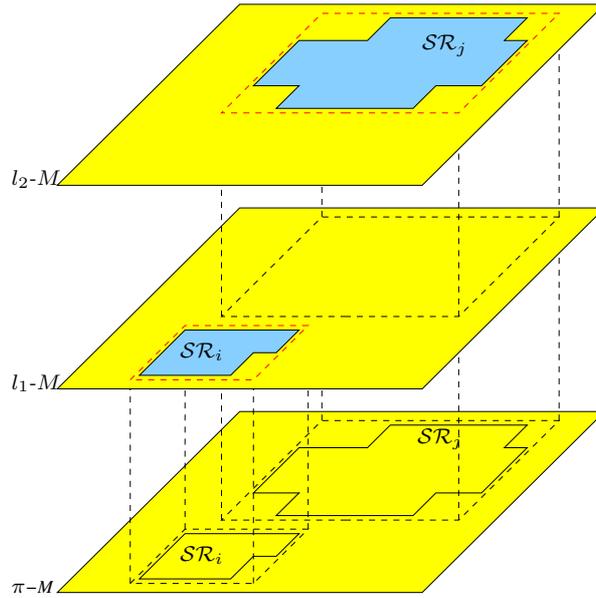


Figura 6.7: O fato  $\neg(l_1\text{-SR}_i \wedge l_2\text{-SR}_j)$  extraído de uma malha  $M$

**Definição 6.3.** *Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional ( $l > 3$ ),  $l_1\text{-}M$ ,  $l_2\text{-}M$  e  $l_3\text{-}M$  três camadas distintas da tesselação  $M$ . Se o conjunto de pontos resultante da intersecção das regiões  $l_3\text{-SR}_k$  e  $l_2\text{-SR}_j$  está contido no conjunto  $l_1\text{-SR}_i$  ( $l_3\text{-SR}_k \cap l_2\text{-SR}_j \subseteq l_1\text{-SR}_i$ , veja Figura 6.8), então  $(l_3\text{-SR}_k \wedge l_2\text{-SR}_j) \Rightarrow l_1\text{-SR}_i$ .*

Um envelope de um polígono também pode ser interpretado como uma lista de intervalos, um intervalo para cada dimensão do polígono. Para um polígono 2D que tem seu envelope limitado pelos quatro vértices  $(x_1, y_1)$ ,  $(x_2, y_1)$ ,  $(x_1, y_2)$  e  $(x_2, y_2)$  tem-se dois intervalos  $[\min\{x_1, x_2\}, \max\{x_1, x_2\}]$  e  $[\min\{y_1, y_2\}, \max\{y_1, y_2\}]$  denotando a variação das coordenadas dos seus pontos.

**Definição 6.4.** *Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional e  $l_1\text{-SR}_k$  uma região de uma dada camada da tesselação  $M$ . O envelope da região  $l_1\text{-SR}_k$ , denotado por  $l_1\text{-Env}_{\text{SR}_k}$ , é uma lista de pontos (vértices)*

$$l_1\text{-Env}_{\text{SR}_k} = ((x_1^1, x_1^2, \dots, x_1^n), \dots, (x_r^1, x_r^2, \dots, x_r^n)), \text{ para } r = 2^n.$$

**Definição 6.5.** *Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional e  $l_1\text{-SR}_k$  uma região de uma dada camada da tesselação  $M$ . O envelope da região  $l_1\text{-SR}_k$ , denotado por  $l_1\text{-Env}_{\text{SR}_k}$ , é também representado por uma lista de intervalos*

$$l_1\text{-Env}_{\text{SR}_k} - \delta = (\text{SR}_k - \delta^1, \dots, \text{SR}_k - \delta^n)$$

onde

$$\text{SR}_k - \delta^i = [\min\{x_1^i, x_2^i, \dots, x_r^i\}, \max\{x_1^i, x_2^i, \dots, x_r^i\}], \text{ para } r = 2^n.$$

**Definição 6.6.** *Sejam  $l_1\text{-SR}_i$  e  $l_2\text{-SR}_j$  duas regiões em camadas distintas da tesselação  $M$ , diz-se que  $l_1\text{-SR}_i \subseteq l_2\text{-SR}_j$  se, e somente se,  $\text{SR}_i - \delta^k \subseteq \text{SR}_j - \delta^k$ , para  $k = 1, \dots, n$ .*

**Definição 6.7.** *Sejam  $l_1\text{-SR}_i$  e  $l_2\text{-SR}_j$  duas regiões em camadas distintas da tesselação  $M$ , diz-se que  $l_1\text{-SR}_i \cap l_2\text{-SR}_j \neq \emptyset$  se, e somente se,  $\text{SR}_i - \delta^k \cap \text{SR}_j - \delta^k \neq \emptyset$ , para  $k = 1, \dots, n$ .*

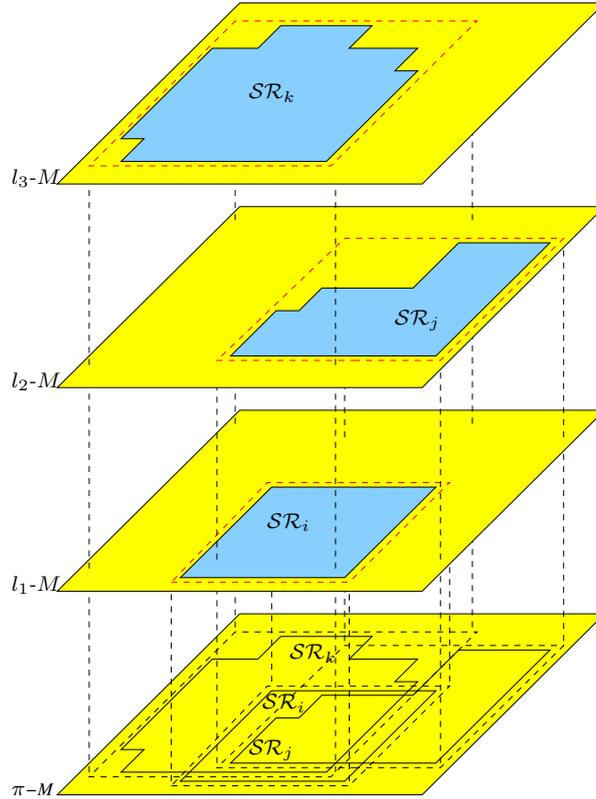


Figura 6.8: O fato  $(l_3\text{-SR}_k \wedge l_2\text{-SR}_j) \Rightarrow l_1\text{-SR}_i$  extraído de uma malha  $M$

**Definição 6.8.** Sejam  $l_1\text{-SR}_i$  e  $l_2\text{-SR}_j$  duas regiões em camadas distintas da tesselação  $M$ , diz-se que  $l_1\text{-SR}_i \cap l_2\text{-SR}_j = \emptyset$  se, e somente se,  $SR_i\text{-}\delta^k \cap SR_j\text{-}\delta^k = \emptyset$ , para  $k = 1, \dots, n$ .

O método da detecção por envelope pode não resultar em uma solução ótima, dependendo do tipo de polígono gerado pela sub-divisão, já que este corresponde a uma aproximação exterior do polígono formador da sub-região. Entretanto, por ser um método bastante eficiente, é factível utilizá-lo preliminarmente a outros métodos de verificação de inclusão de pontos, para eliminar verificações desnecessárias.

Por outro lado, pode-se utilizar idéia análoga para a aproximação interior (veja Figura 6.9) de cada polígono formador de uma dada sub-região.

**Definição 6.9.** Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional e  $l_1\text{-SR}_k$  uma sub-região de uma dada camada da tesselação  $M$ . A coleção de intervalos interiores da região  $l_1\text{-SR}_k$ , denotado por  $l_1\text{-Int}_{SR_k}$ , é uma lista de tuplas com  $n$  intervalos

$$l_1\text{-Int}_{SR_k} = ((SR_k\text{-}\delta_1^1, \dots, SR_k\text{-}\delta_1^n), \dots, (SR_k\text{-}\delta_q^1, \dots, SR_k\text{-}\delta_q^n))$$

onde cada tupla  $(SR_k\text{-}\delta_j^1, \dots, SR_k\text{-}\delta_j^n)$  indica uma sub-região  $SR_k\text{-}\delta_j$  distinta<sup>1</sup> no espaço  $n$ -dimensional formada por  $n$  intervalos dentro da região  $SR_k$ . A quantidade  $q$  de tuplas de  $n$  intervalos é determinada de modo que

$$|SR_k| = \varepsilon + \sum_{i=1}^q \prod_{d=1}^n \mathcal{D}(SR_k\text{-}\delta_i^d)$$

<sup>1</sup>Para todas tuplas tomadas duas a duas, existem pelo menos dois intervalos disjuntos, um para cada tupla, correspondentes à mesma coordenada  $i$ .

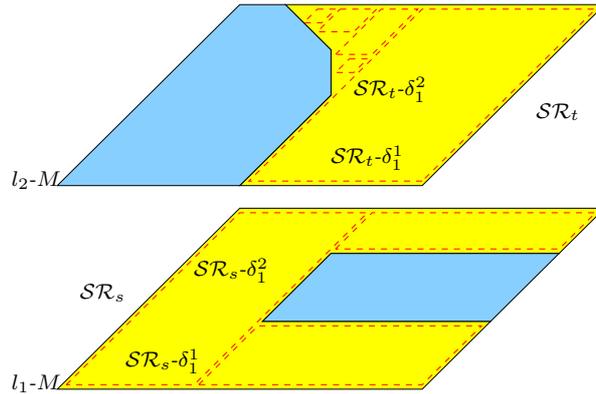


Figura 6.9: Aproximação interior de uma sub-região através de intervalos

onde  $\mathcal{D}(\mathcal{SR}_k - \delta_i^d) = \sup(\mathcal{SR}_k - \delta_i^d) - \inf(\mathcal{SR}_k - \delta_i^d)$  é o diâmetro do intervalo  $\mathcal{SR}_k - \delta_i^d$ ,  $\sup(\mathcal{SR}_k - \delta_i^d)$  ( $\inf(\mathcal{SR}_k - \delta_i^d)$ ) é o extremo superior (inferior) de  $\mathcal{SR}_k - \delta_i^d$ , e  $\varepsilon$  é o limite de erro admissível para a aproximação interior da sub-região.

**Definição 6.10.** Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional,  $l_1$ - $\mathcal{SR}_s$  e  $l_2$ - $\mathcal{SR}_t$  regiões de camadas distintas da tesselação  $M$ . Diz-se que  $\mathcal{SR}_s - \delta_i \cap \mathcal{SR}_t - \delta_j$  é a tupla resultante de

$$(\mathcal{SR}_s - \delta_i^1 \cap \mathcal{SR}_t - \delta_j^1, \dots, \mathcal{SR}_s - \delta_i^n \cap \mathcal{SR}_t - \delta_j^n).$$

**Definição 6.11.** Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional,  $l_1$ - $\mathcal{SR}_s$  e  $l_2$ - $\mathcal{SR}_t$  regiões de camadas distintas da tesselação  $M$ . Diz-se que  $\mathcal{SR}_s - \delta_i \cup \mathcal{SR}_t - \delta_j$  é a lista de tuplas

$$(\mathcal{SR}_s - \delta_i^1, \dots, \mathcal{SR}_s - \delta_i^n, \mathcal{SR}_t - \delta_j^1, \dots, \mathcal{SR}_t - \delta_j^n).$$

**Definição 6.12.** Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional,  $l_1$ - $\mathcal{SR}_s$  e  $l_2$ - $\mathcal{SR}_t$  regiões de camadas distintas da tesselação  $M$ . Diz-se que  $\mathcal{SR}_s \cap \mathcal{SR}_t \neq \emptyset$  se, e somente se

$$\bigcup_{\mathcal{SR}_s - \delta \in l_1 - \text{Int}_{\mathcal{SR}_s}} \left( \bigcup_{\mathcal{SR}_t - \delta \in l_2 - \text{Int}_{\mathcal{SR}_t}} (\mathcal{SR}_s - \delta \cap \mathcal{SR}_t - \delta) \right) \neq \emptyset.$$

**Definição 6.13.** Sejam  $M$  uma tesselação de  $l$ -camadas e  $n$ -dimensional,  $l_1$ - $\mathcal{SR}_s$  e  $l_2$ - $\mathcal{SR}_t$  regiões de camadas distintas da tesselação  $M$ . Diz-se que  $\mathcal{SR}_s \subseteq \mathcal{SR}_t$  se, e somente se

$$l_1 - \text{Int}_{\mathcal{SR}_s} = \bigcup_{\mathcal{SR}_s - \delta \in l_1 - \text{Int}_{\mathcal{SR}_s}} \left( \bigcup_{\mathcal{SR}_t - \delta \in l_2 - \text{Int}_{\mathcal{SR}_t}} (\mathcal{SR}_s - \delta \cap \mathcal{SR}_t - \delta) \right).$$

Outro método que pode ser utilizado é a verificação da inclusão de ponto-em-polígono (FOLEY et al., 1990; PREPARATA; SHAMOS, 1985).

