MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

PROCEDIMENTOS EFICIENTES PARA REGIONALIZAÇÃO DE UNIDADES SOCIOECONÔMICAS EM BANCOS DE DADOS GEOGRÁFICOS

Marcos Corrêa Neves

Tese de Doutorado em Sensoriamento Remoto, orientada pelo Prof. Dr. Gilberto Câmara e pela Profa. Dra. Corina da Costa Freitas, aprovada em 19 de dezembro de 2003.

INPE
São José dos Campos
2003

CAPÍTULO 1
INTRODUÇÃO

1.1 - MOTIVAÇÃO

Nas últimas décadas houve um crescimento acentuado da população humana, chegando a triplicar no decorrer dos últimos setenta anos[1]. O crescimento populacional, e a sua conseqüente pressão sobre os recursos naturais, combinado com a escassez de recursos financeiros têm demandado por estudos e planejamento cada vez mais complexos, que contemplem diversos fatores ambientais e sociais.

Em estudos de diferentes campos do conhecimento, tais como geografia, geologia, estatística, economia, sociologia, epidemiologia, biologia e meio ambiente, a consideração da posição espacial dos objetos, observações ou eventos é essencial para uma adequada representação e compreensão de muitos fenômenos (Bailey & Gatrell, 1995; Goodchild et al., 1996).

Fischer et al. (1996) destacam alguns importantes fatores para se adotar a perspectiva espacial nas análises de fenômenos: i) O espaço fornece uma simples, mas útil, estrutura para a manipulação de grandes volumes de dados. ii) A perspectiva espacial permite o acesso fácil para informações relacionadas à localização de objetos e eventos. iii) Ela permite aos objetos e eventos de vários tipos serem associados, dentro de uma formalização dos processos em um sistema de informação geográfica. iv) Tanto em aplicações ambientais como sociais, a distância entre os objetos e eventos são freqüentemente um fator importante na determinação da interação entres eles.

---

[1] Demorou milhares de anos para que a população humana atingisse a dois bilhões de pessoas, o que ocorreu por volta de 1930. Porém, quarenta e seis anos depois, a população dobrou, atingindo em 1996 à marca de 5,7 bilhões de pessoas (Marsh & Grossa Jr., 1996).

Se por um lado, existe um aumento da demanda por análises mais elaboradas e complexas, envolvendo a perspectiva espacial, também temos o aumento da disponibilidade de informação espacial. A disseminação do uso de meios eletrônicos na sociedade moderna tem gerado um grande volume de dados, em geral. Sistemas gerenciadores de banco de dados estão presentes na maioria das organizações públicas e empresas de médio e grande porte, contendo os mais diferentes dados. No Brasil, institutos como o IBGE (*Instituto Brasileiro de Geografia e Estatística*), disponibilizam na *internet* grande quantidade de dados (censos demográficos, dados econômicos, produção agropecuária, etc.), que podem estar associados a diferentes unidades territoriais, como setores censitários, municípios, microrregião, mesorregião e estados. Outras instituições também optaram por disponibilizar dados em meios eletrônicos, como o Ministério de Saúde, tornando acessíveis diversos dados relacionados com saúde (gastos públicos, internamentos hospitalares, óbitos, entre outros).

Um outro fator a ser considerado, com relação ao aumento da disponibilidade de dados espaciais, é a popularização do uso de *Sistemas de Informação Geográfica* (SIG). Novas informações espaciais derivadas dos mais diferentes estudos e em diversas áreas do conhecimento são geradas, aumentando à possibilidade de re-utilização de dados geográficos. Para ilustrar a disseminação do uso de SIG, Burrough & McDonnel (1998) apresentam uma estimativa para o número de instalações de SIG no mundo. Eles avaliaram que, no ano de 1995, já existiam cerca de 93.000 sistemas, empregados em diversas áreas.

Avanços importantes também ocorrem na aquisição automática de dados como em sistemas de sensoriamento remoto, que geram uma grande quantidade de dados. A cada ano, novos sistemas orbitais são lançados aumentando o número de sensores e dados, disponíveis em diferentes resoluções espaciais, temporais e espectrais.

Todos os exemplos citados demonstram o incremento significativo na oferta de dados referenciados espacialmente, gerados por diferentes fontes, constituindo uma enorme massa de dados. Este volume de informação representa, potencialmente, uma grande fonte de informação para diversos estudos, mas que necessitam de ferramentas adequadas para explorá-las e transformá-las em informação efetivamente utilizável.

Os SIG são sistemas adequados ao tratamento de grandes volumes de dados espaciais. Eles se desenvolveram a partir dos anos 70, inicialmente em aplicações específicas em diversas áreas, até o surgimento de SIG de propósito geral na década de 90. Uma das possíveis visões para SIG é representada pela seguinte definição: poderoso conjunto de ferramentas para coletar, armazenar, recuperar, transformar e apresentar dados espaciais do mundo real (Burrough & McDonnel, 1998). Independente das diversas formas de se entender os SIG, é importante ressaltar duas características destes sistemas: i) integrar informações espaciais de diferentes fontes (mapas temáticos, imagens de sensoriamento remoto, modelos numéricos de terreno, dados censitários, etc.) em uma única base de dados; ii) oferecer formas de manipular, combinar e analisar o conteúdo da base de dados (Câmara & Medeiros, 1998).

A necessidade de métodos para auxiliar a analise de dados espaciais é anterior ao surgimento dos SIGs. A *Análise Espacial* (*AE*) teve seu desenvolvimento associado ao domínio da geografia (*Geografia Quantitativa*) (Fotheringham et al., 2000) a partir da década de 1950. Em seu estágio inicial, ela utilizava técnicas e procedimentos quantitativos para analisar pontos, linhas, áreas e superfície em mapas ou definidos por coordenadas (Fischer & Getis, 1997). Uma definição para AE é apresentada em Haining (1995): AE pode ser definida como uma coleção de técnicas para a análise de eventos geográficos onde o resultado da análise depende no arranjo espacial dos eventos.

A potencialidade da associação de técnicas de análise espacial e SIG foi percebida, mais claramente, durante o final da década de 80 e algumas propostas de integração começaram a aparecer no início da década de 90. A integração de SIG e AE está baseada na capacidade dos SIG em manipular e visualizar rapidamente dados digitais, fornecendo uma nova estrutura para o desenvolvimento da AE e, por sua vez, a AE abre um campo para a incorporação de novas funcionalidades aos SIGs (Longley & Batty, 1996). Fischer et al. (1996) aponta que, em meados da década de 90, já havia se estabelecido um senso comum na comunidade científica, ligada tanto à AE quanto aos SIG, que o futuro da tecnologia de SIG estaria na incorporação de funcionalidades com maior poder analítico e de modelagem.

Além da incorporação de algumas novas funcionalidades aos SIGs atuais, o resultado do esforço pela integração entre SIG e AE pode ser verificado no desenvolvimento de alguns pacotes específicos de software como o *SpaceStat*[2] (Anselin, 1995) e *SAGE*[3] (*Spatial Analysis in a GIS Environment*) (Ma et al., 1997).

Este trabalho propõe um método para auxiliar a análise de dados espaciais com representação geográfica poligonal, através do uso de um procedimento de classificação dos objetos espaciais. A classificação, por Análise de Cluster, é uma técnica utilizada para separar objetos ou eventos em grupos naturais. Ela serve para descobrir estruturas existentes em um conjunto complexo de dados e para simplificar a análise. Portanto, a classificação também pode ser utilizada para o desenvolvimento de ferramentas para auxiliar a análise de dados espaciais.

Wise et al. (1997) propõe utilizar o procedimento de regionalização como ferramenta exploratória de análise espacial aplicada a dados espaciais com representação poligonal. A regionalização é, basicamente, um procedimento de classificação onde as classes são formadas por objetos espaciais contíguos e semelhantes entre si.

Openshaw & Alvanides (2001) defendem a possibilidade de utilizar a regionalização como um "detector de padrões" em dados espaciais e, assim, atuar como uma ferramenta visual de análise espacial. Porém, os autores alertam para a necessidade do desenvolvimento de algoritmos práticos que possam enfrentar a complexidade computacional associada ao problema de regionalização.

A classificação, através da Análise de Cluster, também é utilizada no contexto da Mineração de Dados (*Data Mining*), e também na sua extensão, aplicada a dados espaciais (*Spatial Data Mining*). Esta área emprega métodos de diferentes áreas do conhecimento com o objetivo de extrair informação útil de grandes massas de dados (Goebel & Gruenwald, 1999; Ng & Han, 1994). Buttenfield et al. (2001) afirmam que

---

[2] Informações sobre o SpaceStat: http://www.terraseer.com/spacestat/docs/V180man.pdf

[3] Informações sobre o SAGE: ftp://ftp.shef.ac.uk/pub/uni/academic/D-H/g/sage/sagehtm/sage.htm.

as aplicações correntes em prospecção de conhecimento, voltadas a dados geográficos, geralmente utilizam representações simples para objetos geográficos e para os relacionamentos espaciais. Os métodos de mineração de dados espaciais deveriam reconhecer tipos mais complexos de dados (linhas e polígonos), uma vez que nem sempre os objetos espaciais podem ser reduzidos a um ponto, e os métodos deveriam considerar os relacionamentos espaciais entre os objetos (distância não-euclidiana, direção, conectividade). Koperski et al. (1997) diz que o crucial desafio para mineração de dados espaciais é a eficiência dos algoritmos empregados devido ao volume de dados espaciais, complexidade dos tipos de dados e aos métodos de acesso aos dados espaciais.

Com o exposto acima, verifica-se que existe uma demanda crescente pela consideração da perspectiva espacial nos estudos de diferentes fenômenos e que as tecnologias aplicáveis a dados espaciais, como os SIG ou a mineração de dados espaciais, ainda apresentam algumas deficiências. Há necessidade de algoritmos mais eficientes que possam viabilizar análise dados espaciais, sobretudo em situações que envolvam grandes volumes de informação e trabalhem no contexto espacial. Este trabalho visa contribuir, especificamente no tratamento do problema de regionalização, propondo um método mais eficiente de agrupamento de unidades de área, que possa ser utilizado como ferramenta auxiliar à análise de dados espaciais com representação poligonal.

1.2 - OBJETIVO DO TRABALHO

Desenvolver um método alternativo para o procedimento de regionalização aplicável a grandes volumes de dados espaciais de forma a permitir sua utilização como ferramenta de desenvolvimento de novas representações espaciais e de análise exploratória de dados espaciais. O trabalho propõe dois caminhos: o primeiro, segue uma abordagem mais tradicional, tratando a regionalização como um processo automático, onde o analista define um conjunto de parâmetros e o sistema produz uma classificação. Dentro desta abordagem, é proposto um método de regionalização que busca atingir um resultado de qualidade, do ponto de vista da homogeneidade entre os objetos de uma

mesma classe, com a rapidez na geração do resultado. O segundo caminho, propõe que a regionalização seja um processo interativo, com a interferência e condução do analista durante o processo de classificação.

1.3 – CONTRIBUIÇÃO DO TRABALHO

Este trabalho oferece as seguintes contribuições:

i) Revisão sobre os procedimentos de classificação e regionalização, sendo apresentadas as diferentes abordagens utilizadas para a regionalização e a descrição de métodos propostos anteriormente.

ii) Análise comparativa entre diferentes métodos de regionalização, apresentando suas características e comportamentos.

iii) Método de regionalização alternativo, aplicável a grandes volumes de dados, bem como a avaliação do seu desempenho em relação a outros procedimentos.

iv) Proposição de uma nova abordagem para a regionalização, na qual o procedimento é visto como uma ferramenta mais abrangente de análise de dados espaciais, substituindo o processo automático, utilizado nas outras abordagens, por um processo interativo.

1.4 – ESTRUTURA DO TRABALHO

O trabalho está dividido em sete capítulos. Uma revisão sobre os procedimentos de classificação, baseados em *Análise de Cluster*, e regionalização são apresentados no Capítulo 2. Conceitos e elementos presentes em um processo de classificação, bem como métodos gerais de classificação são apresentados. Além da classificação, este capítulo apresenta as diferentes abordagens para o procedimento de regionalização e descreve alguns dos principais métodos propostos na literatura.

Uma análise comparativa dos métodos de regionalização é apresentada no Capítulo 3. O ambiente de experimentação, utilizado para a implementação e a análise dos comportamentos dos métodos, também é descrito neste capítulo. Com base em

resultados obtidos em alguns testes no ambiente de experimentação, é traçada uma análise comparativa entre os métodos, a partir da qual, se estabeleceu as justificativas e as diretrizes adotadas no desenvolvimento do método alternativo para o procedimento de regionalização.

No Capítulo 4, são apresentados o procedimento proposto e as técnicas empregadas em seu desenvolvimento. Também são mostrados alguns resultados de testes, realizados para avaliar o comportamento e o desempenho do método proposto e os possíveis ganhos comparativos proporcionados por ele.

O Capítulo 5 apresenta a proposta de condução do procedimento de regionalização como um método semi-automático, com a interação direta do analista, não somente na fase inicial de sua execução (definindo os parâmetros de entrada do algoritmo), mas no desenvolvimento do processo de classificação, alterando e interferindo no resultado. São apresentados alguns requisitos para um sistema deste tipo e funcionalidades básicas implementadas no ambiente de experimentação, como testes iniciais.

Dois exemplos de aplicação são apresentados no Capítulo 6. Eles foram escolhidos com a intenção de exemplificar o uso da regionalização em casos próximos a problemas reais e explorar as diferentes possibilidades apresentadas pelo procedimento.

Finalmente, no Capítulo 7, é realizada uma conclusão geral sobre as contribuições que o trabalho oferece e as possibilidades de futuros desenvolvimentos.

CAPÍTULO 2
CLASSIFICAÇÃO E REGIONALIZAÇÃO

Classificar coisas em categorias é uma atividade humana das mais antigas e comuns. Pessoas, objetos ou eventos encontrados em diversas tarefas diárias são, algumas vezes, tão numerosos que é impraticável um tratamento individualizado, como entidade única (Andeberg, 1973). A separação em categorias é necessária e empregada em uma incontável série de atividades. Descrições de indivíduos, por exemplo, consideram aspectos como idade, profissão, renda, religião, etc. Cada um destes itens é um atributo de um indivíduo. Os atributos, sejam eles referentes a pessoas, animais, plantas ou a objetos, são determinantes em qualquer processo de classificação. Há diferentes tipos de classificação, como a taxonomia utilizada na Biologia, mas no contexto deste trabalho a classificação está relacionada ao uso de técnicas *de Análise de Cluster* (AC) para a formação das classes. O principal objetivo deste capítulo é apresentar uma revisão bibliográfica sobre um tipo especial de classificação, a *Regionalização*. A Regionalização é aplicada a objetos com representação espacial por áreas (poligonal), onde os objetos membros de uma classe formam uma região contígua. A parte inicial do capítulo (Seção 2.1 a Seção 2.5) é dedicada aos principais conceitos, elementos e técnicas de Análise de Cluster, destacando os aspectos importantes para o desenvolvimento do nosso trabalho. O restante do capítulo trata exclusivamente da Regionalização (Seção 2.6), apresentando as abordagens utilizadas (Seção 2.7) e alguns métodos específicos (Seção 2.8).

2.1 – CLASSIFICAÇÃO POR ANÁLISE DE CLUSTER

O termo *classificação* tem um sentido amplo e é utilizado simultaneamente em vários processos e áreas do conhecimento envolvendo o agrupamento de objetos ou

observações em categorias. Gordon (1981) define e discute três importantes formas de classificação: *Identificação*; *Análise Discriminante*; e *Análise de Cluster (AC)*. A *Identificação* é um processo ou ato de associar um novo objeto, ou observação, a uma categoria apropriada dentre um conjunto de categorias existentes. Os atributos essenciais das classes são conhecidos. Exemplo desta forma de trabalho é a utilizada por biólogos na taxinomia, classificando animais e plantas. Na *Análise Discriminante*, a estrutura da categoria é parcialmente conhecida. A informação disponível pode variar de uma quase completa descrição da classe até o mero conhecimento do número de classes. O elemento que distingue a *Análise Discriminante* da *AC* é a utilização de amostras das classes. Portanto, na abordagem utilizada na *Análise Discriminante,* a estrutura é parcialmente conhecida e as informações que faltam são estimadas a partir das amostras das classes de objetos.

Enquanto que a preocupação central, tanto na *Identificação* quanto *na Análise Discriminante*, é classificar as novas observações, na abordagem por *AC* pouco ou nada é conhecido a respeito da estrutura das classes e o objetivo é, portanto, descobrir a estrutura das categorias a partir do próprio conjunto de observações. Este problema é freqüentemente definido como a procura por grupos naturais, ou seja, agrupar observações de modo que o grau de associação natural seja alto entre observações de um mesmo grupo e baixo entre membros de grupos diferentes. Neste trabalho, serão empregadas exclusivamente técnicas de *AC* e, portanto, passaremos a utilizar o termo *classificação* com o sentido mais restrito.

Apesar de ter aplicação e objetivo bem definidos, a *Análise de Cluster* não define um único método com regras de utilização claras e precisas, e sim, é um termo que abrange um conjunto de procedimentos heurísticos e alguns elementos de estatística aplicada (Andeberg, 1973). A *AC* compreende, portanto, um conjunto de diferentes métodos utilizados para descobrir estruturas em um conjunto complexo de dados.

O objetivo principal e comum aos métodos *AC* é separar objetos ou observações em *classes naturais* de forma que os elementos pertencentes a um mesmo grupo tenham um alto grau de semelhança ou similaridade, enquanto que, quaisquer elementos pertencentes a grupos distintos, tenham pouca similaridade entre si. Assim, a tarefa

básica dos métodos de *AC* é classificar um conjunto de objetos em subconjuntos (classes) segundo um ou mais critérios apropriados. Os critérios mais comuns adotados em *AC* são: homogeneidade e separação. A homogeneidade refere-se a objetos pertencentes a uma mesma classe, que devem ser tão similares quanto possível; a separação, por sua vez, está relacionada a objetos de diferentes classes, que devem ser tão distintos quanto possível (Maravalle et al., 1997).