O problema é determinar se um ponto  $p$  é interior, exterior ou está na fronteira do polígono simples  $P = p_1 p_2 \dots p_n$ , conforme a Figura 6.10. Uma solução para este problema é considerar uma semi-reta  $L$  partindo de  $p$  e determinar seus pontos de intersecção com a linha poligonal. Se  $p$  coincidir com um destes pontos de intersecção, conclui-se que ele está na fronteira de  $P$ . Senão, basta contar quantas vezes a semi-reta atravessa a

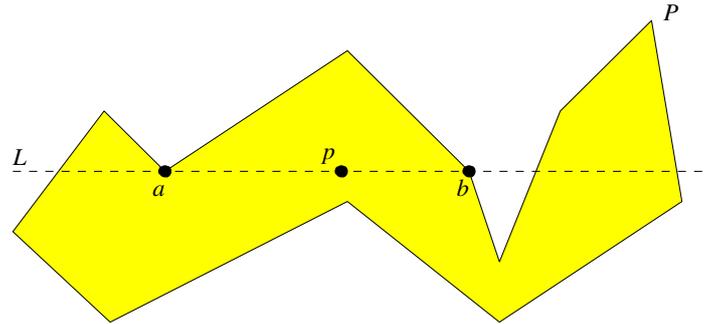


Figura 6.10: Problema de Ponto em Polígono

poligonal: se o número de cruzamentos for ímpar, o ponto  $p$  é interior; caso contrário,  $p$  é exterior.

Quando um ponto de intersecção é um vértice, não é necessariamente verdade que  $L$  passe do interior para o exterior (ou vice-versa) naquele ponto. Por exemplo, na Figura 6.10, o ponto  $a$  não deve ser contado como ponto de intersecção, enquanto  $b$  deve.

A Figura 6.11 mostra cada caso possível e o número de intersecções correspondente.

Casos						
Intersecções	0	2	1	0	2	1

Figura 6.11: Casos de cruzamento e número de intersecções

Nota-se que este método apesar de ótimo é bastante custoso, pois esta verificação deve ser feita em todo ponto da região que está sendo confrontada. Ainda, para o caso multi-dimensional, esta verificação deve ser feita tomando a combinação duas a duas de todas as projeções ortogonais.

### 6.1.1 Conhecimento Quantitativo Probabilístico

A abordagem quantitativa é inspirada na teoria da probabilidade, mais especificamente na definição de probabilidade condicional (DRAKE, 1967).

No modelo *ICTM*, um evento corresponde a uma região da área geográfica e a probabilidade de ocorrência de um determinado evento indica a razão entre o tamanho da região e a totalidade da área.

**Definição 6.14.** A probabilidade de ocorrência de uma região  $l_1\text{-SR}_s$  denotada por  $P(l_1\text{-SR}_s)$ , é tal que

$$P(l_1\text{-SR}_s) = \frac{|l_1\text{-SR}_s|}{a_1 \times \dots \times a_n}.$$

Na Figura 6.12, pode-se conferir uma tesselação  $M$  composta por 100 células idênticas, que representam uma discretização de uma área geográfica. Nesta área ocorrem as regiões  $l_1\text{-SR}_s$  e  $l_2\text{-SR}_t$  em áreas de diferentes tamanhos, apresentando alguma superposição.

Pode-se inferir que os números de células comuns às duas regiões (identificadas como o conjunto  $l_1\text{-SR}_s \cap l_2\text{-SR}_t$ ) podem ser expressos como probabilidades de cada uma delas.

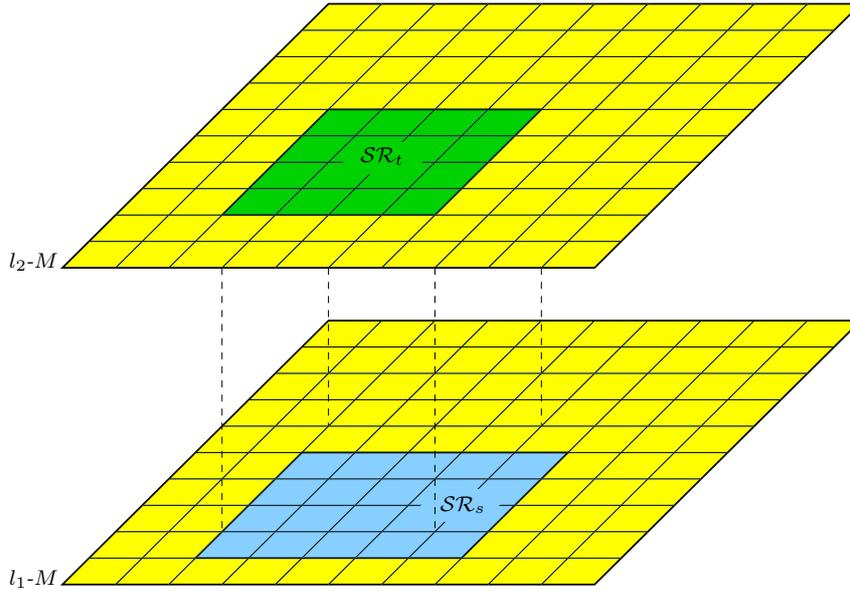


Figura 6.12: Matriz de Ocorrências

O número de células ( $N$ ) referentes a cada região é:  $N(M) = 100$ ;  $N(l_1-SR_s) = 20$ ;  $N(l_2-SR_t) = 16$  e  $N(l_1-SR_s \text{ e } l_2-SR_t) = N(l_2-SR_t \text{ e } l_1-SR_s) = 12$ .

Ainda, pode-se constatar que  $N(l_1-SR_s \text{ e } l_2-SR_t)$  vale  $\frac{12}{20}$  de  $l_1-SR_s$ . A mesma região, expressa como  $N(l_2-SR_t \text{ e } l_1-SR_s)$ , corresponde a  $\frac{12}{16}$  de  $l_2-SR_t$ . Então, pode ser afirmado que  $N(l_1-SR_s \text{ e } l_2-SR_t)$  ou  $N(l_2-SR_t \text{ e } l_1-SR_s)$  corresponde a 60% de  $l_1-SR_s$  e 75% de  $l_2-SR_t$ .

Em termos de  $l_1-SR_s$ , as probabilidades de ocorrência de  $l_1-SR_s$ , uma vez constatada a presença de  $l_2-SR_t$ , expressa a probabilidade condicional  $P(l_1-SR_s|l_2-SR_t)$ . Analogamente, em termos de  $l_2-SR_t$ , as probabilidades de ocorrência de  $l_2-SR_t$ , uma vez constatada a presença de  $l_1-SR_s$ , expressa a probabilidade condicional  $P(l_2-SR_t|l_1-SR_s)$ .

**Definição 6.15.** A probabilidade de ocorrência de  $l_1-SR_s$ , uma vez constatada a presença de  $l_2-SR_t$ , está para a probabilidade de ocorrência de  $l_2-SR_t$ , uma vez constatada a presença de  $l_1-SR_s$ , assim como a probabilidade de  $l_1-SR_s$  está para a probabilidade de  $l_2-SR_t$ , de modo que

$$\frac{P(l_1-SR_s|l_2-SR_t)}{P(l_2-SR_t|l_1-SR_s)} = \frac{P(l_1-SR_s)}{P(l_2-SR_t)}.$$

**Definição 6.16.** A probabilidade condicional de uma região  $l_1-SR_s$  dada a presença da região  $l_2-SR_t$  é

$$P(l_1-SR_s|l_2-SR_t) = \frac{P(l_1-SR_s \text{ e } l_2-SR_t)}{P(l_2-SR_t)}$$

onde:  $P(l_2-SR_t)$  é a probabilidade de ocorrência da sub-região  $l_2-SR_t$  e  $P(l_1-SR_s \text{ e } l_2-SR_t)$  é a probabilidade das sub-regiões  $l_1-SR_s$  e  $l_2-SR_t$  ocorrerem simultaneamente.

A relação conhecida como teorema de Bayes é resultado de uma aplicação particular da definição de probabilidade condicional. Aplica-se esta relação no caso onde eventos, no contexto deste projeto chamados de sub-regiões,  $l_1-SR_1, l_1-SR_2, \dots, l_1-SR_{SR}$  formam um lista de eventos mutuamente exclusivos, conforme a Figura 6.13.

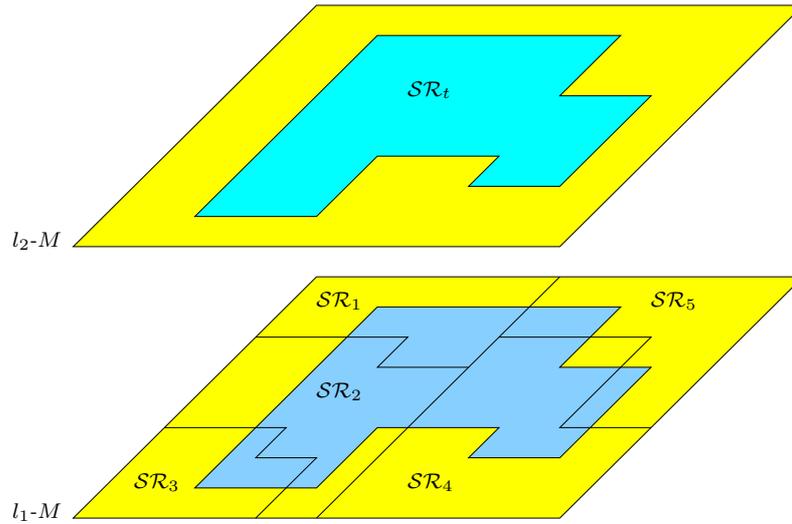


Figura 6.13: Coleção de eventos mutuamente exclusivos

**Definição 6.17.** O teorema de Bayes estabelece a relação entre as sub-regiões e é dado da seguinte maneira

$$P(l_1-SR_i|l_2-SR_t) = \frac{P(l_1-SR_i)P(l_2-SR_t|l_1-SR_i)}{\sum_{i=1}^{SR} P(l_1-SR_i)P(l_2-SR_t|l_1-SR_i)} = \frac{P(l_2-SR_t|l_1-SR_i)}{\sum_{i=1}^{SR} P(l_2-SR_t|l_1-SR_i)}$$

com  $P(l_2-SR_t) \neq 0$  e  $l_1-SR_1, \dots, l_1-SR_{SR}$  uma coleção de sub-regiões de  $l_1-SR_s$ .

## 7 IMPLEMENTAÇÃO DO MODELO

Nesta seção, apresentam-se as considerações sobre a implementação do modelo *Topo-ICTM* (SILVA et al., 2003), cuja formalização matemática foi apresentada no capítulo 4.

O modelo *Topo-ICTM* constitui-se na versão bidimensional de uma camada do modelo genérico *ICTM* (AGUIAR; COSTA; DIMURO, 2004), para considerar dados de entrada oriundos de imagens de satélite (coordenadas  $x$ ,  $y$  e  $z$ ) e realizar uma categorização de acordo com a monotonicidade do relevo.

Para aplicação do modelo, estabelece-se uma malha que representa a região a ser analisada. A cada célula da malha é associado um valor de altura (da elevação do terreno). A análise da monotonicidade da função que mapeia a topografia do terreno (função relevo) da região considerada está embutida nas regras do modelo, que detecta a situação de um segmento de área em relação à área total, através do estado assumido pela célula. Este estado é definido pelas regras do modelo, que analisa a variação das alturas. O modelo é regulado por dois aspectos:

- Pelo número de pontos da malha topográfica que representa a área analisada, que é uma característica determinada pela tomada de informações obtidas das imagens de satélite;
- Pela resolução espacial (raio da vizinhança da célula), que é um parâmetro interno do sistema.

### 7.1 Etapas de funcionamento do *ICTM*

#### 7.1.1 Definição da Malha

A definição da malha a ser utilizada no modelo é uma etapa muito importante para a precisão dos resultados e pode ser determinante para o sucesso do processamento. Isso acontece porque o tamanho de cada região está diretamente associado ao refinamento do modelo.

O tamanho e dimensão da malha são determinados pela resolução espacial das imagens de satélite. De acordo com os pontos da imagem em que se conhecem os dados de elevação, define-se a dimensão (número de linhas e colunas) da matriz cujos elementos representam as células. Através dessa definição, o sistema saberá qual o tamanho da região que será analisada.

#### 7.1.2 Caracterização Topográfica das Áreas Geográficas

Tomam-se os dados topográficos a partir do modelo digital de elevação, ou seja, a altura  $h(x, y)$  descrita como uma função da longitude  $x$  e latitude  $y$ . Estes dados to-

pográficos são os valores iniciais para a construção da matriz absoluta.

## 7.2 Formação da Matriz Absoluta

Após a definição da malha e da obtenção dos dados topográficos, monta-se a matriz absoluta. Observa-se que muitos dos dados levantados são irrelevantes devido à abundância de informações proporcionada pelo modelo digital de elevação. É então realizada uma média dos valores levantados pela tomada dos dados topográficos. Assim, o resultado desta média de valores forma uma matriz de valores absolutos, e são esses valores que servem de fonte dos dados para o sistema.

Além disso, é necessária uma simplificação dos dados. Isso é feito através de uma normalização dos valores absolutos, dividindo cada valor da matriz pelo maior destes valores.

### 7.2.1 Utilização de Técnicas Intervalares para Controle dos Erros

Utilizam-se intervalos numéricos (MOORE, 1979) para representar os erros de discretização dos valores associados às células do modelo. Após a formação da matriz relativa é necessário gerar uma matriz intervalar dos valores relativos no eixo  $x$  e no eixo  $y$ .

A implementação desta matriz intervalar foi realizada conforme descrito a seguir. O sistema identifica, para cada célula, a variação de valores de altura em relação às suas células vizinhas à oeste e à leste, no eixo  $x$ , e ao norte e ao sul, no eixo  $y$ .

O programa compara o valor da célula analisada, diminuindo o valor da mesma pelo valor de suas células vizinhas, obtendo o módulo dessa variação. Esse resultado é dividido por dois, diminuído do valor da célula em questão para gerar o primeiro valor do intervalo, e somado ao valor da célula para gerar o segundo valor do intervalo. Com isso, consegue-se garantir que todos os valores serão representados dentro do intervalo formado.

### 7.2.2 Determinação dos Estados das Células do Modelo

Para a determinação dos estados das células do modelo, são utilizados quatro registradores de declividade (norte, sul, leste e oeste), já que o modelo de vizinhança adotado é inspirado no modelo de Von Neumann (VON NEUMANN; BURKS, 1966) (veja a Figura 7.1). Estes quatro registradores são utilizados para a determinação do estado da célula em relação às suas vizinhas. Para isso, cada registrador indica se o valor da função (de acordo com os valores da matriz espectral) está crescendo no respectivo sentido. A análise da declividade é efetuada nas direções das quatro células vizinhas.

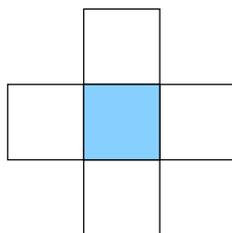


Figura 7.1: Vizinhança de Von Neumann

A implementação nesta fase foi realizada de forma que o sistema gere quatro matrizes de monotonicidade. A matriz de monotonicidade ao norte, ao sul, ao leste e à oeste.

Para cada célula da malha, o sistema compara se o limite superior do intervalo da célula em questão é menor ou igual ao limite inferior do intervalo da célula vizinha ao leste, atribuindo os valores 0 ou 1 aos registradores de declividade, conforme a função relevo seja não crescente ou não decrescente (respectivamente). Este raciocínio é empregado para a geração das demais matrizes (norte, sul e oeste), observando o sentido da análise e os valores do intervalo.

Após o cálculo e a formação das quatro matrizes de monotonicidade, o sistema gera uma outra matriz, denominada de matriz de *status*. Essa matriz nada mais é do que a matriz que contém o estado de cada célula que detecta a situação de cada segmento em relação à área total. Essa matriz é calculada através da soma dos registradores de declividade multiplicados pelos seus respectivos pesos, previamente definidos. É importante salientar que cada estado calculado é, necessariamente, um dos estados previstos no modelo.

Cada célula poderá assumir um e somente um estado apresentado na Tabela 4.2, cujo valor associado é calculado de acordo com a equação:

$$Status\ xy = (1 \times reg.e) + (2 \times reg.s) + (4 \times reg.w) + (8 \times reg.n).$$

### 7.3 Identificação das Regiões Topográficas

Para cada célula, além dos quatro registradores de declividade (*reg.n*, *reg.s*, *reg.e* e *reg.w*), há também um registrador para indicar se a célula é limite de uma área, ou seja, se a função está passando de crescente para decrescente (ou vice-versa). Esse registrador é chamado de registrador *limit*.

Assim, para identificar as células limites das regiões topográficas faz-se a análise do registrador *limit* das células da malha. Toda a célula que tiver este registrador setado com o valor 1 é considerada como célula limite de uma sub-área, ou seja, esta é uma célula que compreende uma área cujas características topográficas estão tomando uma nova configuração. Por convenção, todas as células que compõem a borda da matriz têm o seu registrador de limite definido com o valor 1, pois as células da borda limitam uma área, que é a área total.

Uma análise da matriz de limites permite detectar a existência de configurações de relevo conhecidas. A Figura 4.3, na página 56, mostra exemplos de esquemas de configuração de relevo conhecidos. Observa-se que, sempre que o registrador *limit* tem o valor 1, o sinal da declividade da função relevo muda, fazendo com que esta função passe de crescente para constante (ou decrescente), ou de decrescente para constante (ou crescente) ou constante para crescente (ou decrescente).

A presença de células limite permite uma subdivisão da área considerada em categorias de declividade. Isso é realizado pela implementação de um algoritmo recursivo geralmente utilizado para preenchimento de polígonos. O algoritmo percorre a matriz, verificando os registradores de limite das células em busca de células que não são limítrofes (valor zero no registrador). Ao encontrar a primeira célula que satisfaz essa condição, ele armazena a posição (linha e coluna) da mesma em uma pilha (que indica as células que devem ser verificadas) e dá início ao processo de contagem. O algoritmo verifica se as células vizinhas à célula em questão também possuem o registrador limite com o valor zero. Em caso afirmativo, suas posições também são armazenadas e a célula que deu origem a contagem é descartada (excluída da pilha) depois da análise das suas células vizinhas. Este processo é realizado para toda célula que está na pilha. No término deste

procedimento, o contador de regiões é incrementado e o processo é reiniciado fazendo nova busca de alguma célula ainda não analisada, parando somente após a constatação de que não existem mais células com o registrador limite igual a zero em toda malha.

Como exemplo, veja a matriz de limites produzida pelo modelo no processo de categorização de uma dada área R, mostrada na Tabela 4.2. A categorização de declividade para a área R é mostrada na Figura 4.4 e esta foi obtida pelo sistema *Topo-ICTM*.

## 7.4 Considerações Finais Sobre a Implementação

O protótipo do sistema *Topo-ICTM* foi implementado de forma seqüencial, no sistema operacional Linux, através da utilização da linguagem C/C++.

Além disso, está em andamento a implementação de uma versão paralela do modelo voltada para utilização no cluster de PC's do grupo de pesquisa GMFC - Grupo de Matemática e Fundamentos da Computação da UCPel.

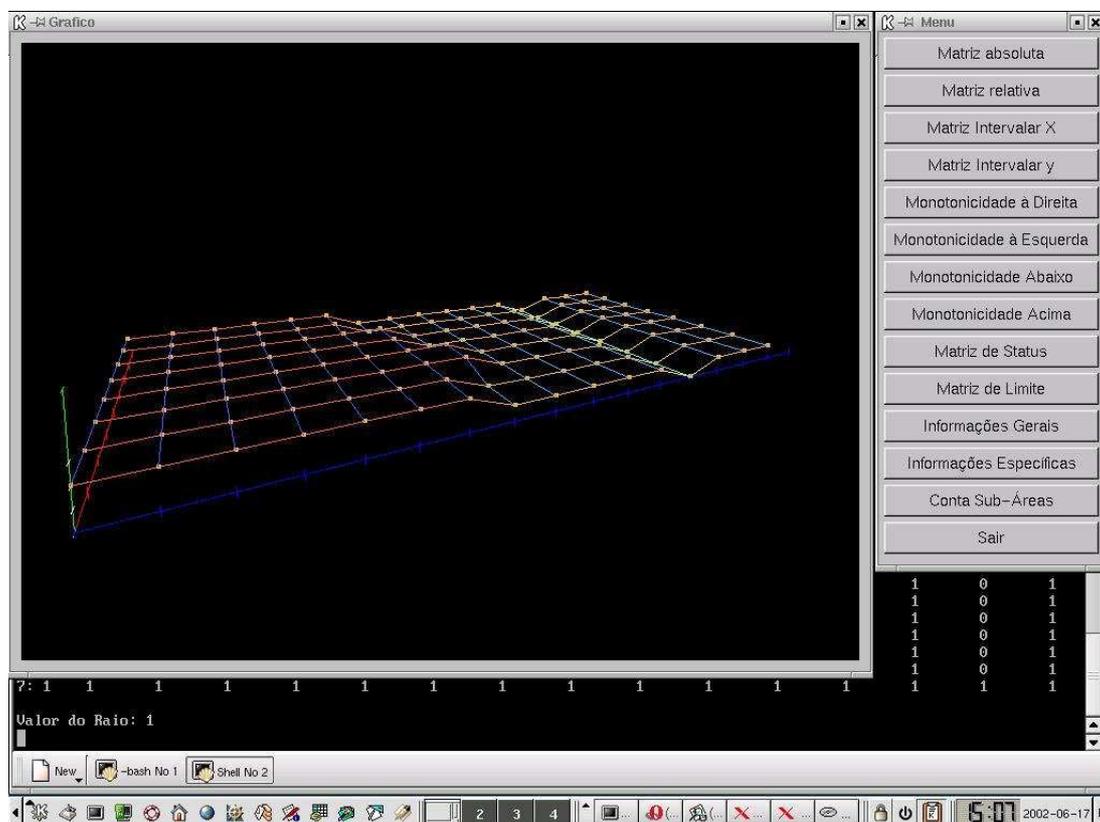


Figura 7.2: Interface do Protótipo

Essa implementação utiliza a biblioteca de paralelização MPI a fim de obter um ganho de performance considerável no *software*, possibilitando a realização de análises em grandes regiões, com uma quantidade elevada de informações. Assim, será possível também realizar uma comparação entre a implementação seqüencial e a implementação paralela do sistema.

Quanto à interface do protótipo, implementada em OpenGL, esta contém apenas uma janela com um menu, onde todas as informações sobre as matrizes geradas estão disponíveis, bastando o usuário clicar nos botões e verificar o resultado e um sistema para a representação gráfica da função que mapeia o relevo.

A Figura 7.2 mostra a interface do protótipo do sistema *Topo-ICTM*. Ela é composta por três janelas: (i) a janela de visualização, que apresenta as informações visuais do relevo, sub-áreas e limites; (ii) a janela do console, que apresenta informações das células e da malha, e (iii) cardápio de opções (à direita) com as operações disponíveis no sistema.

## 8 RESULTADOS EXPERIMENTAIS

Os resultados apresentados a seguir utilizam os Modelos de Elevação Digital (DEM) de resolução 1000m e 500m do quadrante com coordenadas:

- Canto superior Esquerdo:  $X = 427559m$ ,  $Y = 6637852m$ .
- Canto inferior Direito:  $X = 480339m$ ,  $Y = 6614507m$ .

Estas são coordenadas UTM (Universal Transversa de Mercator), Fuso 22S (Zona UTM 22, hemisfério sul) e Datum SAD69 (Datum da América do Sul)

A Figura 8.1 representa o DEM de resolução 1000m horizontalmente e 500m verticalmente, compreendendo 24 linhas e 53 colunas, obtido a partir da digitalização de cartas topográficas na escala 1:1.000.000.

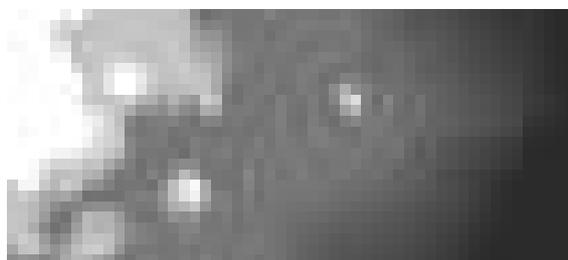


Figura 8.1: Imagem do DEM de resolução 1000m

A Figura 8.2 representa o DEM de resolução 500m horizontalmente e verticalmente, compreendendo 49 linhas e 106 colunas, obtido a partir da interpolação de cartas topográficas digitalizadas na escala 1:1.000.000.



Figura 8.2: Imagem do DEM de resolução 500m

Os resultados são apresentados nas Tabelas 8.1 (DEM 1000m) e 8.2 (DEM 500m). Pode-se observar claramente que no *ICTM* o número de categorias é inversamente proporcional ao raio da vizinhança.