2.2 – ELEMENTOS DA CLASSIFICAÇÃO POR ANÁLISE DE CLUSTER

A qualidade da classificação obtida pela utilização de *AC* depende de uma série de definições coerentes por parte do usuário. A seguir, são apresentados alguns elementos importantes presentes no desenvolvimento dos procedimentos de *AC*. Eles serão apresentados em uma seqüência que obedece, normalmente, a ordem de execução de uma análise. Esta seqüência foi proposta por Andeberg (1973) e aqui é apresentada com algumas alterações:

i) **Definição dos objetos**: O primeiro passo da análise é a escolha dos objetos que serão utilizados na classificação. Estes objetos podem ser observações, amostras, casos, eventos, etc. Mais importante do que a definição do tipo de objeto, já que esta escolha é quase uma imposição do problema a ser analisado, é a escolha coerente do conjunto de objetos a serem classificados. Outro aspecto relevante a ser considerado é: caso o procedimento de classificação seja executado sobre um subconjunto da população, de forma que os grupos identificados sejam utilizados para a classificação do conjunto total de objetos, cuidados adicionais têm que ser observados na seleção das amostras para garantir a aleatoriedade e independência dos dados.

ii) **Escolha das variáveis**: Neste passo, são definidos as características e atributos que serão utilizados na categorização dos objetos. A escolha das variáveis é muito importante e terá uma influência direta nos resultados finais obtidos com a classificação. Variáveis em número elevado podem sobrecarregar desnecessariamente o processamento do método de classificação e aumentar em demasiado o número de classes resultantes em alguns procedimentos. Por outro lado, se alguma variável

importante e determinante para a análise for desconsiderada, classes idealmente distintas serão agrupadas.

iii) **Homogeneização das variáveis**: Em dados reais, as diversas variáveis associadas aos objetos são freqüentemente heterogêneas, variando em tipo e escala, ocasionando prejuízos para uma boa classificação. As varáveis podem ser expressas por valores numéricos (idade, porcentagem, etc.), binários (sim ou não, 1 ou 0), nominais (cor dos olhos, filme favorito) e ordinais (patente militar). Combinar variáveis de diferentes tipos para estabelecer um índice de similaridade não é uma tarefa simples ou direta. Mesmo com relação a variáveis numéricas, elas podem apresentar médias e dispersões muito diferentes entre si, e ainda, expressar valores em unidades de medida distintas. Normalmente estes problemas são solucionados com a transformação e padronização das variáveis.

iv) **Medidas de similaridade**: A maioria dos métodos propostos para a AC utiliza uma medida de similaridade para avaliar o grau de semelhança entre dois objetos durante o processo de agrupamento. Muitas vezes, esta medida é apresentada como sendo a "*distância"* entre dois objetos, considerando os valores dos seus atributos. A combinação entre a escolha das variáveis, transformações das variáveis (homogeneização) e a medida de similaridade adotada são o que determina, operacionalmente, o termo "associação natural". Na próxima seção, serão apresentadas algumas medidas de similaridade utilizadas mais freqüentemente.

v) **Critérios de agrupamento**: A escolha de um critério de agrupamento para a realização do procedimento de classificação está diretamente ligada à definição de uma classe e ela é traduzida por uma expressão matemática que avalia e direciona a execução da classificação. Existem vários critérios possíveis de serem utilizados e eles são dependentes da aplicação. Em alguns casos, a escolha do critério é uma opção mais natural, porém em outras situações, é aconselhável experimentar diferentes critérios antes de uma escolha definitiva. A utilização de diferentes critérios pode também ser útil na melhor compreensão da estrutura dos dados. Andeberg (1973) diz que: agrupar um conjunto de dados várias vezes, com diferentes critérios, é uma boa forma de revelar as várias facetas de uma estrutura de dados.

vi) **Escolha do algoritmo:** Existem diversos algoritmos propostos e vários podem produzir resultados semelhantes, em nível de qualidade. A melhor escolha é dependente do objetivo da análise e das características dos dados. Alguns algoritmos podem ser impraticáveis em certas situações devido à demanda computacional para a sua execução. Outros algoritmos possuem como característica a flexibilidade, permitindo a mudança de parâmetros, tais como, número de variáveis e objetos. A escolha do algoritmo está, na prática, limitada pela disponibilidade de procedimentos disponíveis no sistema de análise utilizado. Dois procedimentos freqüentemente utilizados serão apresentados na Seção 2.5.

vii) **Número de classes**: a definição do número de classes é um problema freqüente. Como a estrutura dos dados não é conhecida, também não se sabe quantos agrupamentos naturais existem em um conjunto de dados. Existem procedimentos que fornecem um esquema hierárquico, *dendrograma* (Gordon, 1981), que permite ao usuário investigar diversas classificações, variando o número de classes, de *1* a *n* (onde, *n* é o número de objetos existentes no conjunto de dados). Existem ainda, algoritmos específicos que, em função do conjunto de dados, sugerem um número apropriado de classes.

Alguns dos passos do procedimento de classificação, apresentados acima, serão detalhados nas próximas seções, realçando algumas escolhas que foram utilizadas nas implementações e no desenvolvimento de métodos de classificação presentes neste trabalho.

2.3 – MEDIDAS DE SIMILARIDADE

Coeficiente de Similaridade é a métrica que avalia a semelhança entre dois objetos. Os coeficientes de similaridade entre todos os objetos do conjunto de dados podem ser condensados em uma representação matricial. Nesta matriz, cada elemento, s_{ij}, representa a medida de similaridade entres os objetos *i* e *j*. Esta matriz será aqui chamada de *Matriz de Similaridade (S)*, porém, ela é também referenciada como *Matriz*

de Dissimilaridade ou ainda *Matriz de Proximidade* (Gordon, 1981). Se Ω é o conjunto de objetos a serem classificados, o coeficiente de similaridade é tal que:

(i) $s_{ij} \geq 0$ para todo *i* e *j* pertencente a Ω;

(ii) $s_{ii} = 0$ para todo *i* pertencente a Ω;

(iii) $s_{ij} = s_{ji}$ para todo *i* e *j* pertencente a Ω..

As regras (ii) e (iii), implicam que a *Matriz de Similaridade* pode ser especificada por uma matriz diagonal inferior contendo $n(n-1)/2$ elementos:

$$S = \begin{bmatrix} & & & \\ s_{21} & & & \\ s_{31} & s_{32} & & \\ \vdots & \vdots & \ddots & \\ s_{n1} & s_{n2} & \cdots & s_{n,n-1} \end{bmatrix} .$$

Na maioria dos problemas de classificação, cada um dos *n* objetos é descrito por um conjunto de *p* atributos ou variáveis. Desta forma, o conjunto de dados também pode ser representado por uma matriz *n x p*:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} ,$$

onde x_{il} é o valor é o valor do atributo *l* para o objeto *i*.

Utilizaremos esta matriz nas expressões matemáticas que denotam as métricas de similaridade que serão apresentadas a seguir. A métrica mais comum é a *Distância Euclidiana*, dada pela seguinte forma:

$$s_{ij} = \left[\sum_{l=1}^{p} (x_{il} - x_{jl})^2 \right]^{1/2} .$$

Outra métrica utilizada é a *distância quarteirão* (*city block*), dada por:

$$s_{ij} = \sum_{l=1}^{p} \left| x_{il} - x_{jl} \right| \quad .$$

As duas medidas apresentadas acima, são aplicáveis diretamente às variáveis do tipo numérica. As diferenças entre estas duas métricas são ilustradas na Figura 2.1, utilizando um espaço bidimensional. Cada objeto é representado por um ponto no plano, considerando a existência de apenas duas variáveis numéricas. A distância euclidiana fornece uma medida mais precisa da distância entre os pontos, enquanto que a medida quarteirão apresenta uma expressão mais simples, que pode representar, em alguns casos, alguma vantagem computacional.

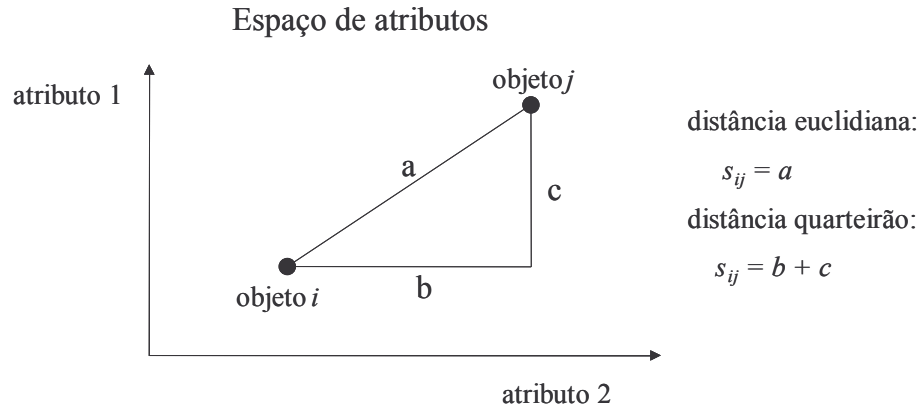


Figura 2.1: Visualização das medidas *euclidiana* e *quarteirão*.

Estas duas medidas, *euclidiana* e *quarteirão*, podem ser entendidas como sendo casos especiais de uma medida geral, chamada de métrica de *Minkowski*, expressa da seguinte forma.

$$s_{ij}^{(\lambda)} = \left[\sum_{l=1}^{p} \left| x_{il} - x_{jl} \right|^{\lambda} \right]^{1/\lambda} \qquad \lambda > 0 \, .$$

As medidas quarteirão e euclidiana são obtidas na métrica de Minkowski fazendo $\lambda = 1$ e $\lambda = 2$, respectivamente. Gordon (1981) diz que a escolha de um valor apropriado para λ, seria em função da ênfase pretendida à maior variação de uma única variável: maiores valores para λ, enfatizam a variável com maior diferença entre $\left|x_{il} - x_{jl}\right|$.

Para permitir uma padronização ou ponderação entre as variáveis, pode-se incluir um novo elemento em qualquer das métricas acima, w_k, que é um vetor com k elementos. Abaixo, é mostrada a pequena alteração na expressão geral da métrica de Minkowski:

$$s_{ij}^{(\lambda)} = \left[\sum_{l=1}^{p} w_k \left| x_{il} - x_{jl} \right|^{1/\lambda} \right]^{1/\lambda} \quad , \lambda > 0 \ .$$

Outras métricas são sugeridas na literatura, assim como formas de homogeneizar as variáveis e trabalhar simultaneamente com variáveis de mais de um tipo. Andeberg (1973) apresenta uma longa discussão sobre este assunto. Neste trabalho, utilizaremos sempre atributos representados por valores numéricos padronizados e a distância euclidiana como medida de similaridade entre os objetos.

2.4 - CRITÉRIOS DE AGRUPAMENTO

Existem vários critérios de agrupamento possíveis de serem utilizados nos procedimentos de classificação e a escolha é dependente do objetivo do trabalho. Um critério de agrupamento comum, o qual foi utilizado para avaliar os métodos de regionalização investigados neste trabalho, é minimizar a seguinte função:

$$f(\Pi) = \sum_{c=0}^{k} SQD_c \quad , \tag{2.1}$$

onde:

- Π é uma partição dos n objetos em k classes.
- SQD_c é a *soma dos quadrados dos desvios* da classe c.

SQD é uma medida da dispersão dos valores dos atributos dos objetos de uma classe em relação aos valores médios dos atributos (centróide) para a classe. Se as classes forem homogêneas, os valores de *SQDs* serão pequenos. Assim a qualidade da partição é inversamente proporcional ao valor da *expressão (2.1)*. O *SQD* de uma classe é expresso por:

$$SQD_c = \sum_{l=1}^{p} \sum_{i=1}^{n_c} \left(x_{il} - \overline{x_l} \right)^2 \quad , \tag{2.2}$$

sendo:

- n_c, o número de objetos membros da calsse c;
- x_{il}, o atributo l do objeto i ;
- p, o número de atributos considerados na análise;
- $\overline{x}_l$, o valor médio do atributo l, para uma classe, dado por:

$$\overline{x_l} = \frac{1}{n_c} \sum_{i=1}^{n_c} x_{il} \ .$$

2.5 – MÉTODOS DE CLASSIFICAÇÃO

Os métodos de *AC* podem ser classificados em duas categorias principais: métodos *hierárquicos* e métodos por *particionamento iterativo (ou relocação iterativa)*. A seguir, veremos algumas características destas duas categorias, detalhando dois procedimentos de classificação específicos. Os dois métodos escolhidos foram implementados para avaliação, já que algumas abordagens de regionalização necessitam de algoritmos de classificação para a sua execução.

2.5.1 – MÉTODOS HIERÁRQUICOS

Os métodos hierárquicos permitem permite obter, no mesmo processo, vários níveis de agrupamento. Uma informação gráfica contendo o histórico das fusões é facilmente gerada em procedimento deste tipo podendo auxiliar na análise dos dados. Este dispositivo gráfico recebe o nome de *dendrograma* (Andeberg, 1973; Gordon, 1981;

Richards, 1995). Os métodos hierárquicos podem ser *por aglomeração* ou *por divisão*. Nos métodos hierárquicos por aglomeração, inicialmente cada objeto é uma classe e, a cada passo do procedimento, as duas classes mais similares são fundidas, até que, ao final, exista somente uma grande classe, contendo todos os objetos. A Figura 2.2 mostra os vários estágios para um processamento fictício. Nesta figura, é possível acompanhar o desenvolvimento do algoritmo para um caso simples, onde existem apenas duas variáveis (espaço de atributos bidimensional).

Os passos típicos de um método hierárquico por aglomeração são:

Passo 1: Iniciar com n classes, cada uma contendo um objeto.

Passo 2: Calcular as coeficientes de similaridades entre os objetos.

Passo 3: Fundir o par de classes com maior similaridade;

Passo 4: Recalcular os coeficientes de similaridade da nova classe com as demais.

Passo 5: Repete os *passos 3* e *4*, $n-1$ vezes.

O cálculo dos coeficientes de similaridade entre as classes no algoritmo hierárquico, apresentado acima e ilustrado na Figura 2.2, pode utilizar um ponto central para representar uma classe com mais de um membro, sendo que os valores dos atributos para este ponto central são os valores médios dos atributos dos objetos membros da classe. Em um outro procedimento hierárquico por aglomeração, chamado de *Ligação Única*, o coeficiente de similaridade entre duas classes é igual ao menor valor possível para o coeficiente de similaridade entre quaisquer dois objetos, sendo cada objeto pertencente a uma das classes envolvidas. Este forma de avaliar a similaridade entre classes elimina o cálculo das médias dos atributos, necessário para estabelecer o centro da classe a cada passo do algoritmo, o que é interessante do ponto de vista do custo computacional, pode produzir classes de forma alongada, efeito denominado de "*encadeiamento"* (Andeberg, 1973). Uma terceira variação do método hierárquico por aglomeração, *Ligação Completa,* utiliza como medida de similaridade

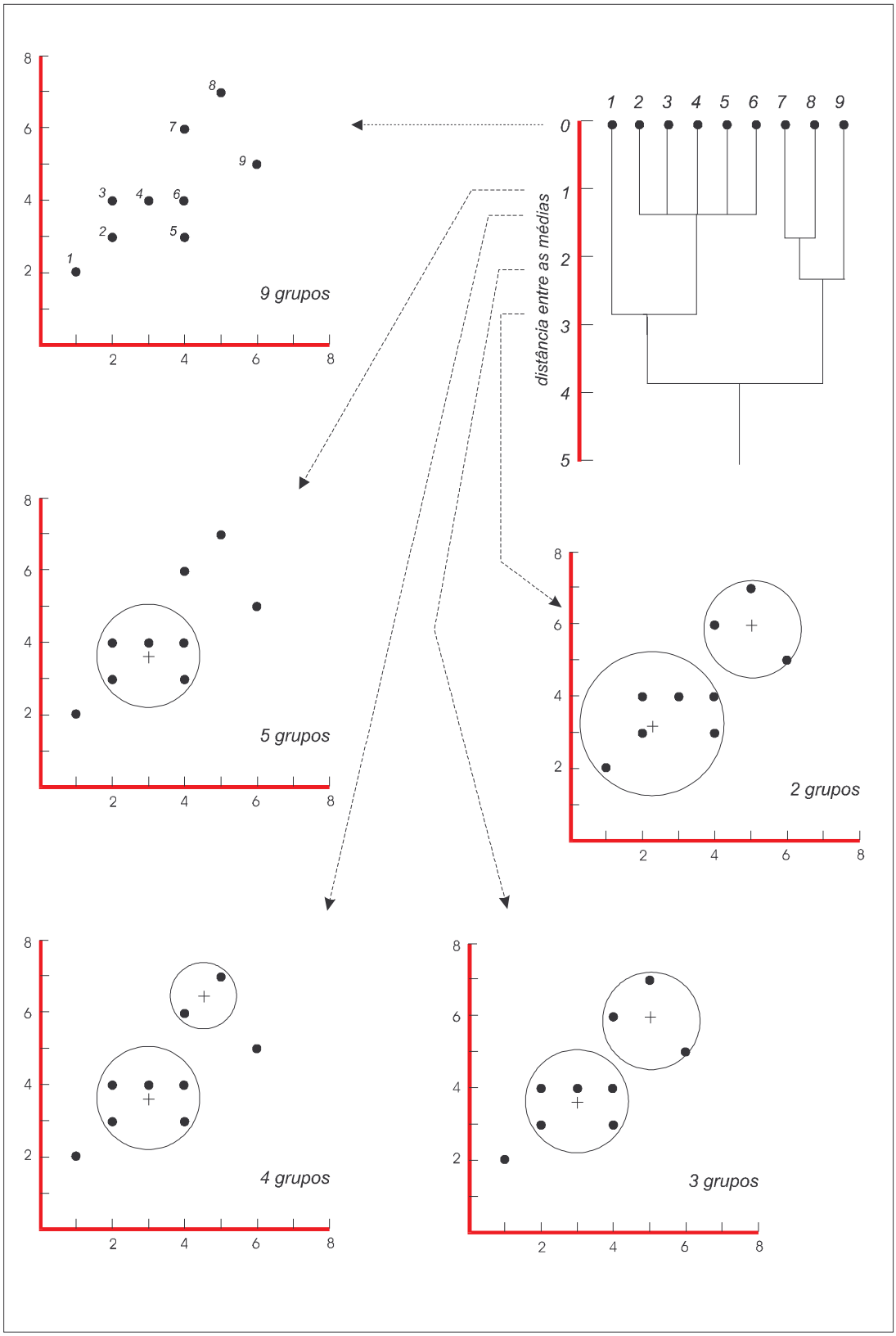


Figura 2.2: Evolução do método hierárquico por aglomeração (adaptado de Richad, 1995).

entre as classes o maior dos valores para o coeficiente de similaridade, medido entre dois objetos de classes distintas (Andeberg, 1973; Gordon, 1981).

No procedimento descrito nesta seção, as classes são fundidas até restar um único grupo. Neste caso, o número de classes, entre 1 a *n*, é escolhido posteriormente. Porém, este procedimento pode ser facilmente alterado, no *passo 5*, de modo que seja interrompido ao atingir um número de classes pré-estabelecido pelo analista.

De forma oposta ao hierárquico por aglomeração, os métodos hierárquicos por divisão, iniciam com todos os objetos pertencendo a um único agrupamento, o qual vai sendo sucessivamente dividido, até que no final, cada classe contenha um único elemento. Esta variação é normalmente mais dispendiosa, do ponto de vista do esforço computacional, e raramente utilizada. Em *Sensoriamento Remoto*, ela é praticamente descartada devido ao grande número de objetos, pixels, existentes nas imagens de satélite (Richards, 1995).

Ng e Han (1994) afirmam que os métodos hierárquicos (tal como o procedimento descrito nesta seção) apresentam uma deficiência relacionada à não revisão dos agrupamentos durante a sua execução, ou seja, uma vez realizado a fusão de dois objetos, passando ambos a pertencer a uma mesma classe, estes objetos não mais são separados até o final do procedimento. De forma similar, nos métodos hierárquicos por divisão típicos, uma vez separados dois objetos, eles não mais serão agrupados em uma mesma classe.

2.5.2 – MÉTODOS DE CLASSIFICAÇÃO POR RELOCAÇÃO ITERATIVA

Os métodos por relocação iterativa buscam encontrar a melhor partição dos *n* objetos em *k* grupos. Freqüentemente, as *k* classes encontradas por esta categoria de métodos são de melhor qualidade (ou seja, grupos internamente mais homogêneos) do que os *k* grupos produzidos pelos métodos hierárquicos. Devido a este melhor desempenho, os algoritmos de relocação iterativa têm sido mais investigados e utilizados (Ng e Han, 1994). Os métodos de relocação mais utilizados são baseados em um ponto central (média dos atributos dos objetos - *k*-médias) (Zhang et al., 2001), mas existem propostas

que utilizam um objeto representativo para a classe (*k-medoids)* (Kaufman e Rousseeuw, 1990).

O método de classificação *k-médias* é bastante difundido, existindo muitas variações propostas na literatura, recebendo diversos nomes, como: *isodata ou migração de médias*. Ele é muito utilizado em *Sensoriamento Remoto* com a finalidade de executar procedimentos de *classificação não supervisionada* de imagens de satélite (Schowengerdt, 1997). Este método exige a definição prévia do número de classes e do posicionamento inicial dos centros das *k* classes no espaço de atributos. As variações e melhorias propostas para o método dizem respeito à definição inicial dos centros das classes e à adoção de avaliações, realizadas durante, ou ao final, do processo de agrupamento (Richards 1995; Zhang et al. 2001).