Tabela 8.1: Resultados para o quadrante no DEM 1000m

Raio de Vizinhança	Nº de Categorias
1	76 (Figura A.1)
2	62 (Figura A.2)
5	36 (Figura A.3)
10	22 (Figura A.4)
20	18 (Figura A.5)
40	18 (Figura A.6)

Além disso, pode-se perceber que para o DEM 500m, cujo ponto congrega uma área 4 vezes menor que o ponto correspondente no DEM 1000m, o número de categorias também seguiu aproximadamente este fator. Neste caso, o DEM com resolução de 500m, especificamente desta região, não apresentou novas regiões a serem categorizadas, apenas detalhou as regiões encontradas. No entanto, em regiões com maiores variações de declividade possivelmente este fator não seja mais verificado.

Ainda, em áreas mais planas os raios de vizinhança maiores foram boas aproximações para as categorias. Por exemplo, observe a região A na Figura A.1, suas representações com raios maiores (Figuras A.2, A.3, A.4, A.5 e A.6) indicam aproximações razoáveis para este tipo de declividade. Entretanto, para regiões com maiores variações de declividades, os melhores resultados (mais detalhados) foram alcançados com raios menores.

Tabela 8.2: Resultados para o quadrante no DEM 500m

Raio de Vizinhança	Nº de Categorias
1	230 (Figura A.7)
2	197 (Figura A.8)
5	143 (Figura A.9)
10	125 (Figura A.10)
20	108 (Figura A.11)
40	83 (Figura A.12)

Regiões com maiores concentrações de categorias (por exemplo, a região B na Figura A.1) são indicadoras que uma análise mais refinada deve ser feita (veja a representação no DEM de 500m de resolução correspondente na Figura A.7). O refinamento do modelo é dado por dois aspectos:

- pela resolução espacial do modelo digital de elevação
- pelo raio da vizinhança da célula

Assim, regiões com várias células limites podem ser melhores estudadas com o aumento da resolução dos dados de altimetria, ou com a redução do raio de vizinhança.

No *ICTM*, o estado de uma célula em relação aos seus vizinhos, em termos de declividade, pode ser verificado instantaneamente, contrastando com as análises usuais presentes nos SIG's (conforme as Figuras 8.3 e 8.4).

Na informação gerada por contornos (curvas de nível) a informação é global, permitindo a visualização do estado da célula em função de células distantes. No *ICTM*, a informação é pontual, associando várias propriedades no mesmo ponto.

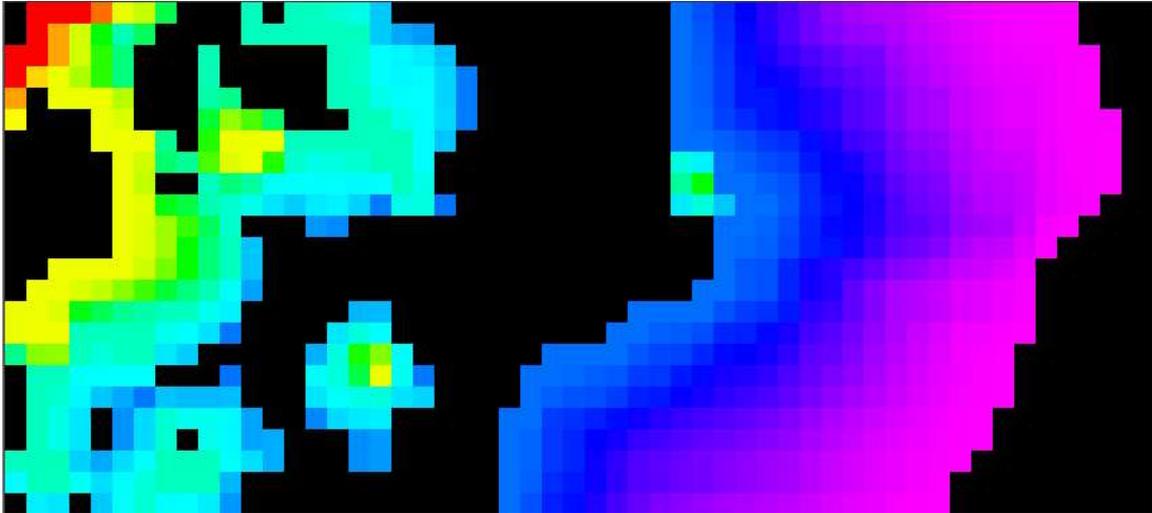


Figura 8.3: Linhas de Contorno no DEM de resolução 1000m

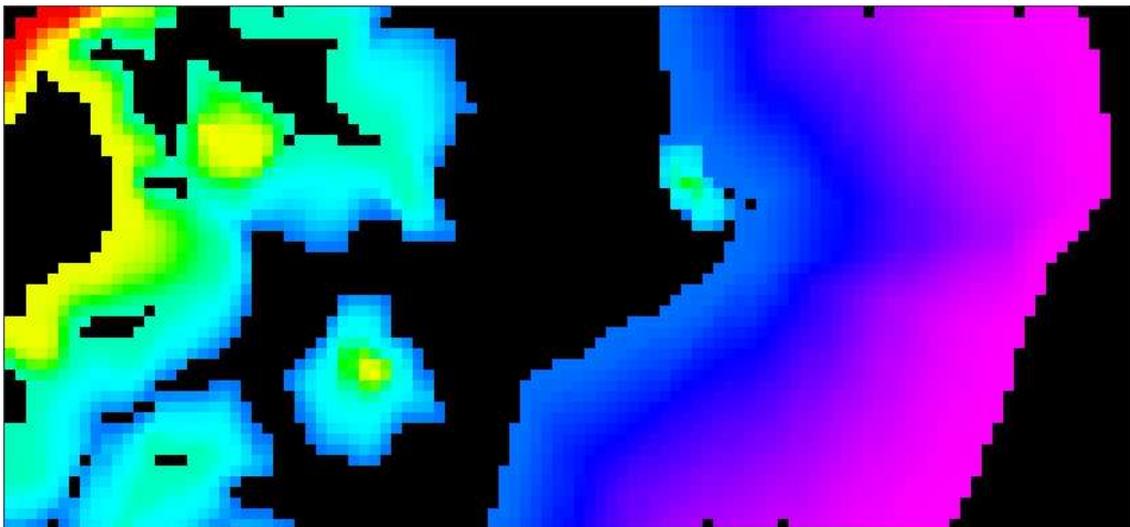


Figura 8.4: Linhas de Contorno no DEM de resolução 500m

Também percebe-se que o tamanho da área, a resolução do DEM referente à esta área e o grau de variação da declividade são as propriedades mais importantes na determinação de um raio de vizinhança apropriado para uma categorização significativa.

Os resultados até aqui obtidos indicam que regiões com menor variação de declividade são receptivas ao *ICTM* com raios maiores. Ao contrário, regiões com grandes variações de declividade sugerem raios menores. A Figura 8.5 indica as classes conforme o grau de declividade do modelo de elevação digital com resolução de 1000m.

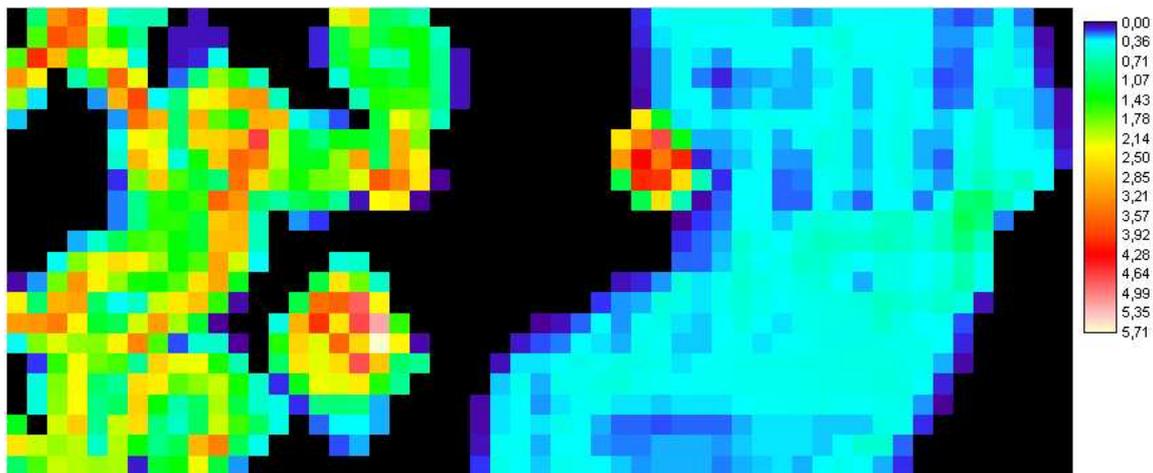


Figura 8.5: Graus de Declividade do DEM de resolução 1000m

Evidentemente, regiões pequenas comportam apenas raios menores, pois raios de vizinhança maiores tendem a suavizar um área muito grande ao redor da célula em questão.

## 9 CONCLUSÕES

O *ICTM* é um modelo baseado em tesselação genérico que é capaz de produzir confiáveis categorizações de regiões de uma dada área geográfica, de acordo com múltiplas características que são conhecidas em um número suficiente de pontos.

A categorização determinada por cada característica é determinada em uma das camadas do modelo, gerando diferentes subdivisões da área analisada. Por exemplo, uma área pode ser analisada de acordo com sua topografia, vegetação, demografia, dados sócio-econômicos etc.

O modelo de tesselação genérico não é restrito à análise de espaços bidimensionais. O conjunto de pontos analisados pode pertencer a um espaço multi-dimensional, determinando o caráter multi-dimensional de cada camada. Um procedimento do tipo projeção da categorização de cada camada, sobre uma camada base, permitirá uma categorização confiável e significativa, que combinará a análise executada para cada característica. Isto permitirá uma análise interessante da dependência mútua dessas características.

No caso específico do *Topo-ICTM*, considera-se apenas uma característica, a declividade da função que mapeia o relevo da área considerada, executando uma análise bidimensional, que considera latitude e longitude.

A dimensão da tesselação pode ser arbitrária, ou escolhida de acordo com um critério específico estabelecido pelo especialista que conduz a aplicação, ou determinada pela natureza dos dados de entrada.

Em qualquer caso, a categorização obtida pode ser refinada pela definição, por exemplo, de uma outra dimensão, ou tomando cada região resultante para ser analisada separadamente. A análise pode ser executada até que um número conveniente de regiões seja obtido, caracterizando o dinamismo do modelo.

O *ICTM* é um modelo inerentemente (puramente) paralelo, que fornece subsídios aos métodos matemáticos e estatísticos comumente utilizados pelas técnicas de categorização. Ao invés de se deter em paralelizar os métodos de interpolação ou os algoritmos de métricas, o *ICTM* abstrai os modelos matemáticos e toda a análise estatística requerida pelas técnicas usuais, utilizando um modelo simples, baseado na verificação local de propriedades, permitindo, assim, o domínio ser facilmente decomposto e preparado para o paralelismo.

Além disso, permite diferentes categorizações conforme o raio de vizinhança, sem necessitar de redefinição da malha (tesselação), como os métodos utilizados para geração de malhas por elementos finitos.

O método de segmentação de áreas geográficas, apresentado no *ICTM*, baseado na monotonicidade da função declividade que mapeia o relevo se mostrou como uma ferramenta bastante poderosa no que diz respeito à quantidade de informação associada a cada célula da tesselação. Como pôde ser observado no capítulo 8, os resultados indica-

ram que a informação contida localmente nos pontos da malha é mais representativa e é instantaneamente inferida a partir da análise dos pontos mais próximos.

A extração de fatos em SIG's é inédita. Normalmente, um SIG permite que sejam aplicados os operadores booleanos para a geração de novos mapas (temáticos). Assim como a utilização de intervalos como algoritmo aproximativo para a extração de fatos.

Como comentado na seção 2.6, o *ICTM* permite a compactação das categorias, dando uma descrição abstrata dos dados presentes nas células quando representa uma categoria por uma coleção de intervalos. Assim, uma descrição compacta das categorias permite maior *performance* dos algoritmos de tomada de decisões.

Quanto à estrutura e funcionamento algorítmico, o *ICTM* pode ser considerado simultaneamente como um método ótimo de construção de categorias, quando o raio de vizinhança é 1 ele encontra exatamente todas as categorias presentes, e aproximativo quando o raio é maior que 1, encontrando as aproximações das categorias.

Esta flexibilidade permite tanto a análise de áreas com propriedades mais homogêneas (no caso do *Topo-ICTM*, regiões com pequena variação de declividade) quanto em regiões com características bem distintas (alto grau de variação de declividade).

O modelo tridimensional associado aos limites e orientação dos vetores de inclinação direcionam a caracterização de subregiões para um caráter de maior qualificação e quantificação, com aplicações em geofísica, agronomia, na área florestal e hidrologia.

A informação gerada, comparada a informações hipsométricas geradas por *softwares* usuais, possui um caráter informativo com maior qualificação, quantificação e precisão, ou seja, se tem um ganho pontual relacionado a informação global.

Dentre as aplicações do modelo *ICTM*, na análise de superfícies topográficas e a batimetria destas superfícies, pode-se destacar os seguintes trabalhos futuros: (i) cálculos de volume e caracterização do perfil para dragagem; (ii) caracterização de unidades geológicas e geofísicas; (iii) utilização das propriedades de declividade e orientação em modelos hidrológicos, análise de transporte de sedimentos, balanços de energia, identificação de padrões de vegetação; (iv) diferenças micro-climáticas; (v) caracterização da profundidade efetiva do solo; (vi) caracterização do regime de nutrientes do solo; (vii) análise do risco de erosão (análise de transporte de sedimentos); (viii) necessidade de práticas especiais de conservação de solo etc.

Outra aplicação deste trabalho poderia ser a análise das variações de crescimento das árvores dentro de uma classe topográfica homogênea. Este tipo de análise é frequentemente utilizada para identificar as variações na disponibilidade de suprimento de água, ar e nutriente do solo, as quais estão intimamente relacionadas com as propriedades do solo, com especial destaque para a textura do solo.

Para o prosseguimento deste trabalho pretende-se utilizar o conceito de autômatos celulares para estender este tipo de análise categorizadora. Isso permitiria não apenas uma análise estática das características do espaço em um momento dado, mas também possibilitaria o avanço na direção de uma simulação dos aspectos dinâmicos desse espaço, em um intervalo de tempo. Assim, permitiria-se a modelagem de processos dinâmicos mais complexos apresentados esquematicamente na Figura 1.1.

Uma vez adicionada esta característica dinâmica no modelo *ICTM*, as seguintes análises poderiam ser efetuadas: (i) simulação da dispersão de óleo no oceano; (ii) simulação da dispersão de sedimentos no estuário; (iii) simulação da dispersão de contaminantes químicos em águas subterrâneas; (iv) simulação do comportamento espacial sócio-econômico, relacionados com as características físicas do meio urbano; (v) determinação do zoneamento agroclimático, atualmente fornecido em intervalos de 10

dias, sendo que os dados meteorológicos são diários; (vi) levantamentos geoquímicos, complementando a análise geoestatística etc.

## REFERÊNCIAS

- AGUIAR, M. S.; COSTA, A. C. R. Autômatos Celulares para Análise da Monotonicidade da Declividade de Áreas Geológicas. In: WORKSHOP BRASILEIRO DE GEOINFORMÁTICA, 3., 2001, Rio de Janeiro. **Anais...** Porto Alegre: Sociedade Brasileira de Computação, 2001. p.87–94.
- AGUIAR, M. S.; COSTA, A. C. R.; DIMURO, G. P. ICTM: an interval tessellation-based model for reliable topographic segmentation. **Numerical Algorithms**, Dordrecht, v.37, n.1-4, p.3–11, 2004.
- AGUIAR, M. S.; DIMURO, G. P.; COSTA, A. C. R.; FINKELSTEIN, A.; KREINOVICH, V. Separating Components in Interval-Valued Images. **Reliable Computing**, Dordrecht, v.10, p.1–15, 2003.
- ANANTHANARAYANA, V.; MURTY, M.; SUBRAMANIAN, D. Efficient clustering of large data sets. **Pattern Recognition**, [S.l.], v.34, p.2561, December 2001.
- BACKER, F. B.; HUBERT, L. J. A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering. **Journal of the American Statistical Association**, [S.l.], n.71, p.870–878, 1976.
- BAEZA-YATES, R. A. Introduction to data structures and algorithms related to information retrieval. In: FRANKS, W. B.; BAEZA-YATES, R. A. (Ed.). **Information Retrieval: data structures and algorithms**. Upper Saddle River, NJ: Prentice-Hall, Inc., 1992. p.13–27.
- BERKHIN, P. **Survey Of Clustering Data Mining Techniques**. San Jose, CA: Accrue Software, 2002.
- BOLSTAD, P. V.; STOWE, T. An evaluation of DEM accuracy: elevation, slope and aspect. **Photogrammetric Engineering and Remote Sensing**, [S.l.], v.60, p.1327–1332, 1994.
- BURROUGH, P. A.; MCDONNELL, R. A. **Principles of Geographical Information Systems**. Oxford, UK: Oxford University Press, 1998.
- BURROUGH, P. **Principles of geography information systems for land resources assessment**. Oxford: Clarendon Press, 1989.
- CAMARA, G. Anatomia de sistemas de informações geográficas: visão atual e perspectivas de evolução. In: ASSAD, E.; SANO, E. (Ed.). **Sistema de informações geográficas: aplicações na agricultura**. Brasília: EMBRAPA, 1993.

CARPENTER, G. A.; GROSSBERG, S. ART-3 Hierarchical Search: chemical transmitters in self-organizing pattern recognition architectures. In: IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, IJCNN, 4., 1990, Washington. **Proceedings...** [S.l.]: IEEE (Erlbaum Assoc. Publishers), 1990. v.2, p.30–33.

COBLENTZ, D.; KREINOVICH, V.; PENN, B.; STARKS, S. Towards Reliable Sub-Division of Geological Areas: interval approach. In: INTERNATIONAL MEETING OF THE NORTH AMERICAN FUZZY INFORMATION PROCESSING SOCIETY, 19., 2000, Atlanta. **Proceedings...** [S.l.: s.n.], 2000. p.368–372.

COBLENTZ, D.; KREINOVICH, V.; PENN, B.; STARKS, S. Towards Reliable Sub-Division of Geological Areas: interval approach. In: REZNIK, L.; KREINOVICH, V. (Ed.). **Soft Computing in Measurements and Information Acquisition**. Berlin-Heidelberg: Springer-Verlag, 2003. p.223–233.

COSTA, F. A. **Introdução ao Sensoriamento Remoto**. Palestras do Projeto FMC, Pelotas, mar. 2003. Disponível em <<http://gmc.ucpel.tche.br/fmc2/papers.htm>>. Acesso em: 22 dez. 2005.

CULLINGFORD, R. E.; KRUEGER, M. W.; SELFRIDGE, M.; BIENKOWSKI, M. A. Automated Construction of Classifications: conceptual clustering versus numerical taxonomy. **IEEE Transactions on Systems, Man and Cybernetics**, [S.l.], v.12, n.2, p.168–181, 1982.

DIDAY, E.; SIMON, J. C. Clustering Analysis. In: FU, K. S. (Ed.). **Digital Pattern Recognition**. [S.l.]: Springer Verlag, 1976. p.47–94.

DRAKE, A. W. **Fundamentals of Applied Probability Theory**. New York: McGraw-Hill, 1967.

DUBES, R. C. How many cluster are best?—an experiment. **Pattern Recognition**, [S.l.], v.20, n.6, p.645–663, 1987.

DUBES, R. C. Cluster analysis and related issues. In: CHEN, C. H.; PAU, L. F.; WANG, P. S. P. (Ed.). **Handbook of Pattern Recognition & Computer Vision**. River Edge, NJ: World Cientific Publishing Co., 1993. p.3–32.

DUDA, R. O.; HART, P. E. **Pattern Classification and Scene Analysis**. New York: John Wiley & Sons, 1974.

FOGEL, D. B. An Introduction to Simulated Evolutionary Optimization. **IEEE Transactions on Neural Networks**, [S.l.], v.5, n.1, p.3–14, January 1994.

FOGEL, L. J.; OWENS, A. J.; WALSH, M. J. **Artificial Intelligence through Simulated Evolution**. New York: John Wiley & Sons, 1966.

FOLEY, J. D.; DAM, A. van; FREINER, S. K.; HUGHES, J. F. **Computer Graphics: principles and practice**. Reading, Mass.: Addison-Wesley, 1990.

FU, K. S.; LU, S. Y. A Clustering Procedure for Syntactic Patterns. **IEEE Transactions on Systems, Man, and Cybernetics**, [S.l.], v.7, p.734–742, 1977.

- GAGNON, P.; BÉDARD, Y. From Surveying to Geomatics – Evaluation of education needs to adapt to a new paradigm (A Canadian Perspective). **Geomatica**, [S.l.], v.50, n.3, p.269–286, 1996.
- GLOVER, F. Future Paths for Integer Programming and Links to Artificial Intelligence. **Computers and Operations Research**, [S.l.], v.13, p.533–549, 1986.
- GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization & Machine Learning**. Reading: Addison-Wesley, 1989. 412p.
- GOWDA, K. C.; DIDAY, E. symbolic clustering using a new dissimilarity measure. **IEEE Trans. Syst. Man Cybern.**, [S.l.], n.22, p.368–378, 1992.
- GOWER, J. C.; ROSS, G. J. S. Minimum Spanning Trees and Single Linkage Cluster Analysis. **Applied Statistics**, [S.l.], v.18, n.1, p.54–64, 1969.
- GRAYSON, R. B.; BLOSCHL, G.; MOORE, I. D. Distributed parameter hydrologic modeling using vector elevation data: thales and tapes-c. In: SINGH, V. P. (Ed.). **Computer Models of Watershed Hydrology**. Boca Raton, FL: CRC, 1994. p.669–696.
- HOLLAND, J. H. **Adaptation in natural artificial systems**. Ann Arbor: University of Michigan Press, 1975.
- HUTCHINSON, M. F.; NIX, H. A.; MCMAHON, J. P.; ORD, K. D. The development of a topographic and climate database of Africa. In: INTERNATIONAL CONFERENCE ON INTEGRATING GIS AND ENVIRONMENTAL MODELLING, 3., 1996, Santa Fe. **Proceedings...** [S.l.: s.n.], 1996.
- JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Englewood Cliffs, NJ: Prentice-Hall Inc., 1988. xiv, 320p.
- JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, [S.l.], v.31, n.3, p.264–323, 1999.
- KING, B. Step-wise clustering procedures. **Journal of the American Statistical Association**, [S.l.], n.69, p.86–101, 1967.
- KIRKBY, M. J.; IMESON, A. C.; BERGKAMP, G.; CAMMERAAT, L. H. Scaling up processes and models. **Journal of Soil and Water Conservation**, [S.l.], v.51, n.5, p.391–396, September-October 1996.
- KIRKPARTICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. **Science**, [S.l.], v.220, n.4598, p.671–680, 1983.
- KNUTH, D. **The Art of Computer Programming**. Reading, MA: Addison-Wesley, 1973.
- KOHONEN, T. **Self-organization and Associative Memory**. New York: Springer-Verlag, 1989. (Springer information sciences series).
- KORTE, G. **The GIS book**. 3.ed. Santa Fe: On World Press, 1994.
- KRAUS, K. Visualization of the quality of surfaces and their derivatives. **Photogrammetric Engineering and Remote Sensing**, [S.l.], v.60, p.457–462, 1994.

LU, S. Y.; FU, K. S. A sentence-to-sentence clustering procedure for pattern analysis. **IEEETSMC: IEEE Transactions on Systems, Man, and Cybernetics**, [S.l.], v.8, p.381–389, 1978.

MACKEY, B. The Role of GIS and Environmental Modeling in the Conservation of Biodiversity. In: INTERNATIONAL CONFERENCE ON INTEGRATING GIS AND ENVIRONMENTAL MODELLING, 3., 1996, Santa Fe. **Proceedings...** [S.l.: s.n.], 1996.

MAO, J.; JAIN, A. K. A Self Organizing Network for Hyperellipsoidal Clustering (HEC). In: IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, ICNN, 5., 1994, Orlando, FL. **Proceedings...** [S.l.]: IEEE, 1994. p.2967–2972.

McQueen, J. B. Some Methods of Classification and Analysis of Multivariate Observations. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967. **Proceedings...** [S.l.: s.n.], 1967. p.281–297.

MENEGUETTE, A. **Introdução ao Geoprocessamento**. Disponível em <[http://www2.prudente.unesp.br/dcartog/arlete/hp\\_arlete/courseware/intgeo.htm](http://www2.prudente.unesp.br/dcartog/arlete/hp_arlete/courseware/intgeo.htm)>. Acesso em: 22 dez. 2005.

MOORE, I. D.; GRAYSON, R. B.; LADSON, A. R. Digital terrain modeling: a review of hydrological, geomorphological and biological applications. **Hydrological Processes**, [S.l.], v.5, p.3–30, 1991.

MOORE, I. D.; HUTCHINSON, M. F. Spatial extension of hydrologic process modelling. In: INTERNATIONAL HYDROLOGY AND WATER RESOURCES SYMPOSIUM, 1991, Canberra. **Proceedings...** [S.l.]: Institute of Australian Engineers, 1991. p.803–808.

MOORE, I. D.; LEWIS, A. D.; GALLANT, J. C. Terrain attributes: estimation methods and scale effects. In: JAKEMAN, A. J.; BECK, M. B.; MCALEER, M. J. (Ed.). **Modeling Change in Environmental Systems**. New York: Wiley, 1993. p.189–214.

MOORE, R. E. **Methods and Applications of Interval Analysis**. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1979. 201p.