Os passos básicos de um procedimento baseado no *k-médias* são:

Passo 1: Escolha de *k* objetos para serem centros iniciais das *k* classes.

Passo 2: Cada objeto é associado a uma classe, para o qual a *distância* entre o objeto e o centro desta classe é menor que as demais.

Passo 3: Os centros das classes são recalculados em função dos atributos de todos os objetos pertencentes à classe.

Passo 4: Volta ao *passo 2*, até que os centros das classes se estabilizem ou não se movam significativamente.

A cada iteração, os objetos são agrupados em função das distâncias aos centros das classes e, por conseqüência, os centros das classes são reavaliados (*passo 3*). Isto provoca, no espaço de atributos, um deslocamento dos centros das classes (atributos médios) à medida que o procedimento avança. O algoritmo é interrompido quando os centros das classes não são mais deslocados, ou há uma insignificante relocação de objetos entre as classes (verificado, por exemplo, por pequena variação no valor de uma expressão matemática que mede a qualidade da partição). A Figura 2.3 ilustra o desenvolvimento do método, para o mesmo caso simples apresentado anteriormente.

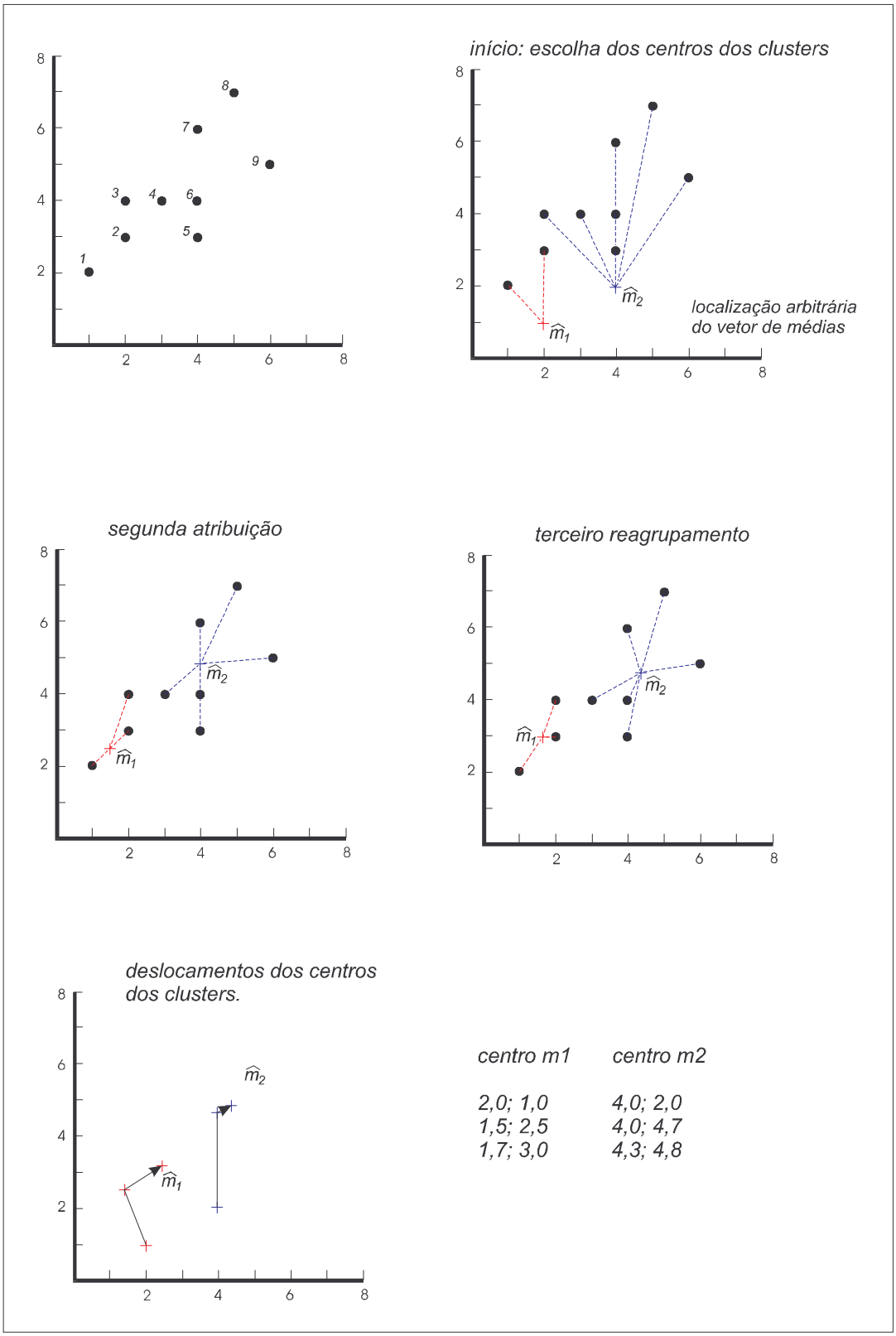


Figura 2.3: Evolução do método por relocação iterativa - *k-médias*.

2.6 – REGIONALIZAÇÃO

Uma restrição freqüentemente imposta a procedimentos de classificação, quando aplicada a objetos espaciais com representação geográfica poligonal, é a exigência de contigüidade entre os objetos de uma mesma classe, ou seja, objetos membros de uma mesma classe devem formar uma região única e espacialmente contígua. Este tipo de classificação é, algumas vezes, tratada como *classificação com restrição de contigüidade* (Gordon, 1996) ou, mais especificamente, como um procedimento de *regionalização* (Wise et al., 1997; Openshaw, 1995). A Figura 2.4 mostra o resultado de um procedimento de classificação com restrição de contigüidade, onde os objetos de uma mesma classe aparecem com uma mesma cor.

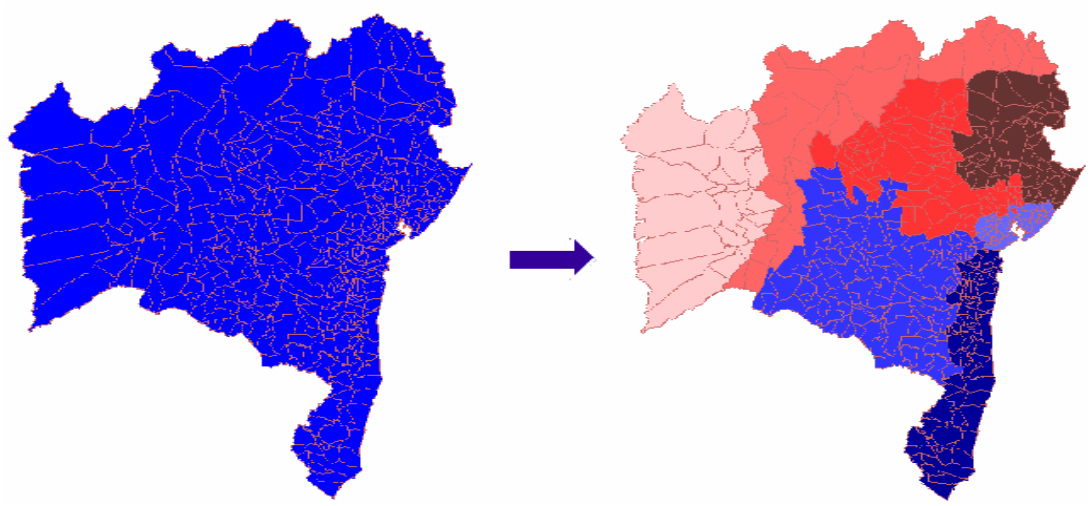


Figura 2.4: Regionalização: classificação com restrição de contigüidade.

A regionalização pode ser utilizada para gerar uma nova repartição do espaço de estudo, resultando em novas unidades de áreas (*regiões*), com dimensões geográficas mais abrangentes e em um número menor de objetos. Alguns motivos para se agrupar unidades espaciais básicas em regiões maiores são: o aumento da representatividade dos valores dos atributos e das taxas associadas às unidades de área; a redução dos efeitos da imprecisão nos valores das variáveis; a redução dos erros associados ao posicionamento geográfico de eventos; e facilitar a análise dos dados (Openshaw, 1995; Wise et al., 1997).

2.7 – ABORDAGENS PARA A REGIONALIZAÇÃO

Existem três abordagens utilizadas para a condução da regionalização. No primeiro tipo de abordagem, o processo é realizado em dois estágios independentes. No estágio inicial, não é considerada qualquer informação espacial e um procedimento de classificação convencional (sem restrição de contigüidade) é executado, utilizando os atributos não-espaciais dos objetos. No segundo estágio, os grupamentos obtidos na primeira fase são reavaliados, observando-se as relações de vizinhança dos objetos. Então, objetos similares agrupados em uma mesma classe na fase inicial, mas sem contigüidade espacial, são separados, no segundo estágio, formando regiões distintas.

Este tipo de abordagem permite identificar, entre os dois estágios, se objetos similares estão, ou não, espalhados pela área de estudo, o que pode ser utilizado como uma rápida avaliação da dependência espacial entre os objetos. Outro aspecto positivo, assinalado por Openshaw and Wymer (1995), refere-se ao controle da similaridade entre os objetos de uma mesma região, garantido pelo primeiro estágio do processo. O inconveniente desta abordagem está na falta de controle sobre o número de regiões resultantes (Wise et al., 1997). Os casos com pequena dependência espacial entre os objetos, por exemplo, tenderão a produzir um número elevado de regiões.

Na segunda abordagem utilizada em procedimentos de regionalização, a similaridade entre os objetos é avaliada considerando simultaneamente a posição geográfica dos objetos e seus atributos não-espaciais. As coordenadas do centróide da área são utilizadas para representar a localização dos objetos e consideradas como atributos adicionais ao problema de classificação. São utilizados algoritmos de classificação onde a avaliação da similaridade contém duas componentes ponderadas, uma para o espaço de atributos e outra para o espaço geográfico. Se o peso dado para a componente geográfica for forte o suficiente, os grupos resultantes do processo de classificação serão contíguos. A definição dos pesos para as duas componentes é uma dificuldade apresentada por este tipo de abordagem, pois o analista terá que executar o procedimento várias vezes, experimentando vários pesos até obter regiões contíguas.

A abordagem por componentes ponderadas é utilizada pelo sistema SAGE (*Spatial Analysis in a GIS Environment*) em seu método de regionalização (Ma et al., 1997), o qual utiliza um procedimento de classificação baseado na técnica de particionamento iterativo *k-médias*, sendo o critério de agrupamento representado por uma função objetivo formada por três componentes independentes: a) *homogeneidade*: regiões formadas por objetos similares, considerando atributos não-espaciais; b) *compacidade*: as coordenadas dos centróides das áreas correspondentes aos objetos membros são próximas; c) *igualdade*: a soma dos valores de um determinado atributo, considerando todos os objetos membros, são semelhantes para todas as regiões (população, por exemplo).

A função objetivo (f_o) do algoritmo de regionalização do SAGE é definida como uma soma ponderada, dada por (Ma et al., 1997):

$$f_o = w_h f_h + w_c f_c + w_i f_i \ ,$$

onde:

- w_h, w_c *e* w_i são os pesos referentes às três componentes (*homogeneidade, compacidade e igualdade)*;

- f_h : é uma medida da dispersão de um ou mais atributos dos objetos nas regiões; É equivalente ao somatório dos *SQDs* (soma dos quadrados dos desvios) das *k* regiões, conforme a Expressão (2.1), apresentada na Seção 2.4.

- f_c : é uma medida da dispersão dos objetos em relação ao centros das classes, tomada no espaço geográfico;

- f_i : é uma medida do equilíbrio da distribuição de uma determinada variável pelas classes (como o total de população humana em cada região), correspondente à soma dos quadrados dos desvios entre o somatório dos valores de um determinado atributo, para os objetos membros das regiões, em relação ao valor de referência (soma do atributo de todos os objetos, dividido pelo número de grupos).

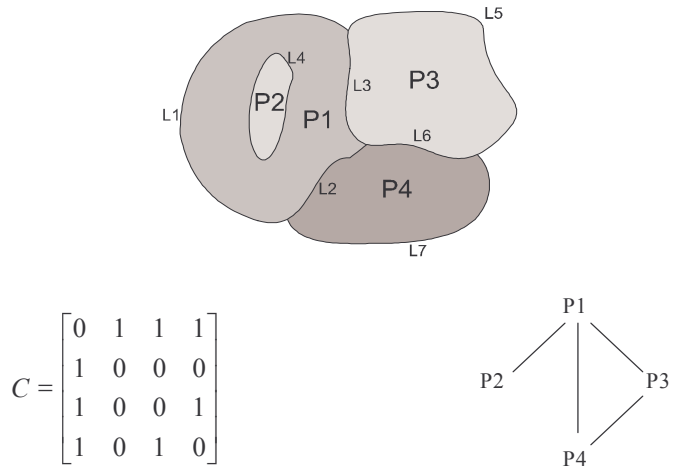
O SAGE utiliza a componente *igualdade* para produzir regiões equilibradas em relação a um determinado atributo, de modo a possibilitar comparações apropriadas entre as regiões, como por exemplo, a taxa de mortalidade por câncer nas diferentes regiões (Wise et al., 1997).

Outros trabalhos que adotam esta abordagem são apresentados em Cliff et al. (1975) e Martin (1998). Openshaw et al. (1998) criticam a abordagem de regionalização por componentes ponderadas devido a necessidade de haver uma apropriada padronização das funções componentes, já que elas são expressas em unidades distintas. Os autores afirmam também que uma estratégia melhor e mais simples é selecionar uma das funções componentes como sendo a função objetivo e tratar os outros critérios como restrições de igualdade ou desigualdade (maior que, menor ou igual a, etc.), definindo de forma explícita seus valores (ex.: população mínima dentro de uma região), criando *funções de penalização* que refletirão a violação das restrições.

Na terceira abordagem utilizada para o procedimento de regionalização, o relacionamento de vizinhança entre os objetos é explicitado por meio de dispositivos auxiliares, como: uma matriz, um grafo ou listas de objetos vizinhos (Figura 2.5). No caso do uso de uma matriz, ela é chamada de *matriz de conectividade (C)*, onde cada elemento, c_{ij}, indica se os objetos i e j são contíguos, ou não. Assim, $c_{ij} = 0$ para objetos não contíguos e $c_{ij} = 1$ para objetos contíguos. De forma equivalente, quando é utilizado um *grafo*, cada objeto é representado por um vértice. Quando os objetos são vizinhos, existe uma aresta ligando os dois vértices correspondentes no grafo (Maravalle et al., 1997; Gordon, 1996).

Dentro desta terceira abordagem, os algoritmos de *AC* tradicionais precisam ser adaptados para o uso em procedimentos com restrição de contigüidade espacial. Nos algoritmos hierárquicos por aglomeração, por exemplo, duas classes mais similares são aglutinadas somente se existir pelo menos dois objetos contíguos, um objeto de cada classe. Nos algoritmos de *relocação iterativa* adaptados para a restrição de contigüidade, um objeto só pode ser relocado em classes que possuam ao menos um objeto contíguo a ele (Gordon, 1996). O método de regionalização AZP (*Automatic*

Zoning Procedure), que veremos na próxima seção, é um método de relocação iterativa que considera as relações de vizinhança dos objetos.



$$C = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Figura 2.5: Representação da relação espacial de vizinhança por matriz e grafo.

Quando a estrutura de vizinhança dos objetos é representada por um grafo, a regionalização se equivale ao problema de particionamento de grafos. A complexidade associada ao problema de particionamento de grafos tem levado a trabalhos buscando a sua simplificação, procurando assim, obter soluções aplicáveis a grafos de maior porte (Karypis and Kumar, 1998). Maravalle and Simeone (1995), Assunção et al. (2000) e Lage et al. (2001) propõem que a partir do *grafo de conectividade* seja gerada uma *árvore geradora mínima (AGM)*. Esta árvore é escolhida de forma a garantir que a soma dos custos associados às arestas seja a menor possível, onde os custos das arestas são inversamente proporcionais às similaridades entre os objetos. A AGM pode ser obtida por algoritmos como: Prim, Kruskal ou Boruvka (Jungnickel, 1999).

As próximas duas seções deste capítulo detalham dois métodos de regionalização que utilizam diretamente a estrutura de vizinhança. Uma comparação entre quatro métodos de regionalização, das diferentes abordagens discutidas nesta seção, será apresentada no capítulo 3.

2.8 – MÉTODOS QUE UTILIZAM A ESTRUTURA DE VIZINHANÇA

Nas abordagens de regionalização realizadas em duas etapas e por componentes ponderadas são empregados métodos de classificação comuns, como o *k-médias* e *hierárquico por aglomeração*, que já foram mostrados na Seção 2.5. Nesta seção, veremos com maiores detalhes dois procedimentos de regionalização representantes da abordagem que utiliza diretamente a estrutura de vizinhança dos objetos: O *AZP* e um método baseado no uso da *Árvore Geradora Mínima*.

2.8.1 – MÉTODO AZP

O *AZP* (*Automatic Zoning Procedure*), proposto inicialmente na década de 70, é um método de regionalização que utiliza a estrutura de vizinhança dos objetos espaciais. Este método começa realizando uma partição aleatória dos n objetos em k regiões e busca, por tentativa e erro, realocar objetos pelas regiões de modo a obter valores cada vez menores para uma função objetivo, sempre respeitando a restrição de contigüidade (Openshaw & Rao, 1995). O *AZP* deu origem a um sistema de zoneamento automático (*ZDES*)[1] e, recentemente, foram propostas algumas melhorias com a intenção de incrementar a qualidade da partição resultante, evitando que o procedimento de busca pela melhor solução fique retido em ótimos locais (Alvanides et al., 2002). Os passos do método *AZP*, conforme (Openshaw & Rao, 1995), são:

Passo 1: gerar uma partição aleatória dos n objetos em k regiões.

Passo 2: construir uma lista de k regiões.

Passo 3: selecionar aleatoriamente uma região k_i da lista de regiões e retirá-la da lista de regiões.

[1]Informações sobre o programa ZDES podem ser obtidas na página:
http://www.geog.leeds.ac.uk/software/zdes/

Passo 4: identificar uma lista de objetos vizinhos à região k_i que podem ser relocados em k_i, sem destruir a contigüidade interna da região doadora.

Passo 5: retirar aleatoriamente um objeto da lista de objetos vizinhos; se ocorrer melhoria na função objetivo, incluir objeto em k_i, retornar ao passo 4; senão, repetir o passo 5 até exaurir a lista de objetos.

Passo 6: Se a lista de regiões não estiver vazia, retornar ao passo 3 para selecionar outra região e repetir os passos de 4 a 6.

Passo 7: Repetir passos de 2 a 6 até não haver mais relocações que promovão melhoria.

2.8.2 – MÉTODO VIA AGM

O método de regionalização via árvore geradora mínima, que será apresentado a seguir, foi proposto em Assunção et al. (2000) e Lage et al. (2001). Ele possui duas fases distintas: a geração da AGM e o particionamento da árvore em *k árvores* disjuntas (*k* é o número de regiões).

FASE 1: GERAÇÃO DA AGM

As informações básicas para a geração da AGM são o grafo de conectividade (*G*) e as medidas de similaridade entre os objetos. Um grafo *G = (V, L)* possui um conjunto de vértices (*V*) e um conjunto de arestas (*L*). A AGM é construída de forma recursiva. O processo tem início a partir de uma árvore T_1, contendo apenas um vértice e, a cada iteração, uma nova aresta e um novo vértice são adicionados à arvore anterior. Na iteração *n*, a árvore T_n contém todos os *n* vértices de *V* e possui um subconjunto de *L*, contendo *n-1* arestas, para o qual a soma dos custos associados às arestas é mínima. Os passos utilizados para a geração da AGM, baseados no algoritmo de Prim, são:

Passo 1: Fazer $T_k = T_1$, contendo um vértice qualquer, v_i, pertencente ao conjunto completo de nós, *V*.