MURTAGH, F. A survey of recent advances in hierarchical clustering algorithms. **The Computer Journal**, [S.l.], v.26, n.4, p.354–359, Nov. 1983.

NAGY, G. State of the art in pattern recognition. **Proc. of IEEE (PIRE)**, [S.l.], v.46, n.5, p.836–862, 1968.

OZAWA, K. A stratificational overlapping cluster scheme. **Pattern Recognition**, [S.l.], v.18, p.279–286, 1985.

PHILLIPS, J. D. Sediment storage, sediment yield, and time scales in sediment denudation studies. **Geographical Analysis**, [S.l.], v.18, p.161–167, 1986.

POLIDORI, L.; CHOROWICZ, J.; GUILLANDI, R. Description of terrain as a fractal surface and application to digital elevation. **Photogrammetric Engineering and Remote Sensing**, [S.l.], v.57, p.1329–1332, 1991.

- PREPARATA, F. P.; SHAMOS, M. I. **Computational Geometry**: an introduction. [S.l.]: Springer-Verlag, 1985.
- RODRIGUES, M. Geoprocessamento. In: ENCONTRO NACIONAL DE ENGENHEIROS CARTÓGRAFOS, 5., 1988. **Anais...** Presidente Prudente: UNESP, 1988. p.144–160.
- RODRIGUES, M. Introdução ao Geoprocessamento. In: SIMPÓSIO BRASILEIRO DE GEOPROCESSAMENTO, 1., 1990. **Anais...** São Paulo: EDUSP, 1990. p.1–26.
- RODRIGUES, M.; QUINTANILHA, J. A. A seleção de software SIG para gestão urbana. In: CONGRESSO BRASILEIRO DE CARTOGRAFIA, 15., 1991. **Anais...** São Paulo: SBC, 1991. p.513–519.
- ROSE, K.; GUREWITZ, E.; FOX, G. A Deterministic Annealing Approach to Constrained Clustering. In: IEEE INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY (ISIT-91), 1991. **Proceedings...** [S.l.: s.n.], 1991.
- RUSPINI, E. H. A New Approach to Clustering. **Information and Control**, [S.l.], v.15, n.1, p.22–32, July 1969.
- SALTON, G. Developments in automatic text retrieval. **Science**, [S.l.], v.253, p.974–980, Aug. 1991.
- SCHAFFER, W. M. Ecological abstraction: the consequences of reduced dimensionality in ecological models. **Ecological Monographs**, [S.l.], v.5, p.383–401, 1981.
- SILVA, J.; SOUZA, M. **Análise ambiental**. Rio de Janeiro: Ed. da UFRJ, 1987.
- SILVA, J. X. d. **Geoprocessamento para análise ambiental**. Rio de Janeiro: Sermograf, 2001. 228p.
- SILVA, R. K. S.; AGUIAR, M. S.; DIMURO, G. P.; COSTA, F.; COSTA, A. C. R. Topo-ICTM: uma ferramenta para categorização topográfica baseada no modelo de tesselação intervalar bidimensional. In: CONFERENCIA LATINOAMERICANA DE INFORMÁTICA, CLEI, 29., 2003, La Paz. **Artículos**. La Paz: Universidad Mayor de San Andrés, 2003. p.1–10.
- SMITH, D.; LEWIS, M. Optimal predictors for compression of digital elevation models. **Computers and Geosciences**, [S.l.], v.20, p.1137–1141, 1994.
- SOKAL, R. R.; SNEATH, P. H. A. **Principles of Numerical Taxonomy**. San Francisco (CA), USA: Freeman, 1963.
- TANAKA, E. Theoretical aspects of syntactic pattern recognition. **Pattern Recognition**, [S.l.], v.28, p.1053–1061, July 1995.
- TEIXEIRA, A. L.; MORETTI, E.; CHRISTOFOLETTI, A. **Introdução aos sistemas de informação geográfica**. Rio Claro: Ed. do Autor, 1992.
- VON NEUMANN, J.; BURKS, A. W. **Theory of Self-Reproducing Automata**. [S.l.: s.n.], 1966. 407p.

WARD, J. H. J. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, [S.l.], n.58, p.236–244, 1963.

WILSON, J. P. GIS-based land surface/subsurface models: new potential for new models. In: INTERNATIONAL CONFERENCE ON INTEGRATING GIS AND ENVIRONMENTAL MODELLING, 3., 1996, Santa Fe. **Proceedings...** [S.l.: s.n.], 1996.

WILSON, J. P.; BURROUGH, P. A. Dynamic modelling, geostatics and fuzzy classification: new sneakers for a new geography? **Association of American Geographers**, [S.l.], v.89, p.736–746, 1999.

WILSON, J. P.; GALLANT, J. **Terrain analysis** : principles and applications. New York: Wiley & Sons, 2000.

ZADEH, L. A. Fuzzy Sets. **Information and Control**, [S.l.], v.8, n.3, p.338–353, 1965.

ZAHN, C. T. Graph-theoretical methods for detecting and describing gestalt clusters. **IEEE Transactions on Computers**, [S.l.], v.20, n.1, p.68–86, 1971.

ZHANG, K. Algorithms for the constrained editing distance between ordered labeled trees and related problems. **Pattern Recognition**, [S.l.], v.28, p.463–474, March 1995.