Passo 2: Encontrar a aresta de menor valor em L que ligue qualquer vértice de T_k à um outro vértice de V, não pertencente a T_k, e acrescentar este vértice à árvore, gerando uma nova árvore, T_{k+1}.

Passo 3: Repetir o passo 2 até que todos os vértices tenham sido incluídos na árvore (T_n).

A Figura 2.6 representa a construção de uma *árvore geradora mínima* para um exemplo hipotético. Nesta figura, as larguras das arestas que ligam os nós vizinhos são inversamente proporcionais aos coeficientes de similaridades entre os objetos, ou seja, objetos vizinhos semelhantes são ligados por arestas finas (baixo custo). Assim, em relação ao grafo original, que representava a estrutura de vizinhança dos objetos, a *árvore geradora mínima* corresponde a um subgrafo conectado, sem circuitos e de custo mínimo, contendo somente as arestas mais finas ("mais baratas"). A Figura 2.7 mostra uma AGM produzida em um experimento a partir de um grafo de conectividade e um conjunto específico de atributos.

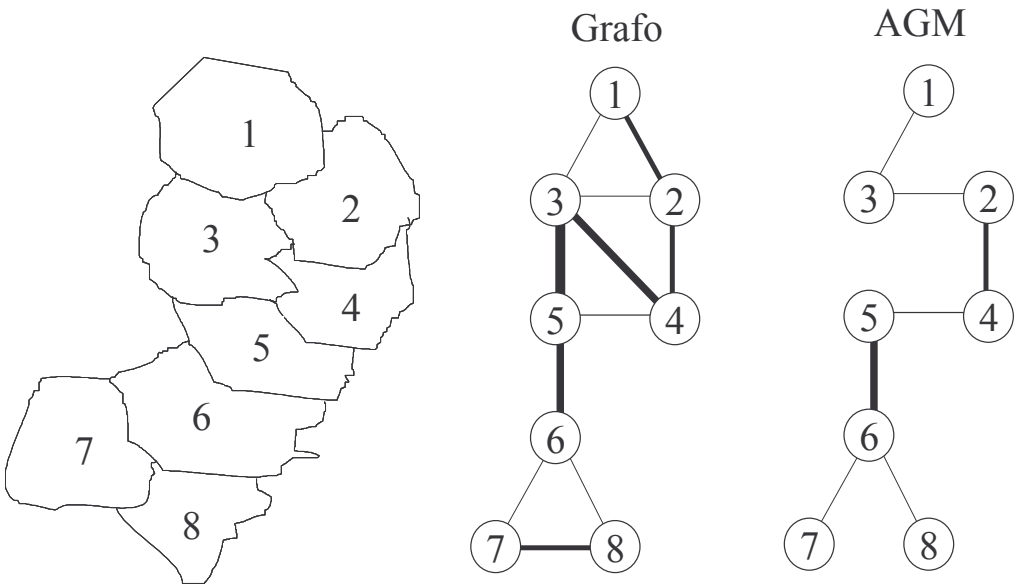


Figura 2.6: Construção da árvore geradora mínima.

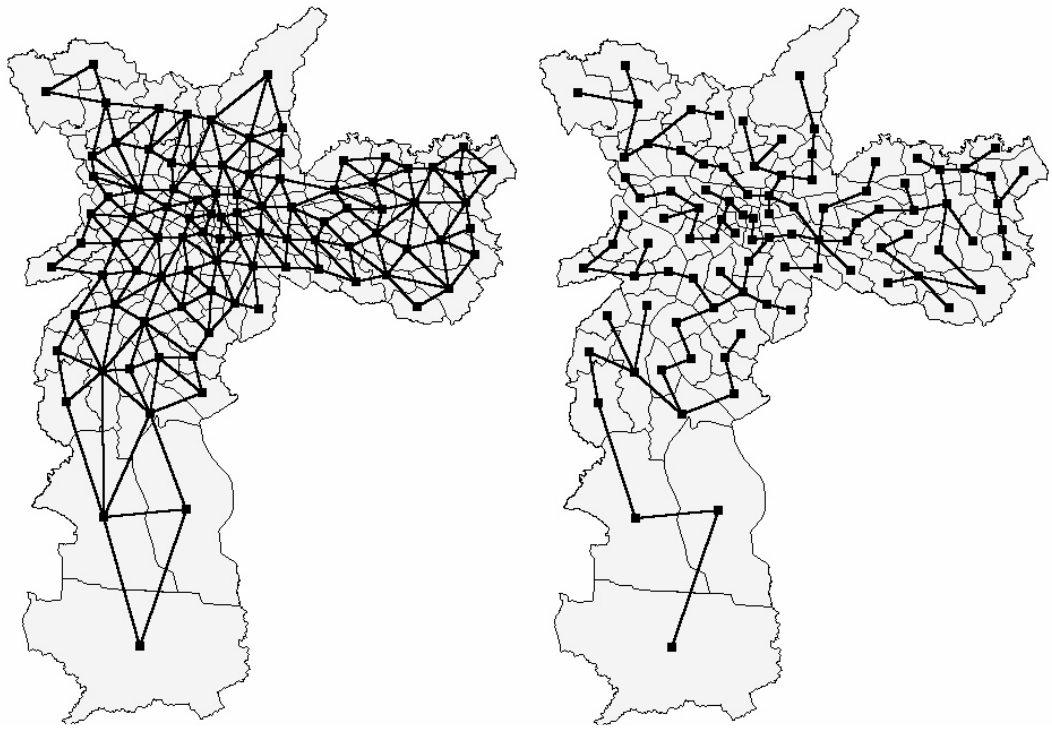


Figura 2.7: Exemplo da AGM, a partir de um grafo de vizinhança e de um conjunto de atributos.

Ao final da execução do procedimento descrito acima, temos *a árvore geradora mínima*, que contém todos os nós, representando todos os objetos. A partir deste ponto, entramos na segunda fase do procedimento de regionalização, onde a árvore será desmembrada. Cada aresta que for eliminada dividirá a árvore em dois subgrafos desconexos, correspondendo a dois clusters espaciais, ou regiões. Portanto, a classificação propriamente dita, se dará com a escolha e eliminação de determinadas arestas a partir da *árvore geradora mínima*. Este processo de eliminação de arestas da árvore corresponde, segundo os autores do método, a uma "poda" da *árvore geradora mínima*.

FASE 2: PODA DA AGM

Após a geração da AGM, o método passa a uma segunda fase. Esta fase consiste em retirar as arestas mais caras. Serão escolhidas *k-1* arestas, para obter *k* regiões

(subárvores). Na fase de poda da AGM, a forma de atribuir custos às arestas é modificada para obter melhores resultados. Esta alteração visa obter regiões mais homogêneas e mais equilibradas em termos de número de objetos por região. O novo custo da aresta é dado por:

$$Custo\ da\ aresta\ l = SQD_T - SQD_l\ ,$$

onde:

i) SQD_T é *soma dos quadrados dos desvios*, associada a árvore *T*, dada por:

$$SQD_T = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2\ ,$$

sendo:

n, o número total de objetos (nós) em *T*;

x_{ij}, o atributo *j* do objeto *i* ;

m, o número de atributos considerados na análise;

$\overline{x}_j$, o valor médio do atributo *j*.

ii) SQD_l é a soma das duas parcelas obtidas da *soma dos quadrados dos desvios* das duas subárvores , T_a e T_b, geradas pela retirada da aresta *l* da árvore *T*:

$$SQD_l = SQD_{Ta} + SQD_{Tb} \quad .$$

Para obter a *soma dos quadrados dos desvios* para as duas subárvores, são calculados os valores médios dos *m* atributos, tal como feito para o cálculo de SQD_T, porém, considerando-se apenas os atributos referentes aos objetos pertencentes a cada subárvore de *T*, T_a e T_b. As podas seguintes são feitas de forma recursiva, escolhendo sempre a aresta de menor custo.

Embora não exista uma forma objetiva e geral para uma avaliação dos procedimentos de classificação, este método possui características que indicam que ele produz bons resultados: a restrição de contigüidade está explícita na AGM; o número de soluções possíveis é limitado pelo pequeno número relativo de arestas existentes na AGM; a avaliação do custo de cada partição através da *soma dos quadrados dos desvios*

privilegia a homogeneidade interna dos grupamentos e é idêntico ao critério de agrupamento utilizado em métodos como *k-médias*, que reconhecidamente, produz bons resultados, como indicado por (Openshaw, 1995).

2.9 – CONCLUSÃO DO CAPÍTULO

Nesta revisão bibliográfica, nós destacamos os métodos AZP (*Automatic Zoning Procedure*) e via *árvore geradora mínima* por eles serem aplicáveis a objetos com representação espacial poligonal e utilizar explicitamente as relações de vizinhança dos objetos. Além disso, o AZP é um método bastante conhecido e serve como um bom parâmetro de comparação, enquanto que o método baseado no uso da AGM foi utilizado como base, para o desenvolvimento do novo método, alterando basicamente, a fase de poda da árvore, com a intenção de tornar o procedimento mais eficiente. No próximo capítulo é realizada uma análise para diferentes procedimentos, escolhidos de forma a abranger as três abordagens utilizadas para a regionalização, com o objetivo de mostrar as principais características e o desempenho proporcionado por cada método.

CAPÍTULO 3
AVALIAÇÃO DE MÉTODOS DE REGIONALIZAÇÃO

Neste capítulo é realizada uma avaliação das abordagens de regionalização. Diferentes métodos foram implementados e submetidos a uma série de testes com a finalidade de avaliar suas principais características. O objetivo principal deste capítulo é demonstrar que a proposta de utilização da AGM como forma de diminuir a complexidade do problema de particionamento de grafos é eficiente e permite a obtenção de uma regionalização de boa qualidade, já que as novas propostas de regionalização, que serão apresentadas nos próximos dois capítulos, também adotam esta solução. Um ambiente de experimentação foi montado para permitir a avaliação dos procedimentos, a visualização gráfica dos resultados e a medição de alguns parâmetros utilizados na análise e comparação entre os métodos, como o tempo de execução dos procedimentos e a qualidade da partição proporcionada pelos métodos. A escolha dos quatro métodos avaliados foi feita de modo que todas as três abordagens, apresentadas no Capítulo 2 estivessem representadas nesta análise.

A primeira seção do Capítulo apresenta o ambiente de experimentação, descrevendo as principais funcionalidades implementadas. A Seção 3.2 apresenta quais parâmetros foram utilizados para quantificar os comportamentos dos diferentes métodos analisados. Na seção 3.3, são avaliados os dois métodos de classificação, apresentados anteriormente (*k-médias* e *hierárquico por aglomeração*). A Seção 3.4 apresenta uma análise individual do comportamento de quatro métodos de regionalização. Na Seção 3.5, é traçado uma análise comparativa dos métodos de regionalização investigados.

3.1 – AMBIENTE DE EXPERIMENTAÇÃO

As avaliações e comparações entre os métodos de regionalização foram realizadas em um ambiente de experimentação. Este ambiente foi desenvolvido no sistema operacional *Windows*, utilizando a linguagem C++, gerenciador de interface QT e uma biblioteca contendo componentes de SIG, chamada *TerraLib*[1]. Esta biblioteca possui o código aberto e foi desenvolvida pela *Divisão de Processamento de Imagem* (DPI) do *Instituto Nacional de Pesquisas Espaciais* (INPE) com o objetivo de facilitar o desenvolvimento de pequenos SIGs voltados para aplicações específicas.

Na construção do ambiente de experimentação foram adicionados alguns elementos (classes específicas, funções necessárias para a implementação dos procedimentos de classificação e regionalização e funções utilizadas para a avaliação) a um outro aplicativo, que continha uma série de funções de visualização de dados espaciais, chamado *TerraView*.

Inicialmente, foram implementados no ambiente: dois procedimentos de classificação, sem restrição de contiguidade; quatro procedimentos básicos de regionalização e algumas variações; algumas funções de análise que fornecem estatísticas das classes e regiões; medida de qualidade da classificação e partição, medida do tempo de execução do procedimento e número de avaliações (esta medida só é aplicada em alguns métodos). Posteriormente, o ambiente foi expandido para ser utilizado em outros experimentos. Destes métodos, o *k-médias* já estava disponível como função na biblioteca TerraLib, sendo implementada apenas uma pequena variação, relativa ao mecanismo de escolha da partição inicial.

Como alternativa ao *k-médias*, foi implementado um segundo método de classificação *hierárquico por aglomeração*. Na versão do método hierárquico implementada, a medida de similaridade entre grupos foi tomada a partir de um ponto central das classes. Estes dois métodos, como dito anteriormente, representam as duas grandes categorias de métodos de classificação em Análise de Cluster: relocação

[1] Informações adicionais sobre a biblioteca TerraLib, em: http://terralib.dpi.inpe.br/

iterativa e métodos hierárquicos. Os algoritmos básicos utilizados nestes dois procedimentos seguem a descrição apresentada no Capítulo 2.

Especificamente para a regionalização, foram implementados alguns procedimentos que permitiram a avaliação das três abordagens utilizadas para a regionalização, sendo que a abordagem que utiliza diretamente a relação de vizinhança entre os objetos, por ser a de maior interesse, foi representada por dois métodos. Os quatro procedimentos avaliados foram:

- realizado *em duas-etapas*.
- *por componentes ponderadas*.
- *AZP*.
- baseado no uso da *AGM*.

Os procedimentos implementados são versões que acompanham a idéia básica dos métodos apresentados no Capítulo 2. O procedimento *por componentes ponderadas*, por exemplo, considerou apenas as componentes fundamentais para a análise: homogeneidade e compacidade.

Além da análise comparativa entre métodos, propriamente dita, o ambiente de experimentação foi utilizado no desenvolvimento de um novo método de regionalização, também baseado na AGM, e na experimentação de uma abordagem de regionalização dirigida pelo usuário, que serão apresentados nos capítulos seguintes. A Figura 3.1 mostra uma visão geral do ambiente de experimentação, destacando a janela de escolha dos métodos de classificação.

Para auxiliar na análise das classificações, o ambiente calcula estatísticas básicas para as classes ou regiões criadas, fornecendo dados como número de membros, médias e variâncias dos atributos para cada classe. Também é possível investigar os demais atributos da base de dados que não foram utilizados na classificação. Para a avaliação da qualidade da classificação ou partição o ambiente fornece valores individuais da homogeneidade para cada classe e uma medida global (índice de qualidade) para o procedimento.

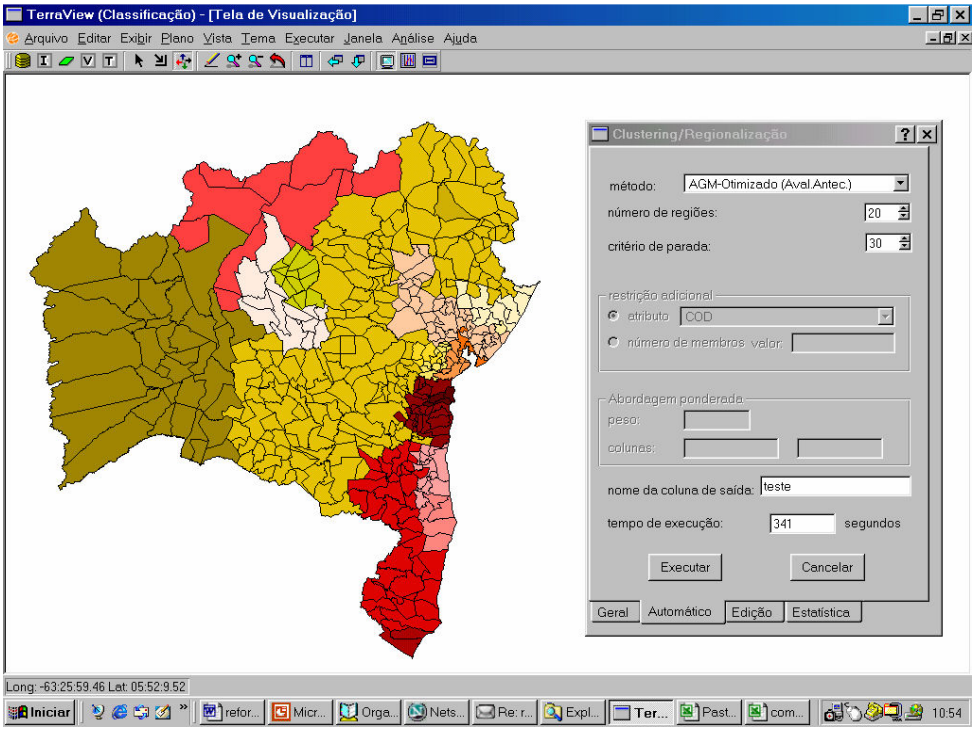


Figura 3.1: Visão geral do ambiente de experimentação.

3.2 – MEDIDAS UTILIZADAS NA ANÁLISE

A análise quantitativa dos procedimentos se baseou em dois aspectos: a qualidade da classificação proporcionada pelos métodos e o esforço computacional requerido. A qualidade de uma classificação é dependente do objetivo da análise e, consequentemente, do critério de agrupamento utilizado. Como o mesmo critério básico de agrupamento, homogeneidade interna das classes, foi utilizado em todos os procedimentos analisados, é possível comparar a qualidade dos métodos. A medida utilizada para a qualidade da classificação, é dada pela expressão:

$$Q(\Pi) = \sum_{i=0}^{k} SQD_i \quad , \tag{3.1}$$

onde:

- Π representa uma partição (ou classificação) dos n objetos em k classes;

- $Q(\Pi)$ é o índice de qualidade da partição;

- SQD_i é a *soma dos quadrados dos desvios* para a classe *i*.

A Expressão 3.1 é idêntica à *função objetivo* apresentada como exemplo de critério de agrupamento no Capítulo 2 (Expressão 2.1). Como o valor de *SQD* de uma classe, é uma medida de dispersão dos atributos dos objetos da classe, este valor é inversamente proporcional a homogeneidade dos objetos membros da classe. A Expressão 3.1 considera simultaneamente a homogeneidade dos agrupamentos, realizando um somatório de todos os *SQDs*. Assim, ela fornece uma medida de qualidade para toda a classificação. Quanto menor for o valor de $Q(\Pi)$, mais homogêneas são as classes e melhor é a qualidade da classificação.

O segundo parâmetro quantitativo utilizado, o esforço computacional, é medido indiretamente pelo *tempo de execução* dos procedimentos, para um mesmo conjunto de dados. A medida de tempo utilizada pelo ambiente de experimentação é obtida por meio de uma função que retorna a contagem do relógio interno do sistema, os valores retornados são inteiros e expressos em segundos. Duas leituras do tempo são realizadas, uma no início e outra ao final dos cálculos. O período de tempo considerado exclui o tempo consumido com a *leitura* e *escrita* em memória dos dados, já que estas operações eram comuns a todos os métodos.

Uma outra medida do esforço computacional implementada no ambiente é o *número de avaliações* realizadas na execução de um procedimento. Esta medida só é válida para alguns dos métodos e será utilizada no próximo capítulo, na comparação entre o método baseado na AGM com busca exaustiva e o método proposto.

3.3 – AVALIAÇÃO DOS MÉTODOS DE CLASSIFICAÇÃO: K-MÉDIAS E HIERÁRQUICO POR AGLOMERAÇÃO

Como o método de regionalização *em duas-etapas* utiliza o resultado de um procedimento de classificação comum (sem restrição de contiguidade) em sua primeira etapa, analisamos os dois métodos (*k-médias e hierárquico por aglomeração*) presentes

no ambiente de experimentação para escolhermos o método de classificação que proporcionasse um melhor desempenho.