## **APÊNDICE A EXEMPLOS DE CATEGORIZAÇÕES**



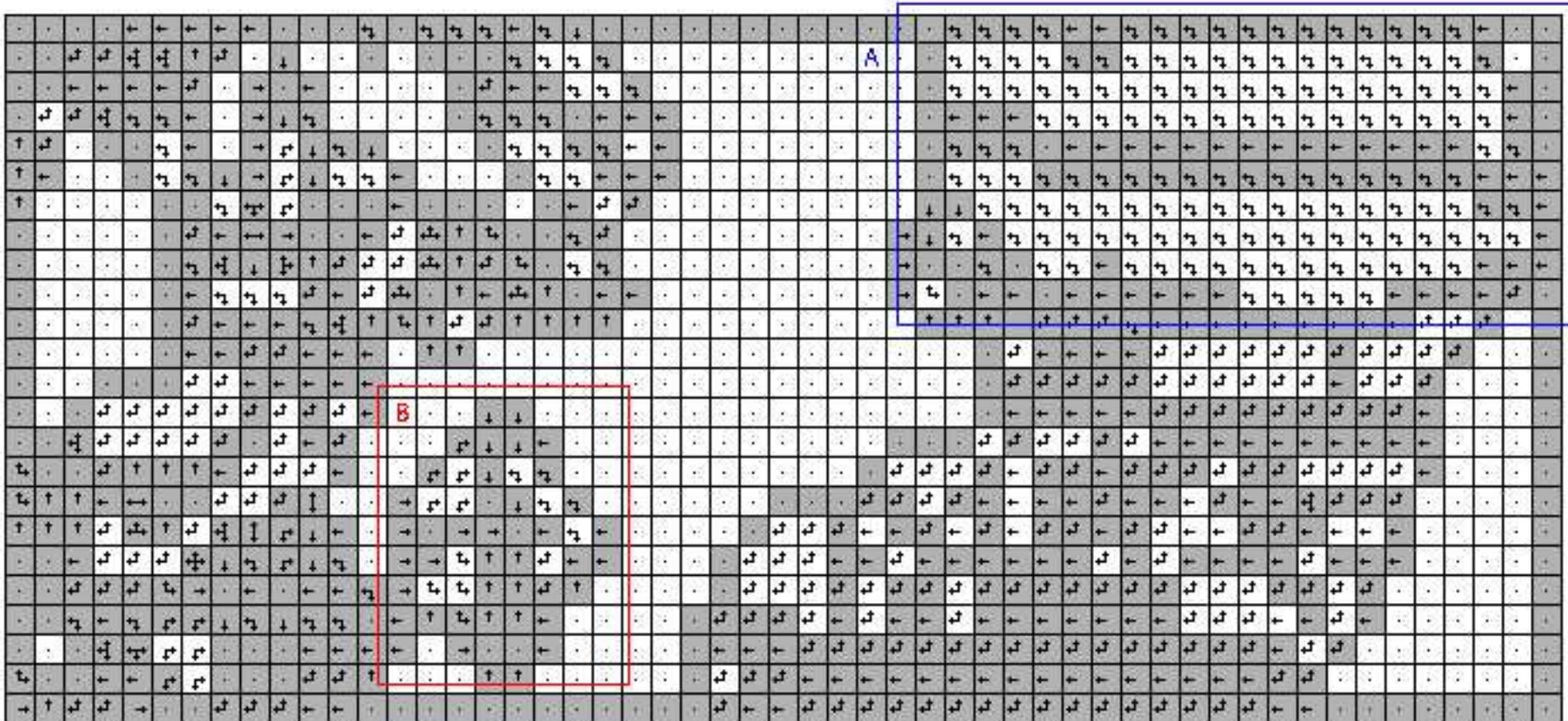


Figura A.1: DEM1000m-raio 1

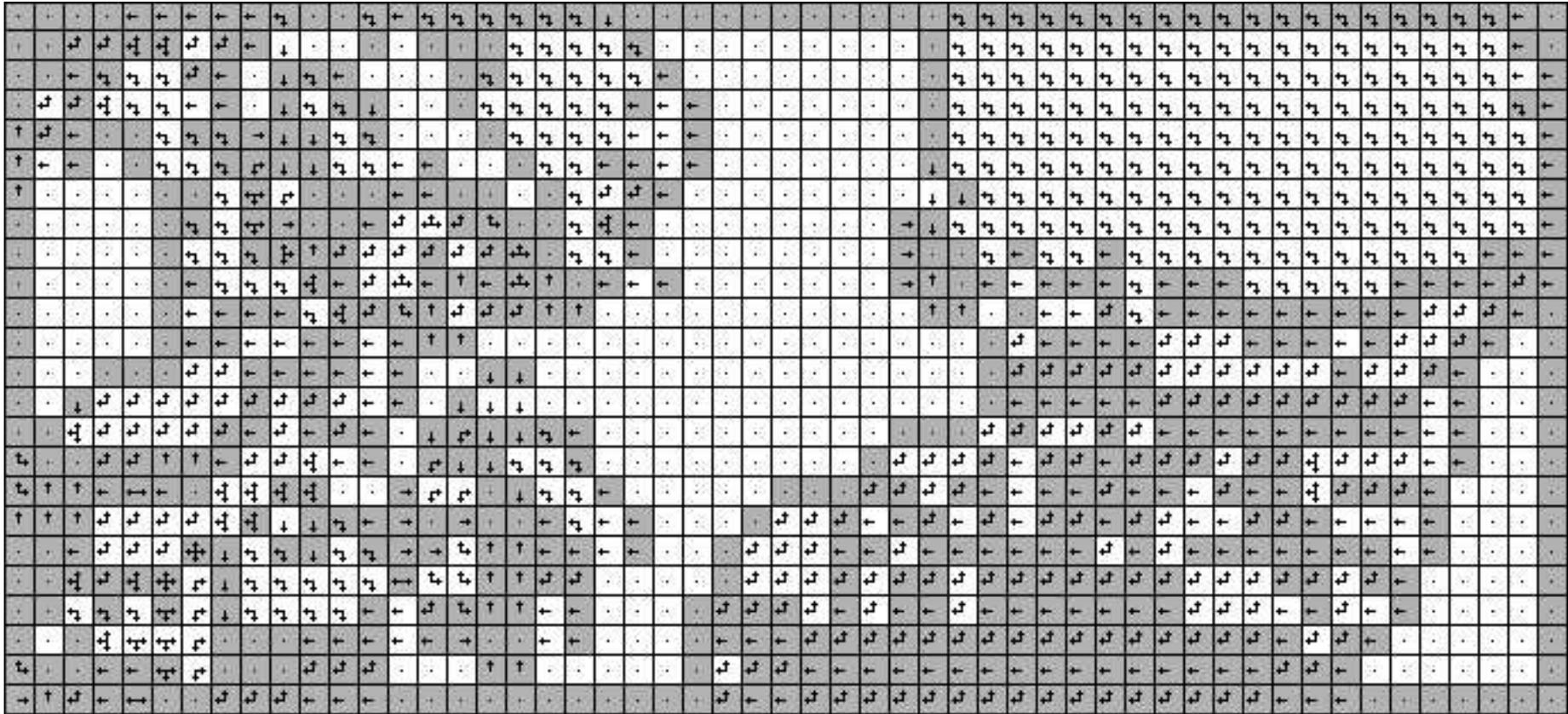


Figura A.2: DEM1000m-raio 2

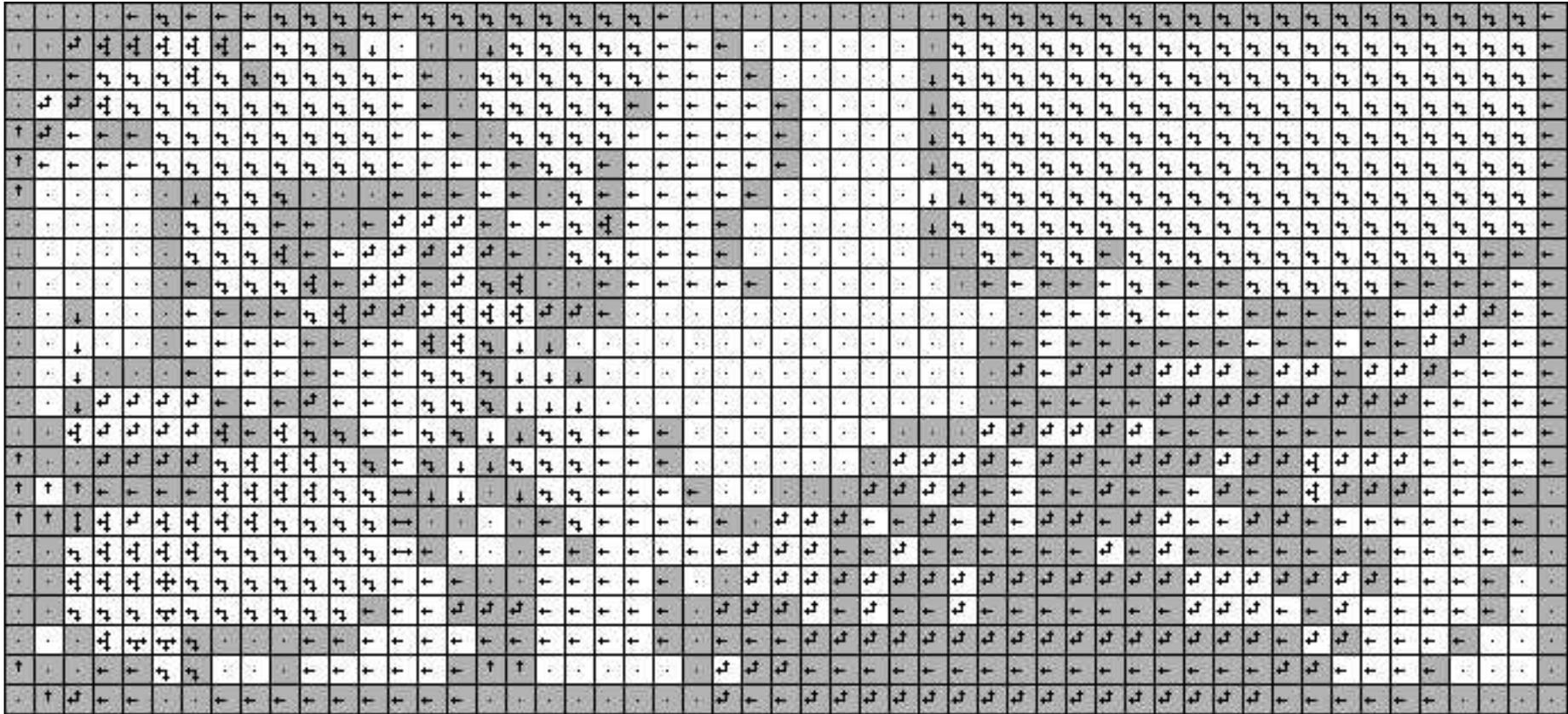


Figura A.3: DEM1000m-raio 5

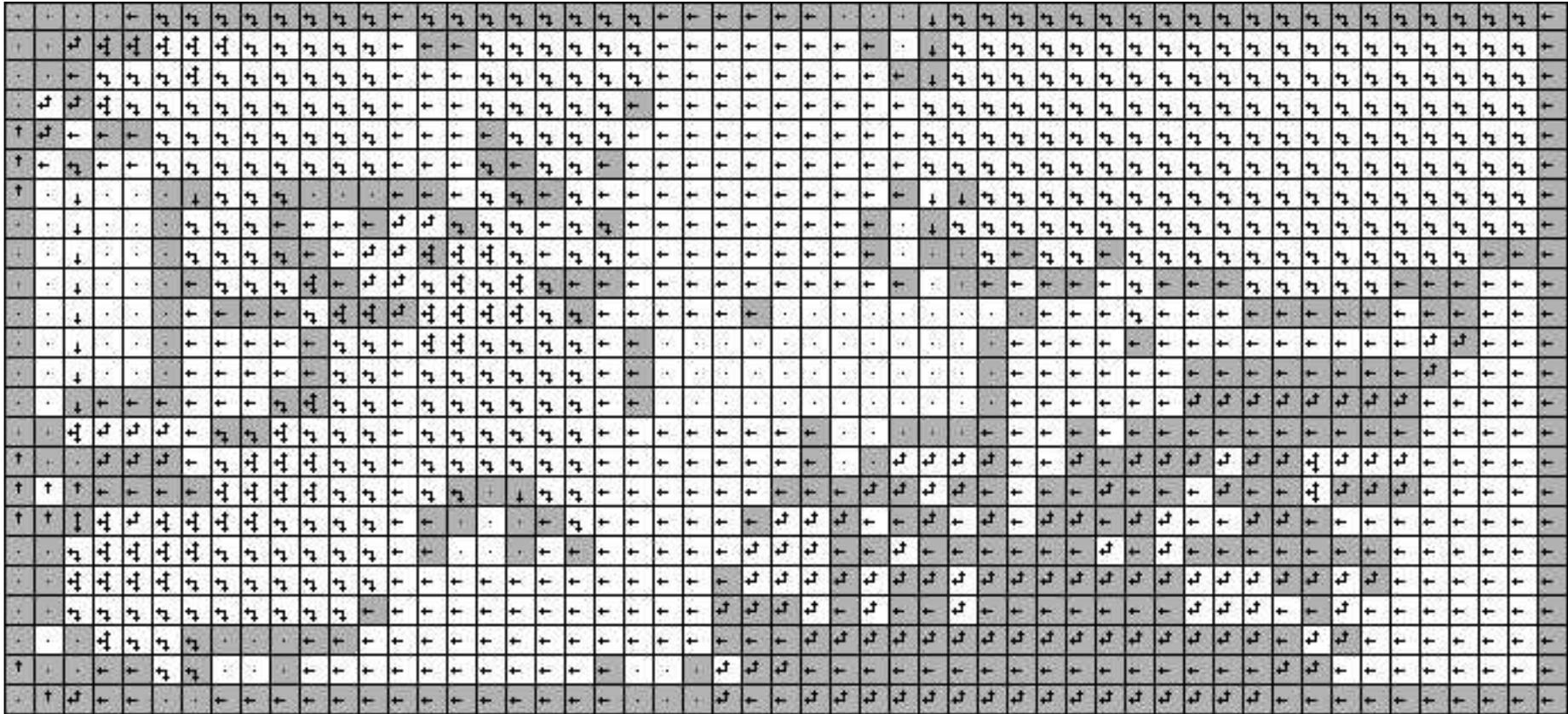


Figura A.4: DEM1000m-raio 10

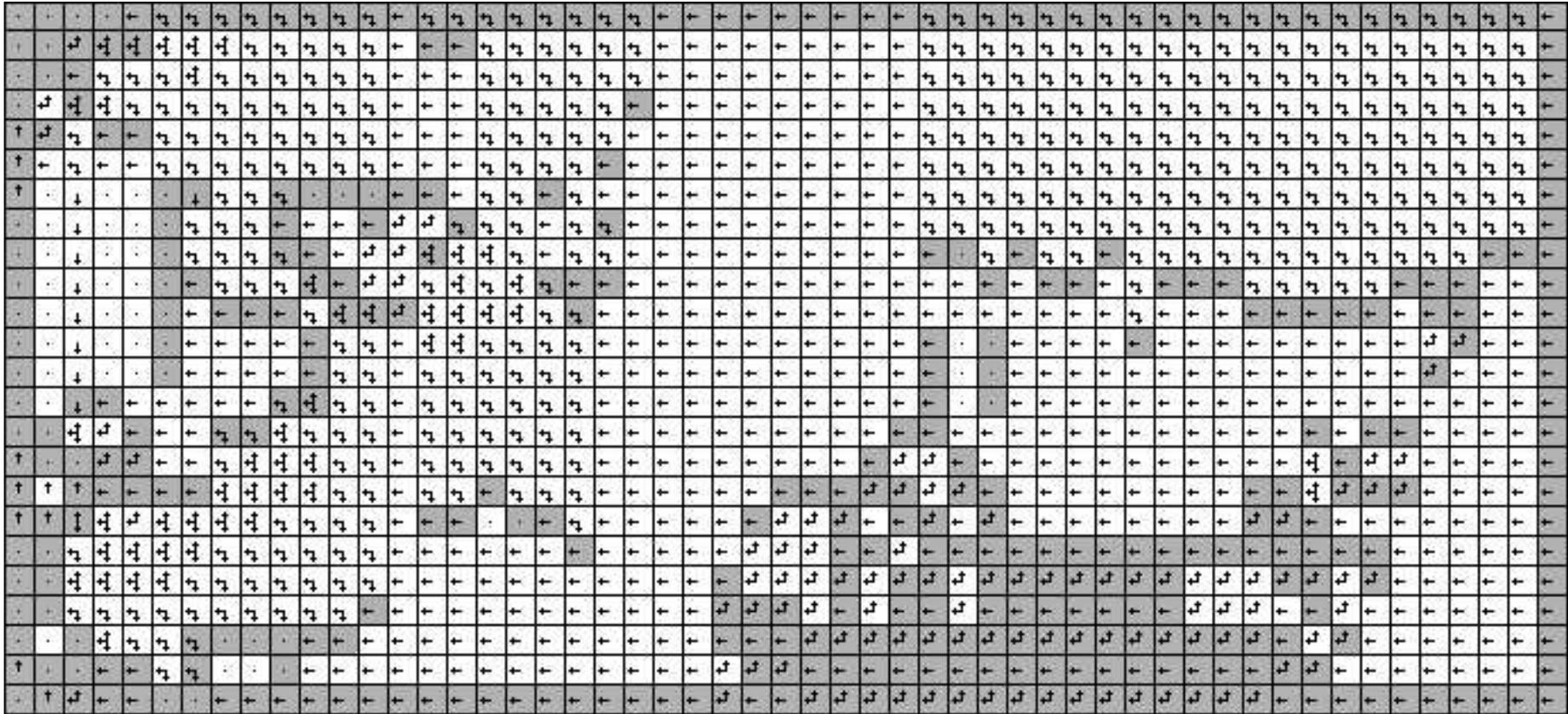


Figura A.5: DEM1000m-raio 20