O método *k-médias* é reconhecido por apresentar bons resultados e ser de rápida convergência para o resultado final. Isto foi comprovado com os experimentos realizados. O método hierárquico, por sua vez, se mostrou lento e com tendência a gerar grupos unitários. A Figura 3.2 apresenta os resultados obtidos para os dois métodos em um experimento, onde foram classificados noventa e seis objetos correspondendo aos distritos do município de São Paulo, em cinco classes. Foram considerados dois atributos: *índice de desenvolvimento humano* e *índice de qualidade de vida* (dados oriundos do mapeamento da exclusão Social da cidade de São Paulo, realizado por Sposati, 1996). Neste exemplo, o critério de parada para o procedimento *k-médias* foi a estabilização dos centróides das classes. A figura mostra que o método hierárquico formou duas classes contendo apenas um membro. Os objetos que deram origem a estas classes unitárias aparecem mais isolados no espaço de atributo, destoam dos demais objetos e correspondem, de certa forma, a casos especiais. A Figura 3.3 mostra dois gráficos de espalhamento, um para cada método. Os objetos são posicionados conforme seus atributos e os objetos aparecem separados por classes em cores diferentes.

Além da tendência em formar classes unitárias, o método hierárquico apresentou uma grande desvantagem em relação ao *k-médias*: o tempo de execução. Enquanto que o método *k-médias* consumiu menos de 1 segundo, o método hierárquico por aglomeração precisou de dez segundos. Esta significativa diferença é explicada, tanto pela rápida convergência do método *k-médias*, quanto pelo número de coeficientes de similaridades calculados pelos métodos durante a sua execução. No *k-médias*, em cada iteração são calculados $n.k$ distâncias no espaço de atributos, correspondente às distâncias entre os n objetos e as k médias. No hierárquico este número é normalmente mais elevado e concentrado na primeira iteração. No início do método são avaliadas as distâncias entre todos os objetos, correspondendo a $(n-1)!$ cálculos de distância. Nas demais iterações são calculadas apenas às distâncias entre a classe criada pela fusão e as demais classes. Com isto, podemos concluir que, comparativamente, a vantagem em termos do tempo de execução do *k-médias* tende a aumentar com o crescimento do

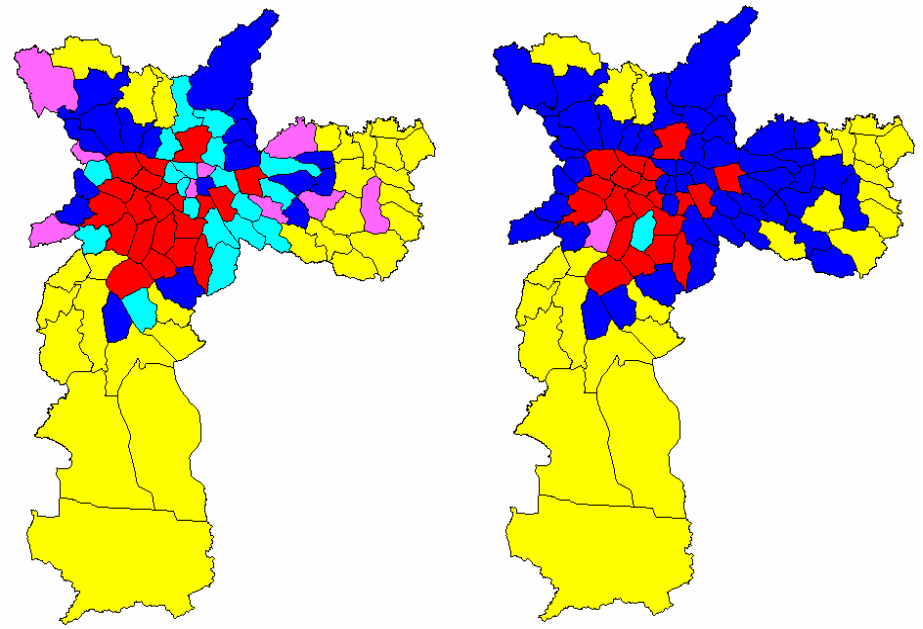
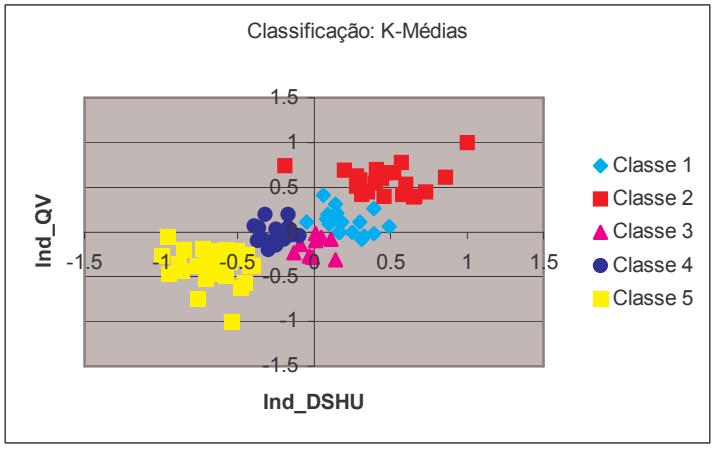


Figura 3.2: Resultado da classificação para os métodos: k-médias (esquerda) e hierárquicos por aglomeração (direita).



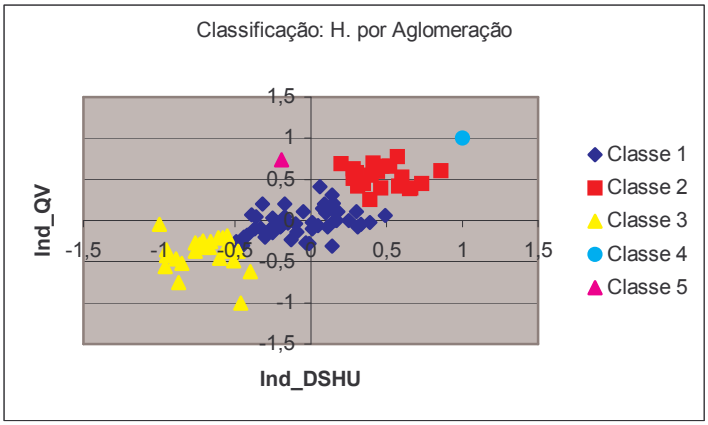


Figura 3.3: Gráficos de espalhamento dos objetos, por classe, para os dois métodos.

número de objetos na análise e, por outro lado, tende a diminuir com o aumento no número de regiões. Como o método de regionalização em duas etapas normalmente gera um grande número de regiões e para controlar este número utiliza-se um pequeno número de classes na primeira fase, a vantagem no tempo de execução tenderá a ser elevada em favor do procedimento *k-médias*.

Devido ao melhor comportamento do método *k-médias* ele foi escolhido para ser utilizado no procedimento de regionalização em duas etapas. Uma vantagem adicional, proporcionada pelo uso do *k-médias*, é podermos comparar quantitativamente a qualidade de todos os quatros métodos de regionalização implementados, já que outros procedimentos (*componentes ponderadas*, AZP e AGM) utilizam o mesmo critério de agrupamento (homogeneidade interna das classes), dado pela Expressão 2.1.

3.4 – AVALIAÇÃO DOS MÉTODOS DE REGIONALIZAÇÃO

Os dados utilizados para analisar o comportamento dos métodos de regionalização, nas próximas seções deste capítulo, são referentes aos 415 municípios do Estado da Bahia. Foram utilizados três atributos normalizados que correspondem às porcentagens das áreas dos municípios com lavoura, pastagem e matas (dados derivados do Censo Agropecuário 1995/1996 – IBGE[2]).

3.4.1 - ABORDAGEM EM DUAS ETAPAS

O primeiro método avaliado foi o procedimento de regionalização realizado em duas etapas. A Figura 3.4 mostra as duas fases deste procedimento. Esta abordagem necessita da estrutura de vizinhança dos objetos para a execução da segunda fase. Portanto, ao tempo de execução do método de classificação *k-médias*, é acrescentando um tempo adicional necessário para a geração da estrutura de vizinhança (ou leitura de arquivo com esta informação) e a posterior reorganização dos objetos em regiões. Para o exemplo, a execução da segunda etapa do método consumiu menos que um segundo para sua execução, indicando que o método não é dispendioso computacionalmente.

[2] Dados do Censo Agropecuário 1995-1996 estão disponíveis em: www.ibge.gov.br.

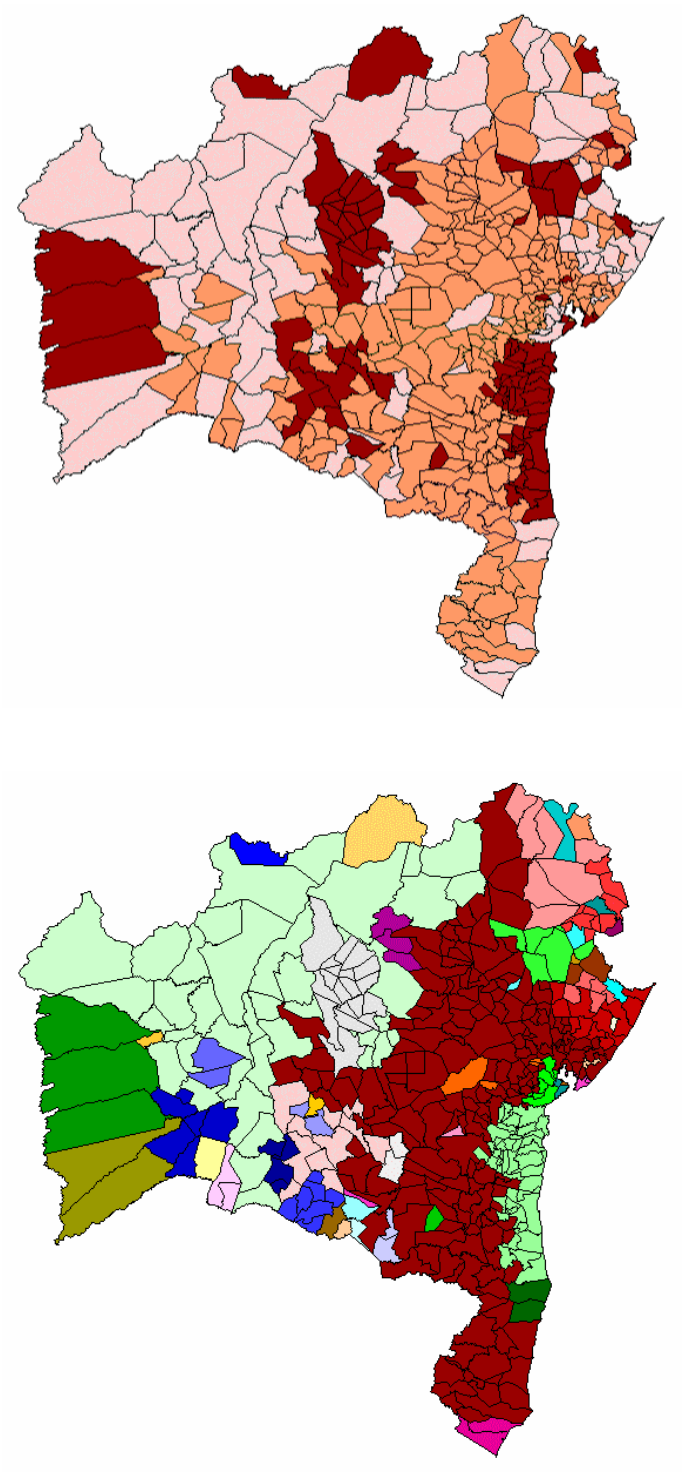


Figura 3.4: Abordagem *em duas etapas*. Acima, as 3 classes resultantes da primeira fase (*k-médias*); abaixo, resultado final (54 regiões).

O grande inconveniente desta abordagem está na falta de controle sobre o número final de regiões. No exemplo apresentado na Figura 3.4, foram gerados três classes na primeira fase e cinqüenta e quatro regiões ao final do procedimento! Os atributos utilizados tinham uma forte dependência espacial e, mesmo assim, foi gerado um número elevado de regiões na segunda fase. Em casos que envolvam objetos e atributos com pequena dependência espacial este problema tende a se agravar, formando um número muito alto de regiões na segunda etapa. Openshaw (1995) afirmou que esta abordagem garante que não haja perda de homogeneidade, já que a "regionalização" só ocorre entre objetos de uma mesma classe. Porém, se o número de regiões for maior que o desejado pelo analista, só restará a ele uma opção, diminuir o número de classes na primeira etapa. Um número menor de classes na primeira etapa significa classes com mais objetos e menos homogêneas e, consequentemente, a partição final perderá qualidade. A falta de controle no número de regiões limita, na prática, a aplicação desta abordagem como método de regionalização. A Tabela 3.1 mostra este descontrole, utilizando uma série de valores para o número de classes na primeira fase, de duas a cinco, e o número resultante de regiões ao final do processo.

Tabela 3.1: Regiões criadas pela abordagem em duas etapas.

Classes	Regiões
2	27
3	54
4	66
5	92

3.4.2 – AVALIAÇÃO DA ABORDAGEM POR COMPONENTES PONDERADAS

O segundo procedimento de regionalização analisado utiliza as coordenadas geográficas dos centróides das áreas como atributos adicionais, atribuindo pesos aos dois grupos de atributos (coordenadas e atributos não-espaciais). Este procedimento seque a abordagem apresentada na Seção 2.7. Esta solução herda a rapidez do método *k-médias*, pois apenas

ocorrem acréscimos de dois atributos na análise e de algumas poucas operações matemáticas. Além disso, este método não necessita da estrutura de vizinhança dos objetos. O tempo de execução, para um procedimento de classificação em vinte grupos, foi inferior a 1 segundo. Um inconveniente desta abordagem está na determinação do peso necessário para que os objetos fiquem agrupados em uma mesma região. Se for utilizado um peso pequeno para as coordenadas geográficas, os objetos de uma mesma classe não serão contíguos. Por outro lado, se o peso da componente geográfica for exagerado, a qualidade da classificação final diminuirá além do necessário.

Na Figura 3.5, é mostrada uma seqüência de resultados utilizando diferentes valores de pesos para ilustrar o processo de definição do peso ideal e sua influência no resultado da classificação. Pode-se observar, pelos índices de qualidade ao lado dos resultados, que quanto mais forte é peso utilizado para a componente geográfica, pior é o índice de qualidade da classificação. Portanto, é importante determinar um peso para a componente geográfica que seja mínimo e suficiente para atender a restrição de contiguidade.

Além da dificuldade de determinar o peso ideal das componentes, existe um outro problema relacionado à variação das dimensões dos objetos. No exemplo, os municípios do interior do estado possuem grande dimensão, enquanto que os municípios do litoral tendem a ter áreas menores, principalmente os municípios no entorno de Salvador. Esta variação faz com que o peso da componente geográfica atue de forma distinta, dependendo do local onde se encontram os municípios. A Figura 3.6 apresenta um detalhe de parte do estado contendo muitos municípios pequenos (pouca extensão). Os municípios estão coloridos conforme sua classificação usando o peso da componente geográfica igual a três. Podemos observar na figura que algumas classes formadas por municípios pequenos ainda não estão contíguas. Isto é provocado pela proximidade geográfica dos centróides das áreas, fazendo com que prevaleça a componente correspondente aos atributos não-espaciais. A contiguidade para todas as regiões, só foi conseguida com o valor do peso igual a vinte (só foram testados valores inteiros). Isto significa que para as demais regiões com municípios mais espaçados, o peso final

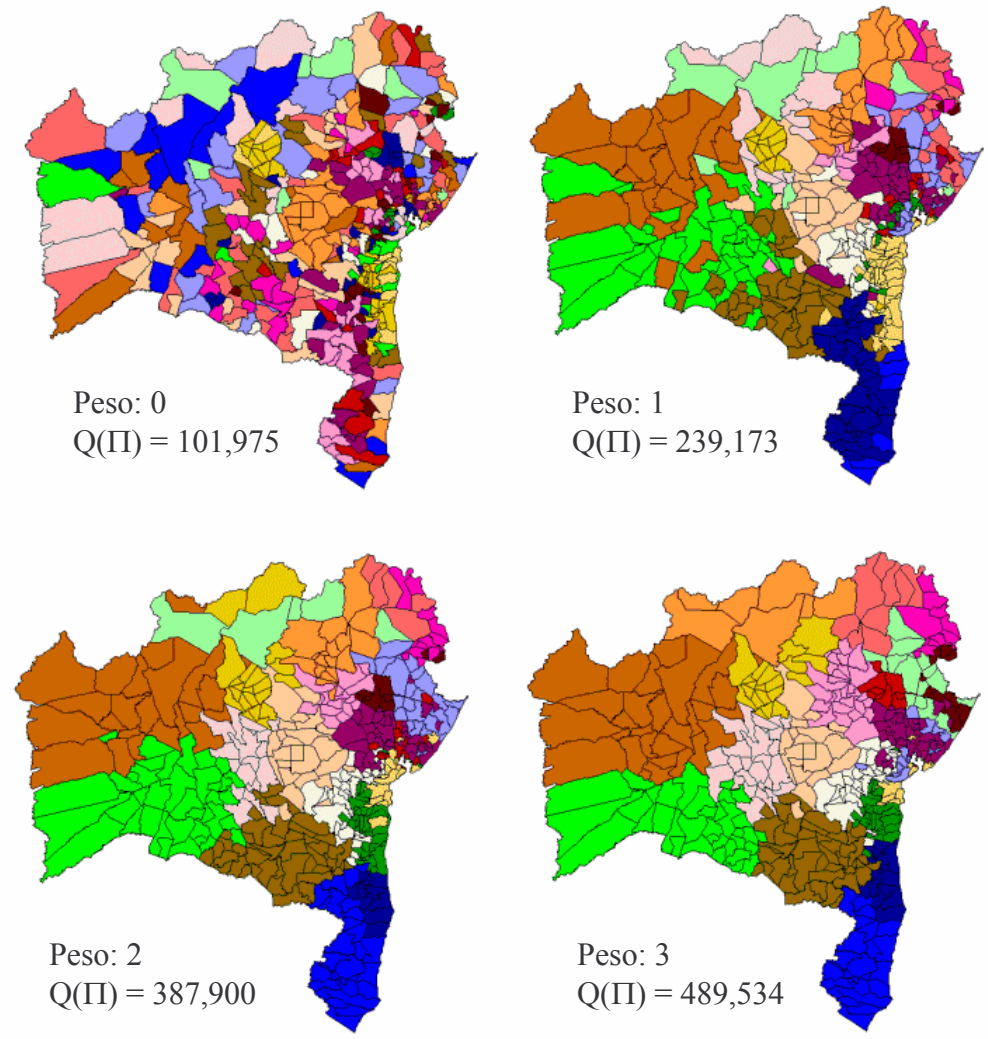


Figura 3.5: Diferentes classificações (em 20 classes) utilizando o método *componentes ponderadas*, com pesos diferentes para a componente geográfica..

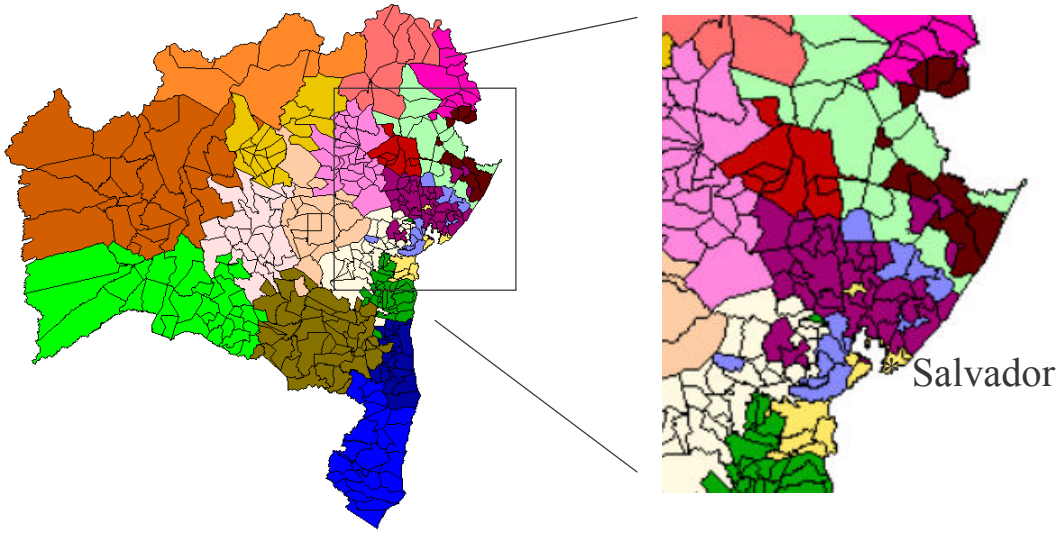


Figura 3.6: Detalhe do resultado do método por componentes ponderadas: peso insuficiente para objetos pequenos (peso 3).

utilizado (vinte) foi elevado, fazendo prevalecer a componente geográfica de forma exagerada, degradando assim, a homogeneidade das regiões.

Ainda utilizando o método por componentes ponderadas, foi realizado um experimento adicional, com o objetivo de obter uma referência para avaliar a degradação da qualidade do agrupamento provocado pelo uso de um peso demasiadamente elevado para a componente geográfica. Foi realizado uma regionalização considerando apenas as coordenadas geográficas, o que é equivalente a tornar nula a componente correspondente aos atributos não-espaciais. O índice de qualidade para tal agrupamento ficou em 903,78. Este valor representa um índice de qualidade muito baixo para a classificação, e serve como um valor de referência inferior para as avaliações dos diversos procedimentos de regionalização. Ele confirma que o peso utilizado no experimento anterior, para que todas as regiões fossem contíguas, acabou por produzir uma partição com pouca homogeneidade e de baixa qualidade.

3.4.3 – AVALIAÇÃO DO MÉTODO AZP

O AZP foi o primeiro método analisado dos dois procedimentos que considera a estrutura de vizinhança diretamente. Este procedimento se enquadra dentro da categoria de métodos de classificação por particionamento ou relocação iterativos, como o próprio *k-médias*, porém é mais complexo, pois considera a restrição de contigüidade durante a sua execução. O procedimento foi implementado conforme descrição na Seção 2.8.1. Neste método, não existe a necessidade de se descobrir qualquer parâmetro, como no caso da abordagem por componentes ponderadas e os resultados, em termos de qualidade da classificação, foram melhores que as soluções apresentadas anteriormente. Porém, o procedimento AZP é bem mais caro, do ponto de vista computacional, que os métodos anteriores (*por componentes ponderadas* e *em duas etapas*). Além disso, o método AZP se mostrou muito sensível a escolha da partição inicial, gerando resultados visualmente bem distintos, para execuções sucessivas, utilizando o mesmo conjunto de dados e partições iniciais diferentes. Isto mostra que o método AZP é sensível a ótimos locais. Para contornar este problema, Openshaw et al. (1998) propõem que se execute o AZP algumas vezes, para então, escolher a melhor partição. Porém, isto pode ser muito demorado para estudos envolvendo muitos objetos.

As melhorias propostas por Alvanides et al. (2002) buscam evitar que o procedimento AZP fique retido em um ótimo local, já que o algoritmo usado é um método descendente (ou subida da montanha). As novas propostas estão baseadas na adoção de outras técnicas de otimização *("simulated annealing" e busca Tabu*). Elas provocarão, sem dúvida, melhorias na qualidade da partição resultante, mas também trarão acréscimos ao tempo de execução do método, porque o processo de busca pela melhor solução deixará de parar na primeira solução encontrada (quando não ocorrer uma melhoria imediata na função objetivo). A Tabela 3.2 apresenta os resultados obtidos para 10 execuções do método AZP, classificando os 415 municípios em 20 regiões, onde pode ser observada uma variação na qualidade das partições obtidas. Os atributos utilizados neste experimento foram os mesmos utilizados nas seções anteriores. A Figura 3.7 apresenta o mapa com o resultado da regionalização que obteve o melhor resultado (qualidade da partição).

Tabela 3.2: Resultados de 10 execuções do método AZP com diferentes partições iniciais (20 regiões).

dados: 415 municípios da Bahia; atributos: porcentagem de área com lavoura, pastagem e matas.

execução	tempo	Q(p)
1	2789	400,50
2	2540	395,47
3	2988	396,02
4	2115	404,79
5	2511	443,58
6	2746	392,43
7	2963	390,19
8	2098	430,22
9	3075	464,18
10	2015	405,77
média	2584	412,32

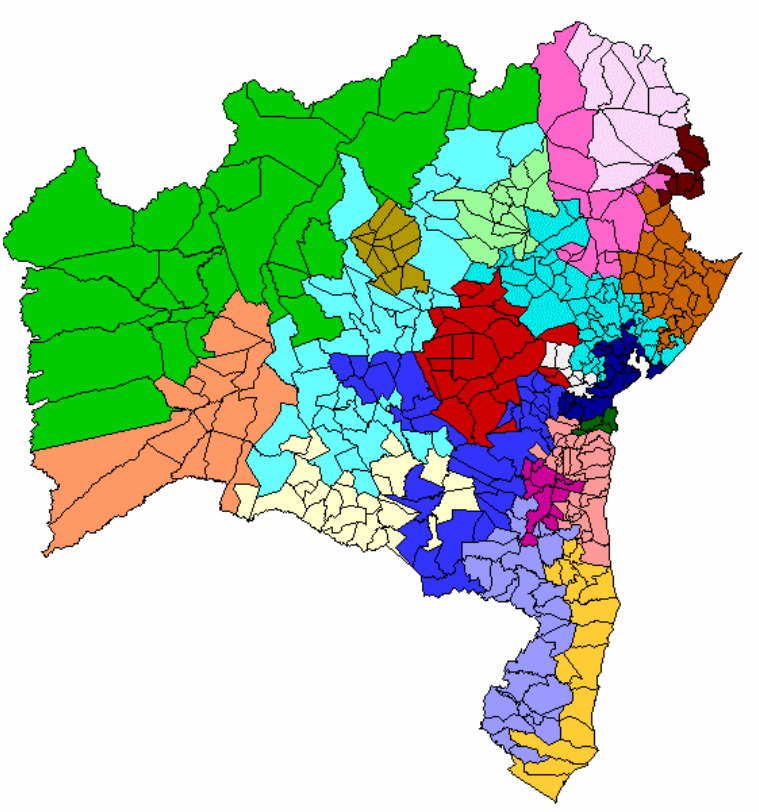


Figura 3.7: Melhor partição (20 regiões) produzida em dez execuções do método AZP.

3.4.4 – AVALIAÇÃO DO MÉTODO VIA AGM

O último método de regionalização avaliado foi o procedimento que utiliza a árvore geradora mínima como um passo intermediário do processo de regionalização. Assim como o AZP, este método utiliza diretamente a estrutura de vizinhança dos objetos. Em vários testes realizados, o método mostrou ser capaz de produzir uma classificação de boa qualidade em um tempo significativamente menor que o método AZP. A Figura 3.8 mostra a classificação obtida pelo método (20 regiões) para o mesmo conjunto de dados utilizados nas análises anteriores. Esta classificação obteve um índice de qualidade de:

$$Q(\Pi_{AGM}) = 380{,}32 \quad .$$

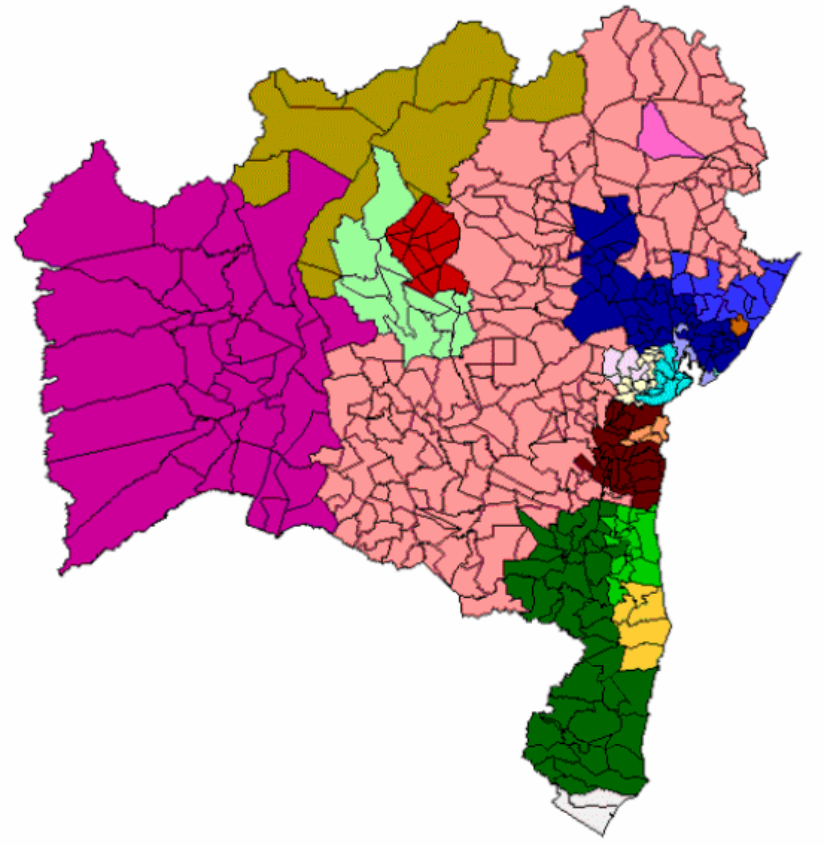


Figura 3.8: Partição resultante (20 regiões) do método baseado no uso da AGM.

O valor para o índice de qualidade do procedimento baseado na AGM é ligeiramente melhor (menor) que o melhor resultado das 10 execuções do método AZP. Quanto ao tempo de execução, o método consumiu 627 segundos, incluindo o tempo gasto para a geração da AGM, que foi de 33 segundos. Este valor corresponde a menos de 5,3% do tempo global do método, que indica que o custo computacional do algoritmo de Prim é relativamente baixo. Com isto, o método baseado no uso da AGM gastou menos da quarta parte (24,4%) do tempo médio gasto pelo procedimento AZP. Em relação à abordagem por componentes ponderadas, o método é evidentemente mais lento, porém, a qualidade da partição é significativamente superior.

Uma outra qualidade do método baseado no uso da AGM é a sua praticidade, pois ele não precisa de qualquer ajuste, como demandava a abordagem por componentes ponderadas; há controle sobre o número de regiões e o resultado final não é dependente da escolha de uma partição inicial, como ocorre com o AZP.

3.5 - COMPARAÇÃO FINAL

A Tabela 3.3 agrupa os resultados para os métodos testados. O método *em duas etapas* não foi incluído nesta tabela já que não foi possível gerar um resultado com o mesmo número de regiões pretendido, e será comparado em outro teste. A Figura 3.9 mostra um diagrama com os valores dos índices de qualidade dos resultados obtidos pelos métodos sobre uma mesma escala. O valor máximo da escala de qualidade corresponde ao método de classificação *k-médias*, sem restrição de contigüidade. Este valor funciona como um limite superior para o índice de qualidade, onde nenhum método de regionalização consegue atingir, pelo simples fato da existência da restrição de contiguidade. Para o valor mínimo da escala de qualidade, utilizamos o valor obtido anteriormente, para o método por componentes ponderadas, anulando a componente dos atributos não-espaciais. Os resultados utilizados na construção desta escala de qualidade não podem ser generalizados. A rigor, eles só valem para o experimento realizado. Mas, de qualquer forma, eles servem para mostrar as tendências gerais de comportamento e situar a qualidade alcançada pelos métodos nos testes em relação a algumas referências.

Tabela 3.3: Comparação entre os métodos: componentes ponderadas, AZP e via-AGM

dados: 415 municípios do estado da Bahia, atributos: id_Past, id_LPerm, e id_LTemp. Classificados em 20 regiões.

	Ponderado	**AZP** (média de 10)	**AZP** (melhor)	**AGM**
qualidade	732,45	412,32	390,19	380,22
tempo de execução	< 1 *	2584	2963	383

* tempo de execução é medido em segundos e por valores inteiros.

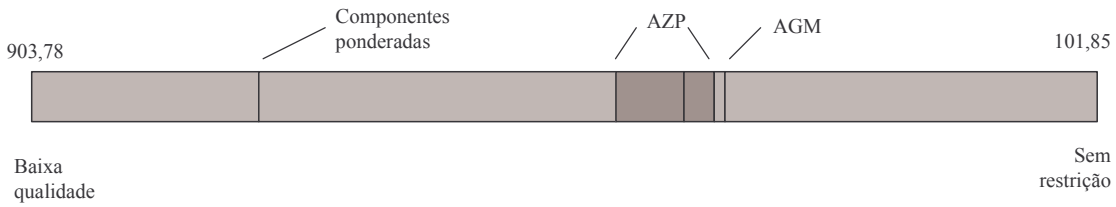


Figura 3.9: Escala relativa, com os resultados obtidos com os diversos métodos.

Por fim, para avaliar a qualidade do agrupamento produzido pelo método *em duas-etapas*, já que este foi excluído da comparação anterior, realizamos uma série de procedimentos de regionalização com os métodos *em duas-etapas* e baseado no uso da AGM. Os números de regiões utilizados para o método via-AGM foram escolhidos de modo que igualasse ao número de regiões produzidas pelo método *em duas-etapas*. A Tabela 3.4 mostra os resultados para o índice de qualidade em todos os procedimentos. Os valores para o índice de qualidade das partições produzidas pelo método via-AGM foram significativamente melhores, em todas as situações experimentadas.

Tabela 3.4: Comparação entre os métodos duas etapas e baseado na AGM.

regiões	**$Q(\Pi_{2etapas})$**	**$Q(\Pi_{AGM})$**
27	915.07	316.203
54	382.16	187.96
66	284.83	156.18
92	210.32	103.83

3.6 – CONCLUSÕES DO CAPÍTULO

Os resultados apresentados pela abordagem em duas etapas mostram que esta solução é pouco viável para procedimentos de regionalização, atendendo, talvez, a alguma aplicação com objetivos específicos. O método não fornece qualquer controle sobre o número de regiões produzidas ao final do processo. A qualidade da partição é questionável, pois outros métodos, quando utilizando o mesmo número de regiões, fornecem particionamentos consideravelmente de melhor qualidade, do ponto de vista da homogeneidade interna das regiões.

A abordagem ponderada tem como melhor característica a rapidez com que o resultado é apresentado. No experimento, gastou menos de um (1) segundo. Além disso, não demanda a estrutura de vizinhança dos objetos. Porém, a qualidade do agrupamento deixou a desejar. Outro inconveniente está na determinação de um peso para as componentes.

O método AZP, um dos dois métodos analisados que utilizam diretamente a estrutura de vizinhança dos objetos, apresentou uma qualidade razoável para os particionamentos. Porém, se mostrou sensível à escolha da partição inicial e a ótimos locais. Um outro aspecto negativo do AZP é o tempo de execução, consideravelmente maior que todas os demais procedimentos analisados.

O método baseado no uso da AGM, que como o AZP utiliza diretamente a estrutura de vizinhança dos objetos, apresentou uma qualidade do agrupamento ligeiramente superior aos resultados obtidos pelo AZP e significativamente melhor que os demais métodos. Com relação ao tempo de execução, é mais lento que as abordagens ponderada e em duas-etapas, mas se mostrou mais rápido que o AZP (cerca de quatro vezes mais rápido, no experimento comparativo).

Um aspecto relevante a favor do método baseado na AGM é a sua praticidade. O único parâmetro a ser definido pelo usuário - além da escolha dos atributos, é claro – é o número de regiões a serem produzidas. Não havendo necessidade de executar o método

várias vezes, como no AZP (para a escolha da melhor partição) e no método por componentes ponderadas (para definir um peso para as componentes).

CAPÍTULO 4

MÉTODO BASEADO NA AGM E OTIMIZAÇÃO

No Capítulo 2 foram apresentados as abordagens e alguns métodos utilizados em procedimentos de regionalização. Vimos que dois métodos apresentados, o AZP e o via-AGM, utilizam diretamente a estrutura de vizinhança dos objetos espaciais. Estes procedimentos apresentaram melhores resultados que os demais, quanto à qualidade da partição, como pudemos verificar nos experimentos realizados no Capítulo 3, porém, com um custo computacional mais elevado. Uma parte dos procedimentos de regionalização, que consideram a estrutura de vizinhança, utiliza grafo para representar a conectividade entre os objetos. O método específico, via-AGM, reduz o grafo de conectividade a uma árvore de custo mínimo. Esta estratégia visa diminuir a complexidade associada ao problema de particionamento de grafo. Este método apresentou um bom resultado, sendo significativamente mais rápido que o método AZP. Porém, no processo de poda da AGM, o método realiza uma busca exaustiva, analisando cada uma das arestas pertencentes à arvore para, ao final de uma iteração, identificar a aresta de custo mais elevado. Neste capítulo, descrevemos o desenvolvimento de um método alternativo, que emprega técnicas de otimização como forma de agilizar a fase de poda da AGM, procurando a aresta de custo mais elevado sem analisar todas as possibilidades e assim oferecer uma alternativa mais viável para ser aplicada em problemas com um número elevado de objetos.

O capítulo está dividido na seguinte forma: Na Seção 4.1, é apresentado o procedimento de regionalização proposto. A Seção 4.2 detalha a estratégia de busca pela melhor divisão de uma região, caracterizando a fase de poda da AGM como um problema de otimização. A adoção de um ponto de partida adequado para a estratégia de exploração é discutida na Seção 4.3. O procedimento de busca completo é apresentado

na seção seguinte (4.4). A Seção 4.5 apresenta uma avaliação do comportamento do método proposto, utilizando como referência o método original. Por fim, são apresentadas as conclusões relativas ao capítulo.

4.1 - MÉTODO PROPOSTO

O método via AGM, apresentado na Seção 2.8.2, possui duas fases bem definidas. Na fase inicial, a partir do grafo de conectividade e dos valores dos coeficientes de similaridade entre os objetos, é gerada uma árvore de custo mínimo. A partir daí, o procedimento de regionalização é tratado como um problema de particionamento de árvore. O procedimento de regionalização que sugerimos neste trabalho, assim como o método original, após a fase de geração da AGM, é um método de classificação do tipo hierárquico por divisão. Inicialmente todos os objetos, representados por vértices da árvore, estão conectados e pertencem a um único grupo (ou região). Então, a cada iteração do método, ocorre uma divisão de uma região em duas, até que o número de regiões previamente estipulado seja atingido. A divisão de uma região ocorre pela eliminação de uma aresta na AGM. Para gerar uma partição de n objetos em k regiões, é necessário eliminar sucessivamente $k-1$ arestas da AGM.

À medida que arestas da AGM original são eliminadas passa-se a ter um *conjunto de árvores desconexas*, sendo que cada árvore do conjunto tem uma correspondência unívoca com uma região. Para representarmos o processo de regionalização utilizamos no método três estruturas de dados: a estrutura *Região (R)*, que contem um conjunto de dados sobre uma região, relevantes para o processo de regionalização; a *Lista de Regiões* (L_{Reg}), que agrupa os dados sobre todas as regiões; e o grafo *CAD*, que representa o conjunto de árvores desconexas (Figura 4.1).