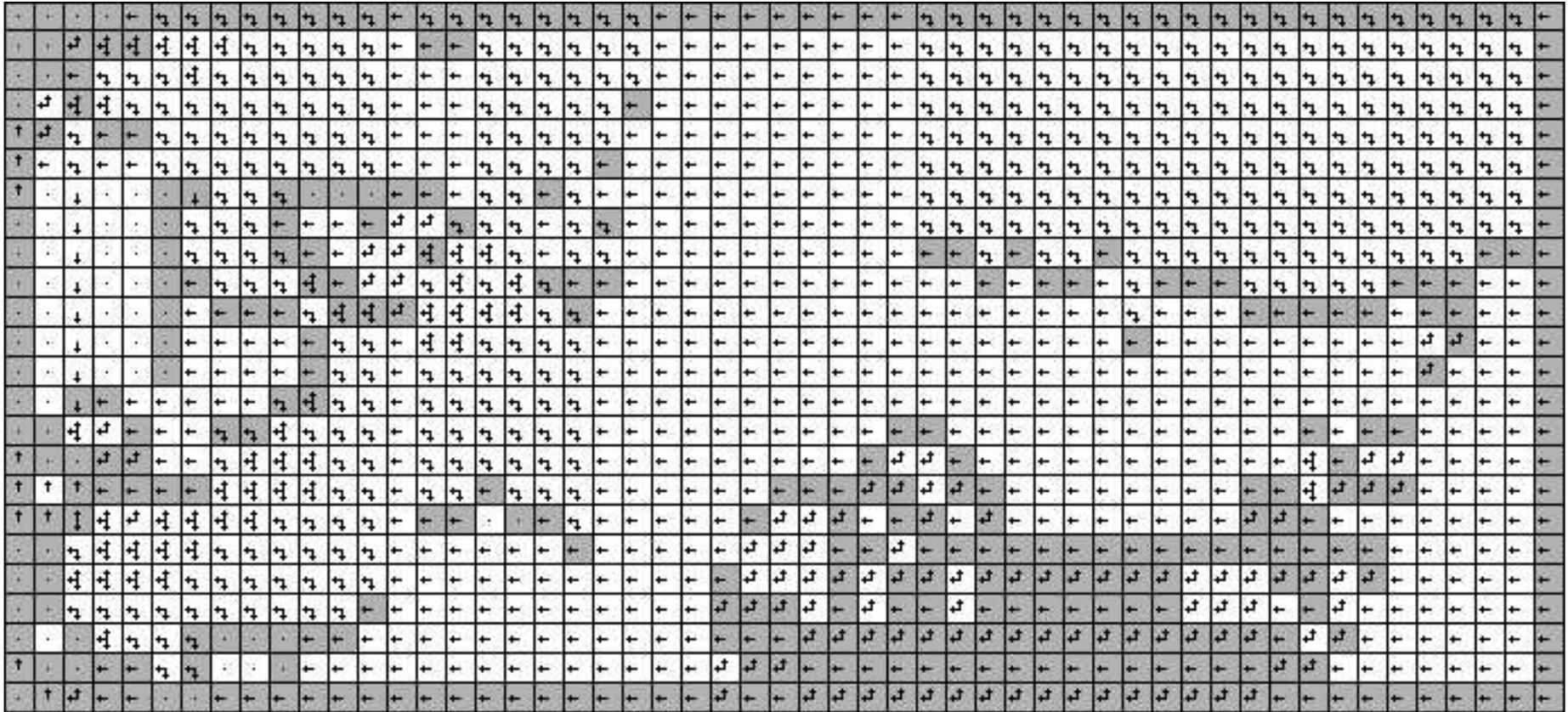


Figura A.6: DEM1000m-raio 40





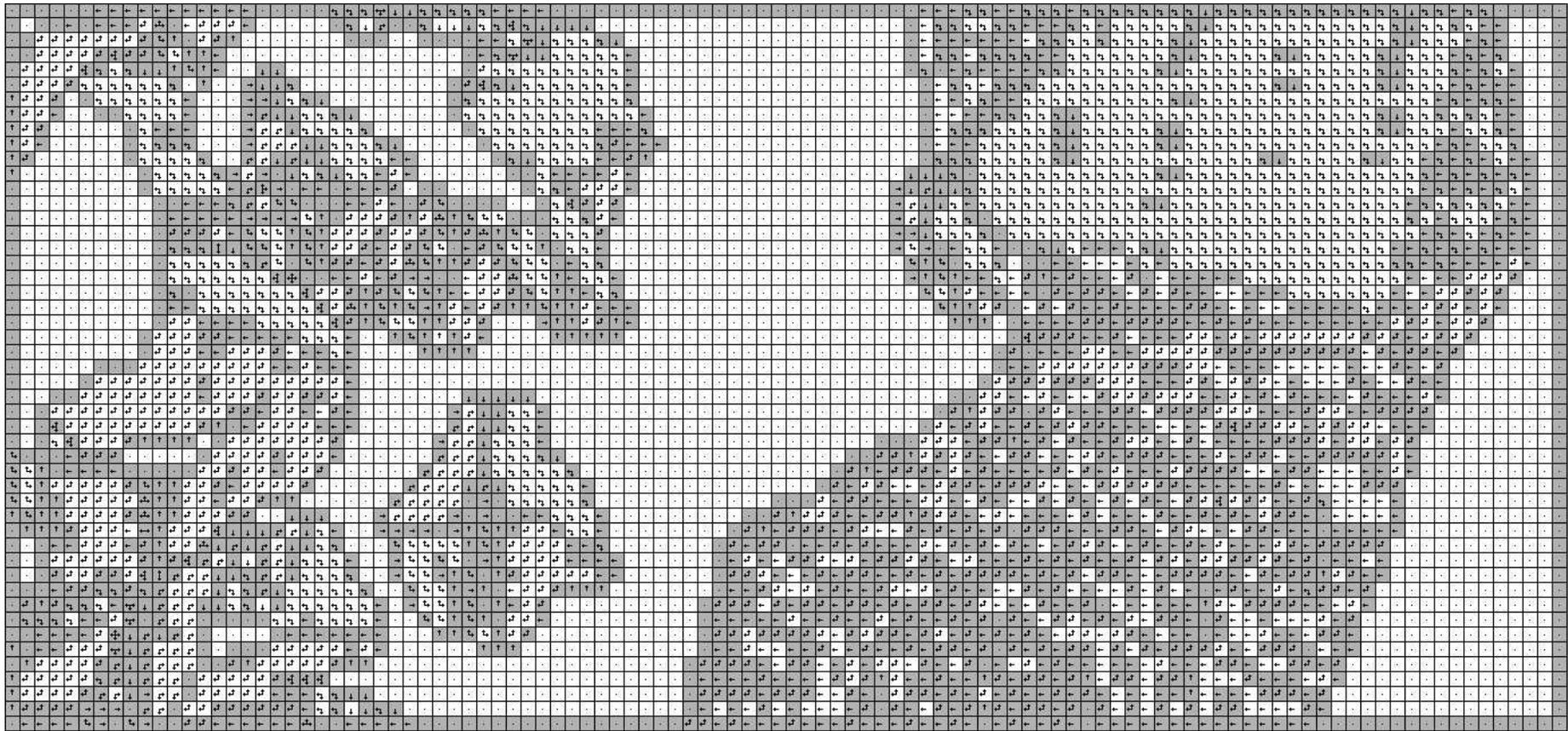


Figura A.7: DEM500m-raio 1

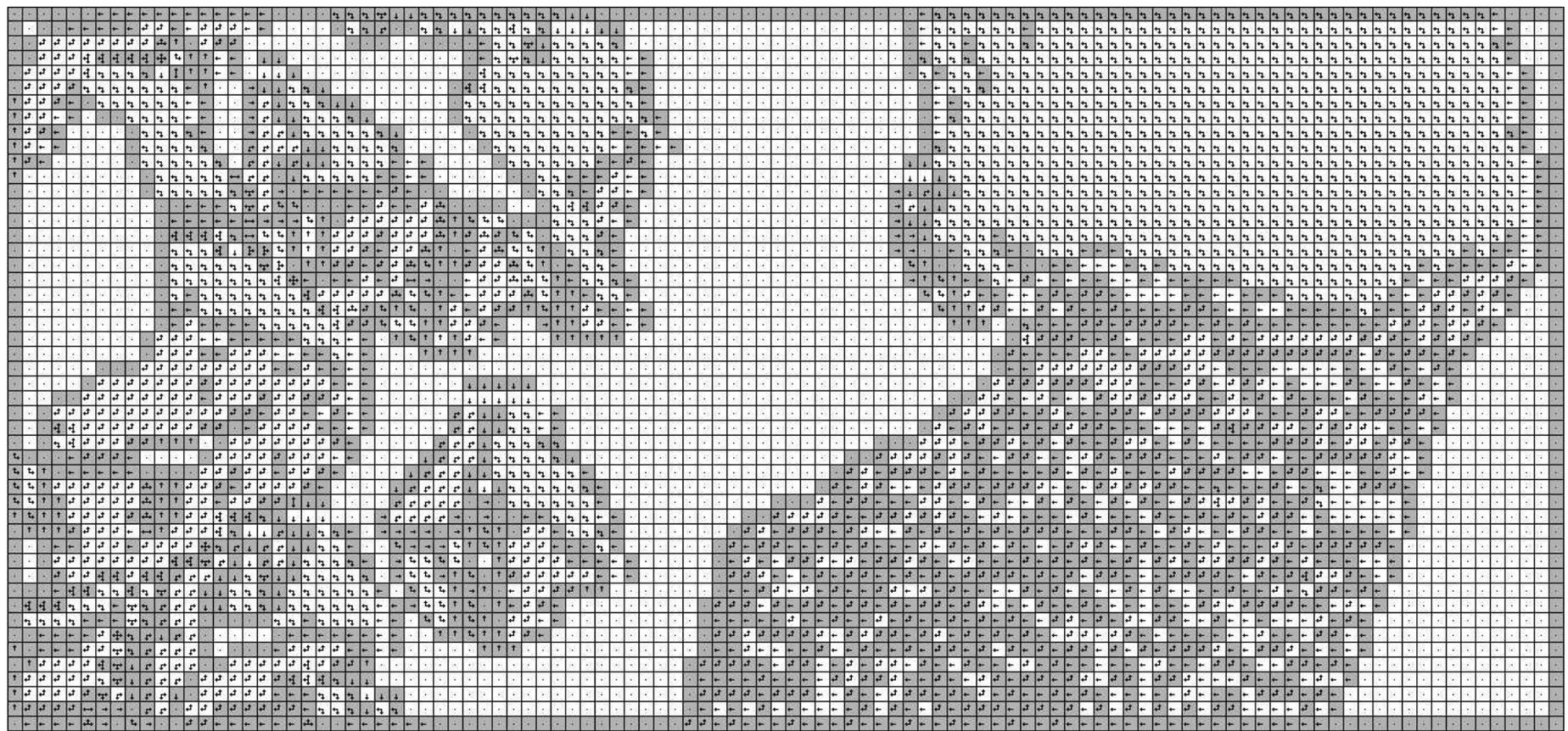


Figura A.8: DEM500m-raio 2

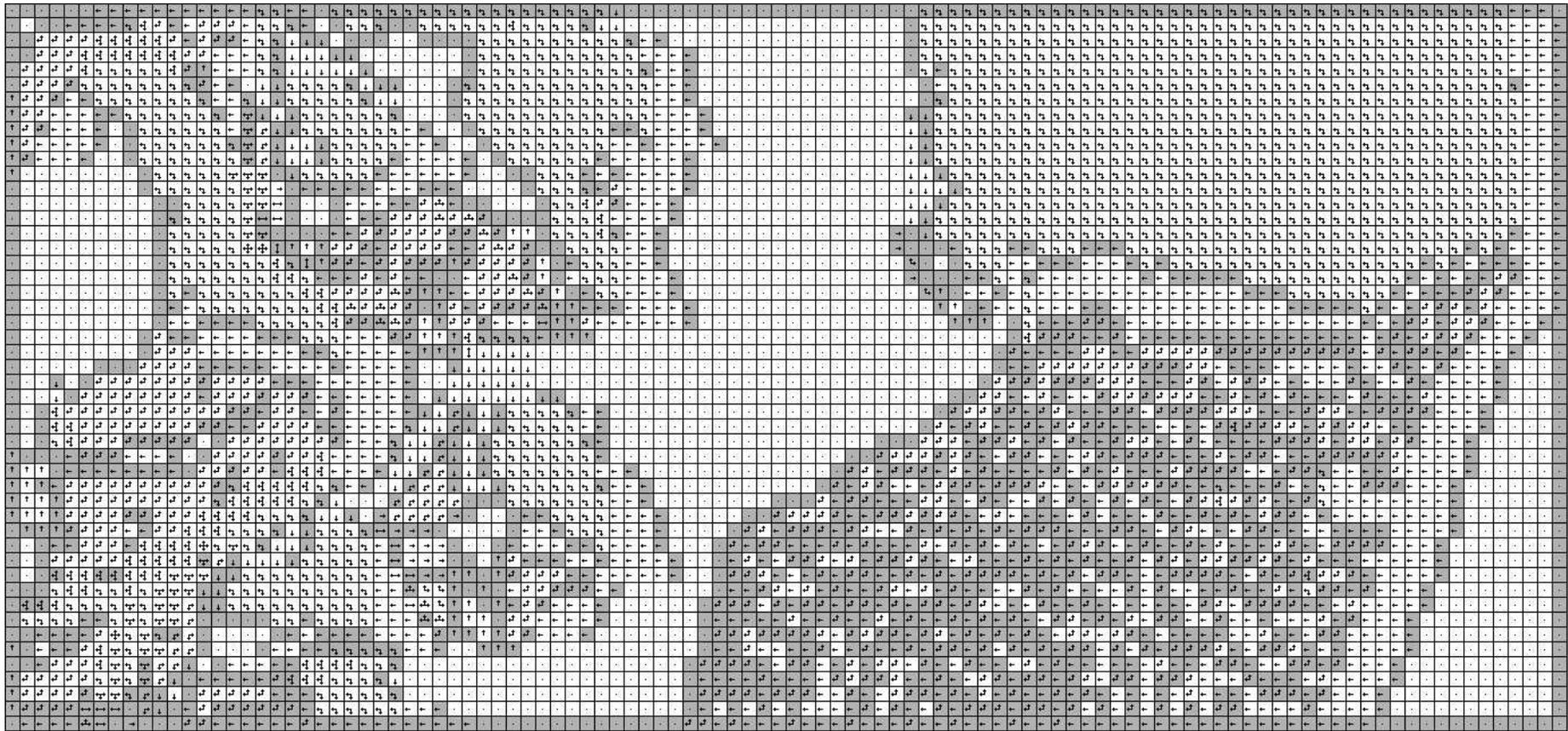


Figura A.9: DEM500m-raio 5



Figura A.10: DEM500m-raio 10



Figura A.11: DEM500m-raio 20



Figura A.12: DEM500m-raio 40