Em cada árvore do grafo CAD existe sempre uma aresta que, ao ser eliminada, melhor dividirá a região correspondente, separando os objetos em dois grupos mais homogêneos. Em cada iteração, o método escolhe dividir a região que proporcionará um maior ganho na qualidade geral da partição, seguindo o mesmo critério adotado para o método apresentado na seção 2.8.2:

$$Q(\Pi) = \sum_{i=0}^{k} SQD_i \ , \tag{4.1}$$

onde:

- Π é uma partição dos objetos em *k* regiões.
- Q(Π) é um valor associado à qualidade de uma partição Π.
- SQD_i é a *soma dos quadrados dos desvios* da região *i*.

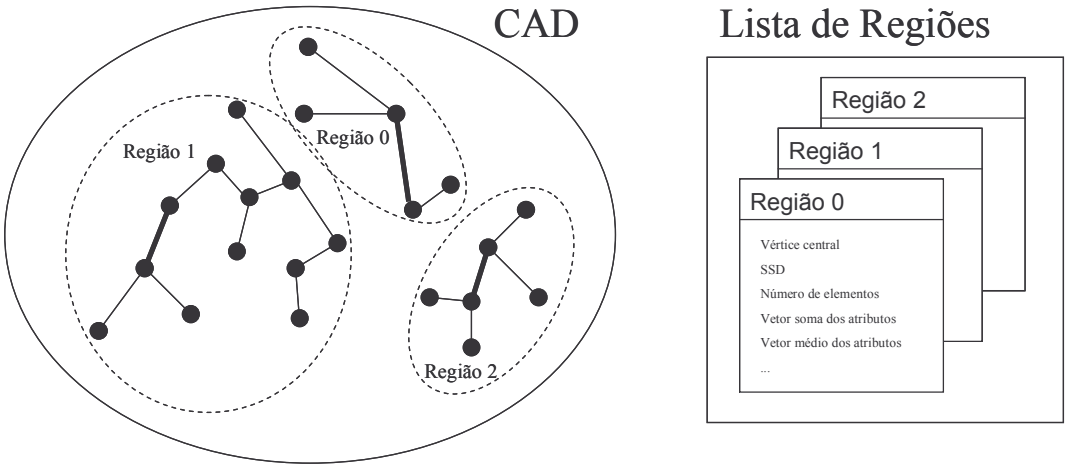


Figura 4.1: Estrutura de dados utilizadas no procedimento.

A escolha da melhor divisão de uma árvore é o problema elementar neste método. Para uma região *R*, a sua subdivisão corresponde a uma eliminação de uma aresta da sua árvore correspondente. Desta forma, para a árvore correspondente à região *R*, A_R *(V, L)*, existe um um conjunto de *n* vértices *(V)*, um conjunto de *n-1* arestas *(L)*, e também, um conjunto *S* , de *n-1* possíveis soluções.

$$S = \{S_1, S_2, ..., S_{n-1}\} \quad ,$$

onde, cada solução S_l corresponde à retirada da aresta *l* da árvore. Cada solução de *S* é avaliada pela função objetivo *f* , conforme a seguinte expressão .

$$f(S_l^R) = SQD_R - (SQD_{Ra} + SQD_{Rb}) \ , \tag{4.2}$$

onde:

- S_l^R é a solução proporcionada pela retirada da aresta *l* da árvore correspondente à região *R*.

- *Ra* e *Rb* são as duas regiões derivadas da divisão de *R*, decorrente da retirada da aresta *l* .

A solução que melhor subdivide a região é a solução ótima S_*^R. Tanto S_*^R como $f(S_*^R)$ são tratados como parâmetros da região *R*. Ao final de uma iteração do método, é decidido qual região efetivamente sofrerá a divisão, escolhendo a região R_i que apresentar o maior valor para $f(S_*^{Ri})$. Quanto maior for o valor de *f*, melhor será o ganho na qualidade geral da partição. Para identificar a solução ótima internamente em cada região é realizada uma estratégia independente de busca, investigando as arestas da árvore. O que difere o método baseado no uso da AGM, apresentado no Capítulo 2, do método agora apresentado é, fundamentalmente, a estratégia de busca pela melhor divisão de uma árvore. Abaixo, serão apresentados os passos do método de regionalização, sem detalhar a estratégia de busca, que será mostrada na próxima seção. A partir da geração da AGM, os passos seguintes do processo de regionalização são:

Passo 1: Iniciar o *CAD = AGM*; Criar a região R_0 correspondente a AGM; identificar a aresta que melhor divide a região, S_*^{Ro} , e incluir a região R_0 na lista de regiões L_{Reg} ; fazer *numReg = 1*.

Passo 2: Enquanto *numReg* < *k*, executar passos de 3 a 5.

Passo 3: Escolher uma região R_i em L_{Reg}, com a maior $f(S_*^{Ri})$; Apagar R_i da lista L_{Reg} e apagar aresta correspondente a S_*^{Ri} em CAD.

Passo 4: Para as duas novas *Regiões* , R_a e R_b, criadas a partir da divisão de R_i: identificar as futuras divisões (S_*^{Ra} e S_*^{Rb}).

Passo 5: Inserir R_a e R_b em L_{Reg} e incrementar *numReg*.

A Figura 4.2 ilustra o funcionamento do método para um caso hipotético, mostrando as primeiras iterações onde duas arestas foram eliminadas do grafo CAD, resultando em 3 regiões. Na primeira iteração só existe uma árvore em CAD (equivalente à AGM) e uma região (R_o) em L_{Reg}. Então, é selecionada a solução S_*^{Ro} e a região é dividida (*passo 3*) retirando a aresta correspondente à S_*^{Ro} de CAD. A retirada da aresta dá origem a duas novas regiões, R_1 e R_2. A região R_0 é eliminada, enquanto que R_1 e R_2 são incluídas em L_{Reg}. Para as novas regiões são calculados os parâmetros e determinadas as soluções correspondentes às futuras subdivisões. A segunda iteração começa no *passo 2*, a região R_2 é escolhida no *passo 3* e ela sofre o mesmo processo de divisão, resultando em R_3 e R_4.

O armazenamento dos dados referentes à solução ótima, S_*^R e $f(S_*^R)$, para cada região criada no processo é necessário, porque estes parâmetros permanecem válidos para uma determinada região, até que esta região sofra, de fato, uma divisão. O método baseado no uso da AGM implementado e utilizado para os testes no capítulo 3, utilizou a estrutura de dados apresentado nesta seção. Porém, para avaliar S_*^R e $f(S_*^R)$, ele investigava todas as arestas da árvore correspondente a R. Com o objetivo de reduzir o número de soluções investigadas, procuramos descobrir a solução ótima para a divisão de uma região sem analisarmos todas as arestas da árvore correspondente. A redução do número de avaliações é conseguida com a utilização de uma heurística que conduz à exploração do espaço de soluções.

4.2 – ESTRATÉGIA DE BUSCA PELA MELHOR DIVISÃO DE UMA ÁRVORE

Para agilizar a definição da aresta que melhor divide uma região, executamos uma busca interna na árvore correspondente à região, com técnicas de otimização combinatória. Visto como um problema de otimização, a procura pela aresta que melhor subdivide uma árvore com n vértices, é equivalente a busca pela solução ótima em um espaço de soluções $S = \{S_1, S_2, ..., S_{n-1}\}$.

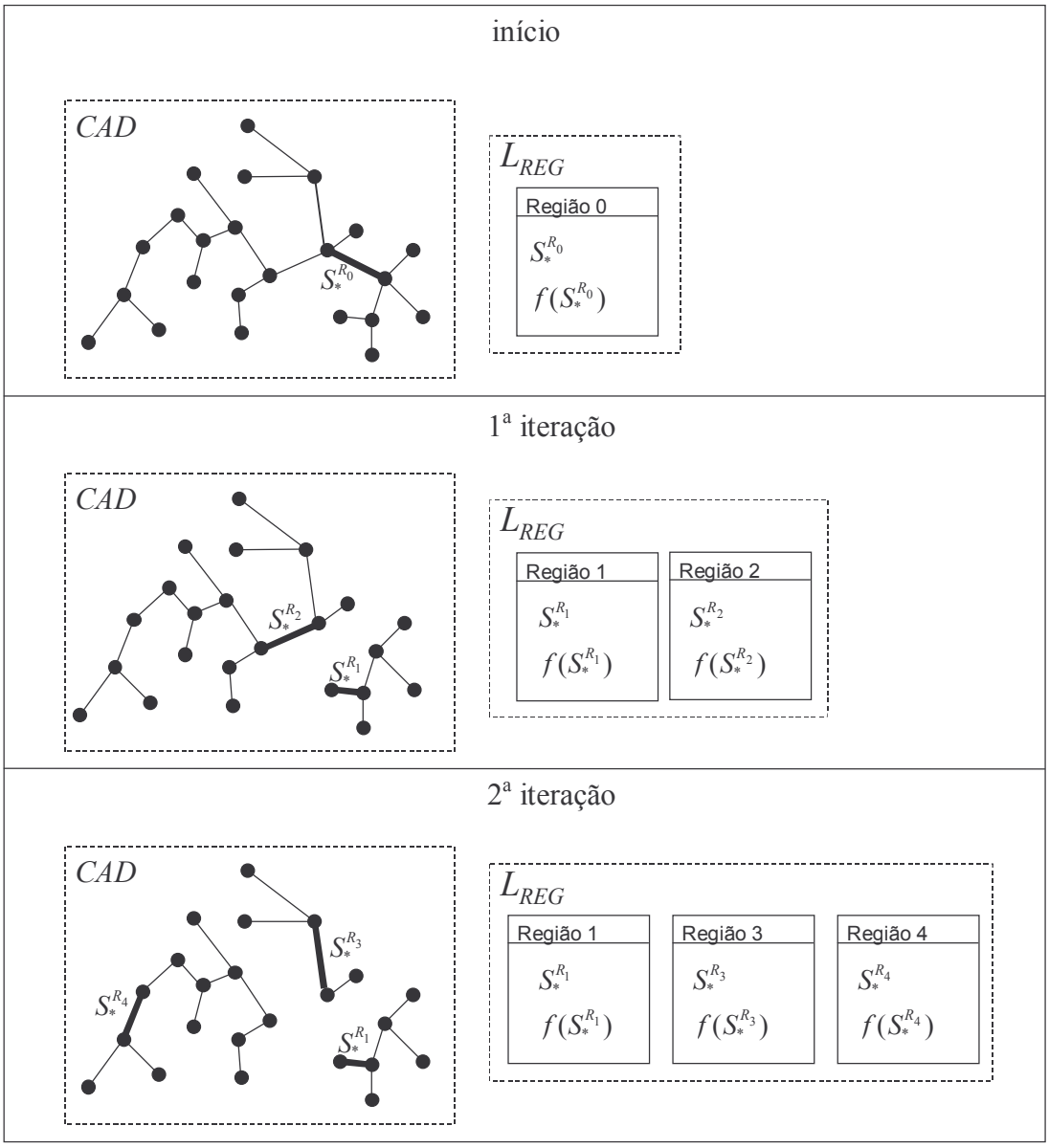


Figura 4.2: Evolução do procedimento.

Na estratégia de exploração do espaço de soluções são utilizados elementos dos métodos de otimização *Expansão pela Vizinhança* e *Busca Tabu* (Laguna, 1994). A estratégia de busca, através da expansão pela vizinhança, conduz a busca pela solução ótima analisando a vizinhança das soluções já visitadas (avaliadas). Em nosso caso, o espaço de soluções é uma árvore homóloga à árvore correspondente a uma região. Assim, uma solução S_i é vizinha à S_j, se as arestas correspondentes, i e j, são incidentes em um mesmo vértice da árvore. A Figura 4.3 ilustra como é realizada a expansão pela vizinhança no espaço de soluções a partir de uma solução já visitada. Inicialmente, a estratégia parte de uma solução S_i. Esta solução é avaliada e, então, tem sua vizinhança expandida. No exemplo, S_i apresenta quatro soluções vizinhas: S_j, S_k, S_l e S_m. Em um segundo passo, estas quatro soluções são avaliadas e uma delas é escolhida para ter a sua vizinhança expandida e assim sucessivamente. Desta forma, o mecanismo de busca pela vizinhança procura obter valores de $f(S_i)$ cada vez maiores, explorando o espaço de soluções pela vizinhança de soluções já analisadas em direção à solução ótima. Além do mecanismo de busca pela vizinhança, a estratégia de exploração incorpora a possibilidade de investigar um certo número de soluções mesmo que não ocorra uma melhora imediata em f. Esta característica permite que a exploração do espaço de soluções não fique facilmente retida em ótimos locais, mantendo a busca, por um certo número de iterações. O critério de parada (*CP*) da estratégia de exploração é o número máximo de iterações, definido pelo usuário, sem que ocorra uma melhoria no valor de f.

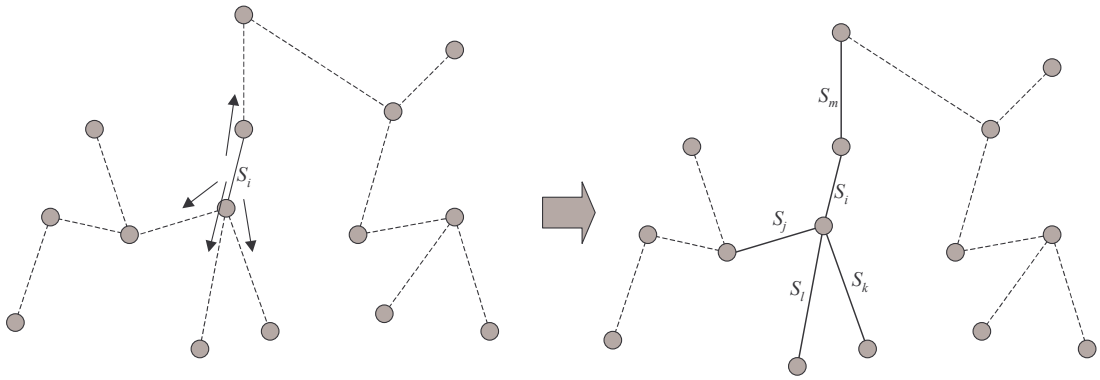


Figura 4.3: Expansão pela vizinhança de uma solução S_i.

Os passos executados pela estratégia de busca são:

Passo1: A partir de uma semente *Vs* (vértice qualquer), inserir soluções correspondentes às arestas incidentes em *Vs* na lista *V**; Fazer $f(S^*) = 0$ e $k^* = k = 0$.

Passo 2: Avaliar as soluções em *V**; armazenar soluções avaliadas na lista *T*; escolher a melhor solução, S_j, em *V**.

Passo 3: Se $f(S_j) > f(S^*)$, então: $S^* = S_j$; $k^* = k$.

Passo 4: Fazer $k = k + 1$. Escolher na lista *T* a solução que terá a vizinhança expandida, gerando um novo conjunto de soluções promissoras (*V**).

Passo 5: Verificar condição de parada ($k - k^* > CP$). Senão satisfeita, voltar ao *Passo 2*.

onde:

- S^* denota a melhor solução encontrada que, ao final do procedimento, representará a solução ótima;
- k registra a contagem de iterações da estratégia de busca;
- k^* indica a última iteração, na qual ocorreu uma melhora no valor de *f*;
- *V** é a lista de soluções promissoras, que serão analisadas na iteração corrente;
- *T* é a lista de soluções avaliadas, mas que ainda não foram expandidas;
- S_j é a melhor solução dentre as soluções promissoras. Esta solução é identificada após serem analisadas todas as soluções em *V**.
- *CP* é o critério de parada da busca, equivale ao número de iterações sem melhoria em *f*.

A escolha de qual solução terá sua vizinhança expandida (passo 4) é realizada através de uma outra função de avaliação, *f'*. O valor de *f'* corresponde ao menor das duas diferenças entre o valor de *SQD* da árvore e os dois valores de *SQDs* das duas subárvores:

$$f' = \min[(SQD_R - SQD_{Ra}), (SQD_R - SQD_{Rb})] \quad .$$

Esta segunda função de avaliação foi criada para evitar que a busca se dirigisse para ótimos locais, freqüentemente presentes em ramos da árvore com um ou poucos elementos. A função *f'* evita que estes ramos sejam verificados, dando prioridade para soluções que dividam a árvore em dois grupos mais homogêneos e equilibrados.

A Figura 4.4 mostra as soluções investigadas em um experimento, para o qual foi escolhido como ponto de partida para a estratégia de busca um vértice distante da solução ótima. Nesta figura, pode-se perceber o caminho percorrido pela estratégia de busca, já que as arestas correspondentes às soluções visitadas estão desenhadas em cor mais escura em relação às demais arestas da AGM, não analisadas. No detalhamento da figura, cada aresta aparece numerada com um valor correspondente à ordem de sua avaliação durante a exploração. A solução ótima foi avaliada na décima quarta iteração correspondendo à vigésima primeira avaliação.

Na Tabela 4.1 são apresentados alguns dados correspondentes às soluções visitadas neste mesmo experimento: os valores de *f* e *f'* utilizados para direcionar o processo de busca, a iteração correspondente, o número de elementos e os valores de *SQD* para cada subárvore. Observa-se na tabela que o procedimento de exploração visita a solução ótima na 13ª iteração, porém a busca prossegue até que a condição de parada seja satisfeita. Neste exemplo, a condição de parada era a ocorrência da 9ª iteração sem melhora na função objetivo *f*.

A Figura 4.5 apresenta um gráfico de barras onde as alturas das barras correspondem aos valores obtidos para a função objetivo *f* em cada iteração do método. Podemos observar um ótimo local neste gráfico, atingido na sétima iteração. Para ultrapassá-lo, a estratégia de busca teve que insistir durante seis iterações até que o valor de *f* voltasse a aumentar. Neste experimento foi utilizado como critério de parada a ocorrência de oito iterações sem melhoria em *f*.

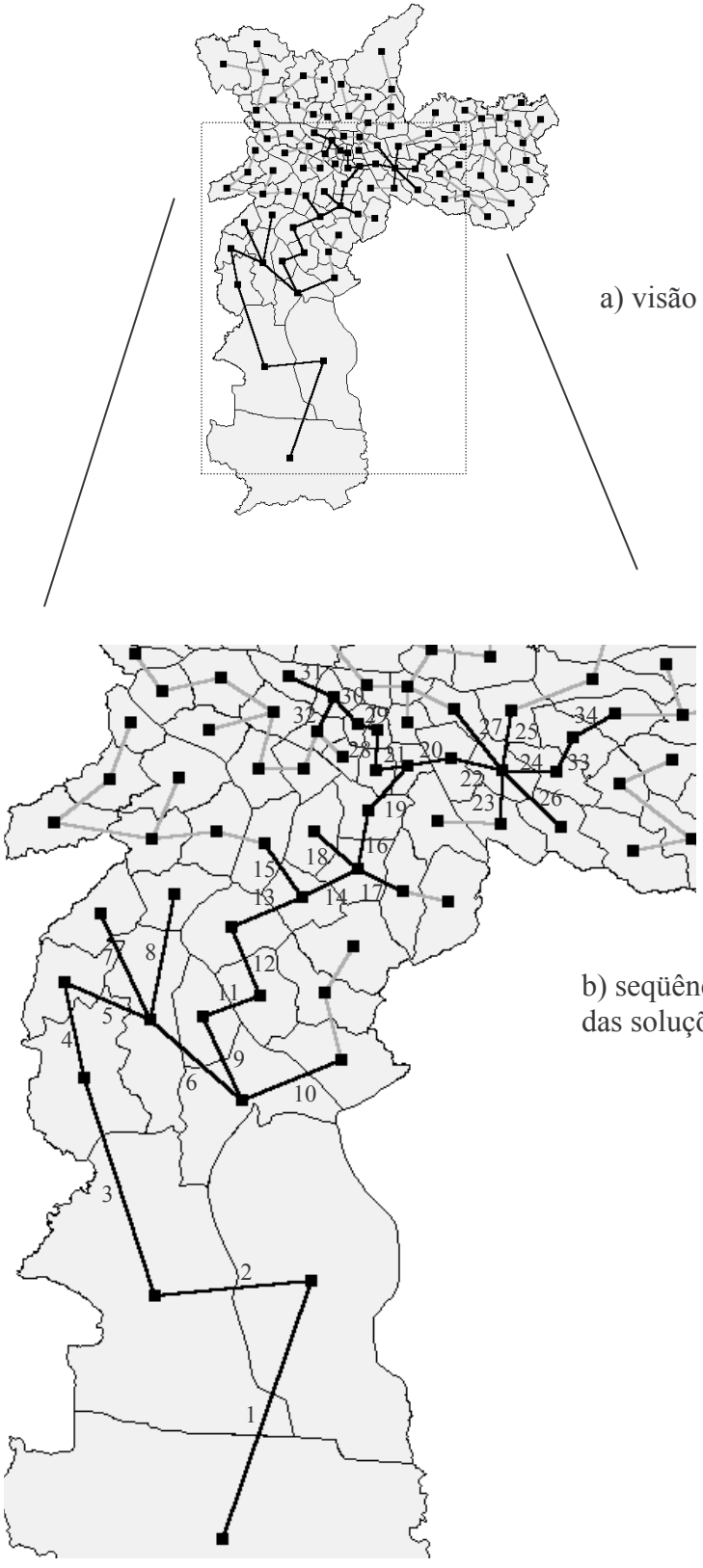


Figura 4.4: Caminho percorrido pela estratégia de busca.

Tabela 4.1: Avaliações realizadas pela estratégia de busca.

dados: 96 distritos do munucípio de São Paulo.

Iteração	aresta	*f*	*f'*	SQD_{Ta}	SQD_{Tb}	N.ELEM.A	N.ELEM.B
0	1	0.41	0.41	0.00	26.19	1	95
1	2	1.03	1.20	0.17	25.41	2	94
2	3	1.75	1.99	0.24	24.61	3	93
3	4	2.36	2.66	0.30	23.94	4	92
4	5	2.74	3.10	0.36	23.50	5	91
5	6	4.18	4.73	0.55	21.87	8	88
	7	0.23	0.23	26.37	0.00	95	1
	8	0.84	0.84	25.76	0.00	95	1
6	9	4.97	5.81	0.84	20.79	12	84
	10	0.51	0.56	26.04	0.05	93	3
7	11	4.76	5.88	1.11	20.73	13	83
8	12	4.30	5.97	1.67	20.63	14	82
9	13	3.52	6.29	2.78	20.31	15	81
10	14	2.08	7.13	5.05	19.47	23	73
	15	0.07	0.57	26.03	0.50	89	7
11	16	0.80	9.06	8.25	17.55	27	69
	17	0.45	0.48	26.13	0.03	94	2
	18	1.24	1.24	25.36	0.00	95	1
12	19	0.58	9.41	8.83	17.19	28	68
13	20	1.40	11.15	15.45	9.75	43	53
	21	5.73	7.65	18.95	1.92	82	14
14	22	1.59	11.07	15.53	9.48	44	52
15	23	0.09	0.09	26.51	0.00	94	2
	24	3.25	5.60	21.01	2.35	73	23
	25	0.44	0.50	26.11	0.05	93	3
	26	0.19	0.19	26.41	0.00	95	1
	27	0.05	4.56	22.04	4.52	74	22
16	28	5.56	7.39	19.22	1.83	83	13
17	29	5.40	7.11	19.50	1.70	84	12
18	30	5.13	6.57	20.03	1.44	85	11
19	31	0.43	0.43	26.18	0.00	95	1
	32	4.20	5.40	21.20	1.19	87	9
20	33	3.45	5.59	21.01	2.14	74	22
21	34	3.82	5.57	21.03	1.75	75	21

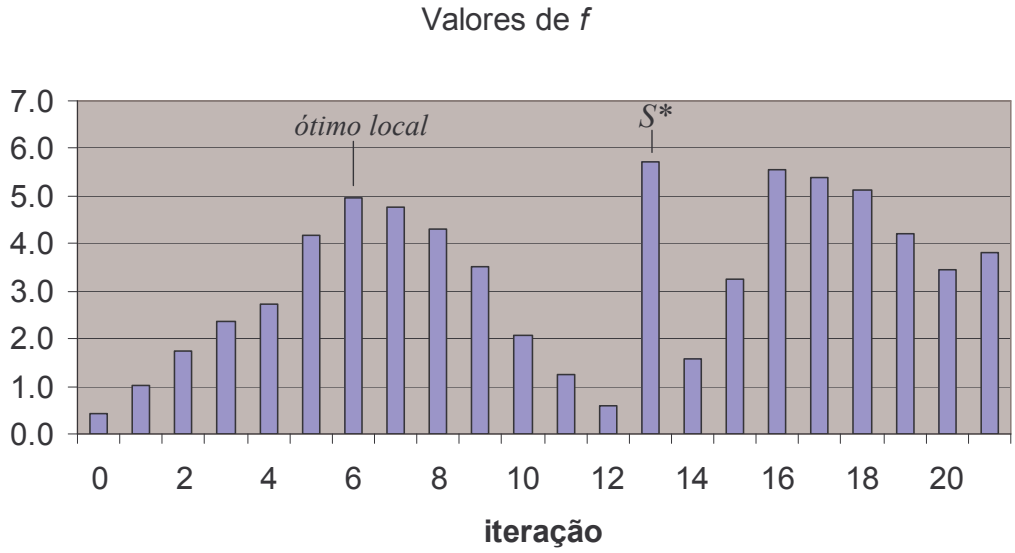


Figura 4.5: Valores da função objetivo, durante a exploração do espaço de soluções.

4.3 – DEFINIÇÃO DE UM PONTO DE PARTIDA CONVENIENTE

No experimento associado à Figura 4.4, foi escolhido como semente, para início do procedimento de exploração, um vértice da AGM distante da solução ótima. Esta escolha visava verificar o comportamento geral da estratégia de busca e sua capacidade de fugir de ótimos locais. Em um caso real, a escolha de um bom ponto de partida pode reduzir o número de iterações necessárias para a localização da solução ótima. Considerando a característica dos métodos *hierárquicos por divisão*, um ponto de partida conveniente é um vértice posicionado ao centro da árvore. Para determinar o nó central é realizada uma outra exploração, percorrendo as arestas até identificar a aresta que melhor divide a árvore em duas subárvores de tamanhos semelhantes. O nó central é um dos dois nós que são ligados por esta aresta, preferencialmente o nó da subárvore com maior número de elementos. Para agilizar a determinação da aresta incidente no nó central utilizamos uma nova estratégia de busca, mais simples, cuja a função objetivo é minimizar a diferença entre o número de nós das duas subárvores.

Como para esta função objetivo não existem ótimos locais é utilizada uma estratégia baseada no *método descendente*[1].

A função objetivo para localizarmos a aresta central é:

$$f'' = \dim(T_a) - \dim(T_b),$$

onde:

dim(T) é uma função que informa o número de vértices de uma árvore *T*.

Utilizamos esta notação porque ela se aproxima da implementação realizada, que utilizou uma Classe *Grafo*, para representar as diversas estruturas (grafos de vizinhança, AGM, subárvores,...) existentes nos métodos de regionalização desenvolvidos e, para esta classe, existe uma função (*método da classe*) que retorna o número de vértices existentes em um grafo. Em alguns casos, podem haver mais de uma aresta com o mesmo valor de *f''*. Nestes casos, a aresta escolhida será aquela primeiramente identificada.

Nosso objetivo, neste novo processo de busca, é minimizar a função *f''*. Diferente do processo utilizado anteriormente, o valor ideal de *f''* que pretendemos alcançar é previamente conhecido e *f''* assume somente valores inteiros. Se a árvore possui um numero par de vértices, o valor da função objetivo pode chegar a zero. Evidentemente que isto depende das ligações entre os vértices existentes na árvore, ou seja, da estrutura da árvore.

A estratégia para a busca pela aresta central é:

Passo 1: Fazer o valor inicial de *f''** grande, *f'' * = dim(AGM)*.

Passo 2: Escolher uma aresta qualquer como a primeira solução, S_i. Insirir S_i na lista de soluções vizinhas (*Lv*).

---

[1] O *método descendente* (ou *subida da montanha*, nos casos de maximização) é um método simples de otimização, onde a busca é interrompida quando nenhuma melhoria imediata no valor da função objetivo é encontrada.

Passo 3: Avaliar f'' para todos as soluções em Lv e escolher a melhor solução S_j.

Passo 4: Se $f''(S_j) < f''(S^*)$, fazer $S^* = Sj$; Expandir a vizinhança de S_j, criando uma nova lista de soluções vizinhas, voltar ao passo 3.

Passo 5: Senão, $f''(S_j) \geq f''(S^*)$, ir para o Passo 6.

Passo 6: A partir da aresta central, definir o vértice central.

No Passo 6, uma vez identificada a aresta central, são consequentemente identificados os dois nós nos quais esta aresta incide. É escolhido como vértice central aquele que pertencer a sub-árvore com maior número de elementos. Em casos onde as subárvores têm o mesmo número de elementos, o nó central será definido arbitrariamente. A Figura 4.6 ilustra o procedimento para uma pequena árvore, onde o processo de busca foi iniciado pela aresta *1* e a Tabela 4.2 mostra os valores de $f''(S)$ e $f''(S^*)$ para todas as iterações do procedimento.

Neste pequeno exemplo, o ganho em relação a avaliação das arestas (8 avaliadas em 14 possíveis) não justificaria a utilização de uma estratégia de busca, mas em problemas envolvendo uma árvore com muitos vértices, o ganho é significativo. Também é importante realçar que os custos computacionais desta estratégia, relativos ao gerenciamento do processo de busca e avaliação das soluções, são mais baratos que os correspondentes na busca pela aresta mais cara, utilizada na fase de poda da AGM. Assim, o custo em localizar o vértice central, é compensado com a economia realizada no procedimento de poda da AGM.

4.4 – ESTRATÉGIA DE BUSCA DEFINITIVA

A Figura 4.7 mostra o caminho realizado pelo procedimento de busca para o mesmo problema anterior, porém partindo do vértice central. A solução ótima é encontrada na terceira iteração, correspondente à oitava solução avaliada. Foi utilizado, neste segundo caso, um critério de parada equivalente a 4 iterações sem melhoria em f.

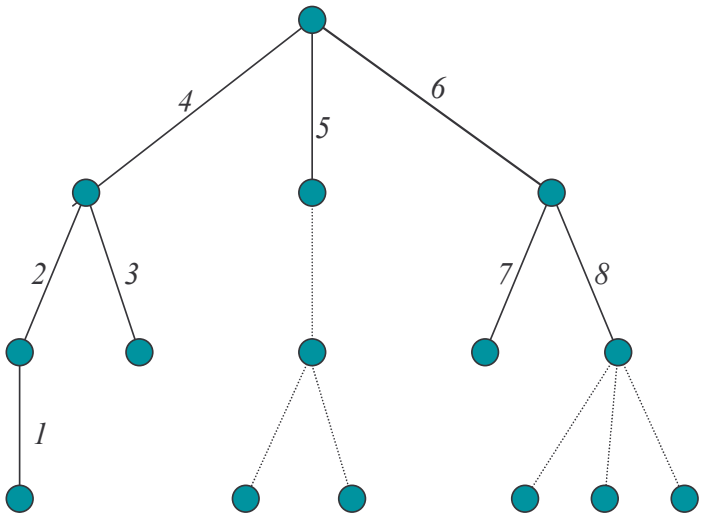


Figura 4.6: Estratégia de busca para a aresta central.

Tabela 4.2: Evolução da busca pela aresta central.

solução	Iteração	dim(Ta)	dim(Tb)	*f´´(S)*	*f´´(S*)*
0	0	-	-	-	15
1	1	1	14	13	13
2	2	2	13	11	11
3	3	1	14	13	
4	3	4	11	7	7
5	4	11	4	7	
6	**4**	**9**	**6**	**3**	**3**
7	5	14	1	13	
8	5	11	4	7	

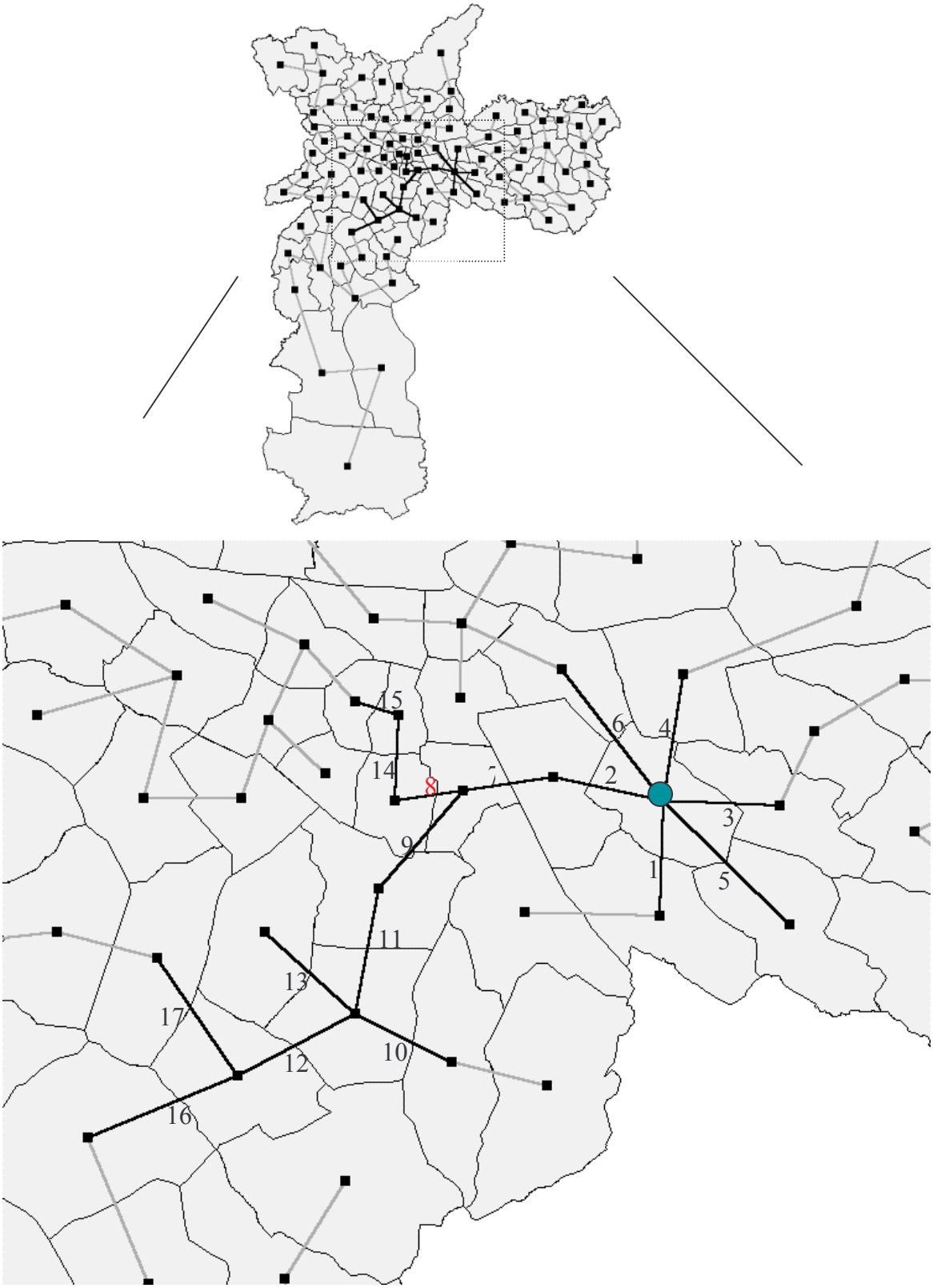


Figura 4.7: Caminho percorrido pela estratégia de busca partindo no vértice central.

A Tabela 4.3 mostra os resultados para este procedimento de busca. Os testes mostraram que a adoção do vértice central da árvore, como ponto de partida do procedimento de busca, tende a diminuir o número de avaliações até a identificação da melhor solução e permite a utilização de CP menores. A alteração na estratégia de busca apresentada na Seção 4.3 ocorre principalmente no primeiro passo, onde *Vs* passa a ser o vértice central da árvore. A estratégia de busca pela solução ótima definitiva, considerando como ponto de partida o vértice central, é:

> **Passo 1:** Definir vértice central da árvore (*Vc*), inserir soluções correspondentes às arestas incidentes em *Vc* na lista *V**; fazer $f(S^*) = 0$ e $k^* = k = 0$.
>
> **Passo 2:** Avaliar as soluções em *V**; armazenar soluções avaliadas em *T*; escolher a melhor solução, S_j, em *V**.
>
> **Passo 3:** Se $f(S_j) > f(S^*)$, então: $S^* = S_j$; $k^* = k$.
>
> **Passo 4:** Fazer $k = k + 1$. Escolher na lista *T* a solução que terá a vizinhança expandida, gerando um novo conjunto de soluções promissoras (*V**).
>
> **Passo 5:** Verificar condição de parada ($k - k^* > CP$). Senão satisfeita, voltar ao *Passo 2*.

A diferença básica entre a estratégia acima e a definida anteriormente na Seção 4.3 está no primeiro conjunto de soluções promissoras (*V**) gerado na estratégia de busca, que agora é obtido pelas soluções correspondentes às arestas incidentes no vértice central, as demais modificações foram apenas para adequar a estratégia a esta nova inicialização. Todo o procedimento para a identificação do vértice central, está embutido no *Passo 1* da estratégia acima.

A definição do nó central é um processo preliminar e independente, executado antes do início da estratégia de busca pela aresta mais cara, propriamente dita. Porém, em procedimentos que utilizam busca Tabu, às vezes são empregadas mais de uma função objetivo. Funções mais simples (menos dispendiosas, do ponto de vista

computacional) são utilizadas para que o procedimento de busca se aproxime rapidamente da solução ótima; e funções mais caras, em uma segunda fase, como refinamento do método de exploração. Isto é, de certa forma, equivalente ao estabelecimento do ponto de partida apropriado para, depois, iniciar o procedimento de busca pela aresta mais cara.

Tabela 4.3: Avaliações realizadas na estratégia de busca, partindo do vértice central.

dados: 96 distritos do município de São Paulo.

iteração	aresta	f	f'	SQD_{Ta}	SQD_{Tb}	n.elem.Ta	n.elem.Tb
0	1	0.09	0.09	26.51	0.00	94	2
0	2	1.59	11.07	9.48	15.53	52	44
0	3	3.25	5.60	21.01	2.35	73	23
0	4	0.44	0.50	26.11	0.05	93	3
0	5	0.19	0.19	26.41	0.00	95	1
0	6	0.05	4.56	22.04	4.52	74	22
1	7	1.40	11.15	9.75	15.45	53	43
2	8	5.73	7.65	18.95	1.92	82	14
2	9	0.58	9.41	17.19	8.83	68	28
3	10	0.80	9.06	17.55	8.25	69	27
4	11	0.45	0.48	26.13	0.03	94	2
4	12	2.08	7.13	19.47	5.05	73	23
4	13	1.24	1.24	25.36	0.00	95	1
5	14	5.56	7.39	19.22	1.83	83	13
6	15	5.40	7.11	19.50	1.70	84	12
7	16	3.52	6.29	20.31	2.78	81	15
7	17	0.07	0.57	26.03	0.50	89	7

A Figura 4.8 apresenta todas as fases do processo de regionalização, resultando em uma classificação dos 96 distritos do município de São Paulo em 10 regiões. Foi utilizado o mesmo conjunto de atributos empregado anteriormente. Neste primeiro experimento completo, envolvendo um número relativamente pequeno de objetos, o tempo de execução do método com busca otimizada (CP = 5) foi cerca de 62% do tempo gasto pelo método com busca exaustiva e cerca de 55% do número total de avaliações de soluções.

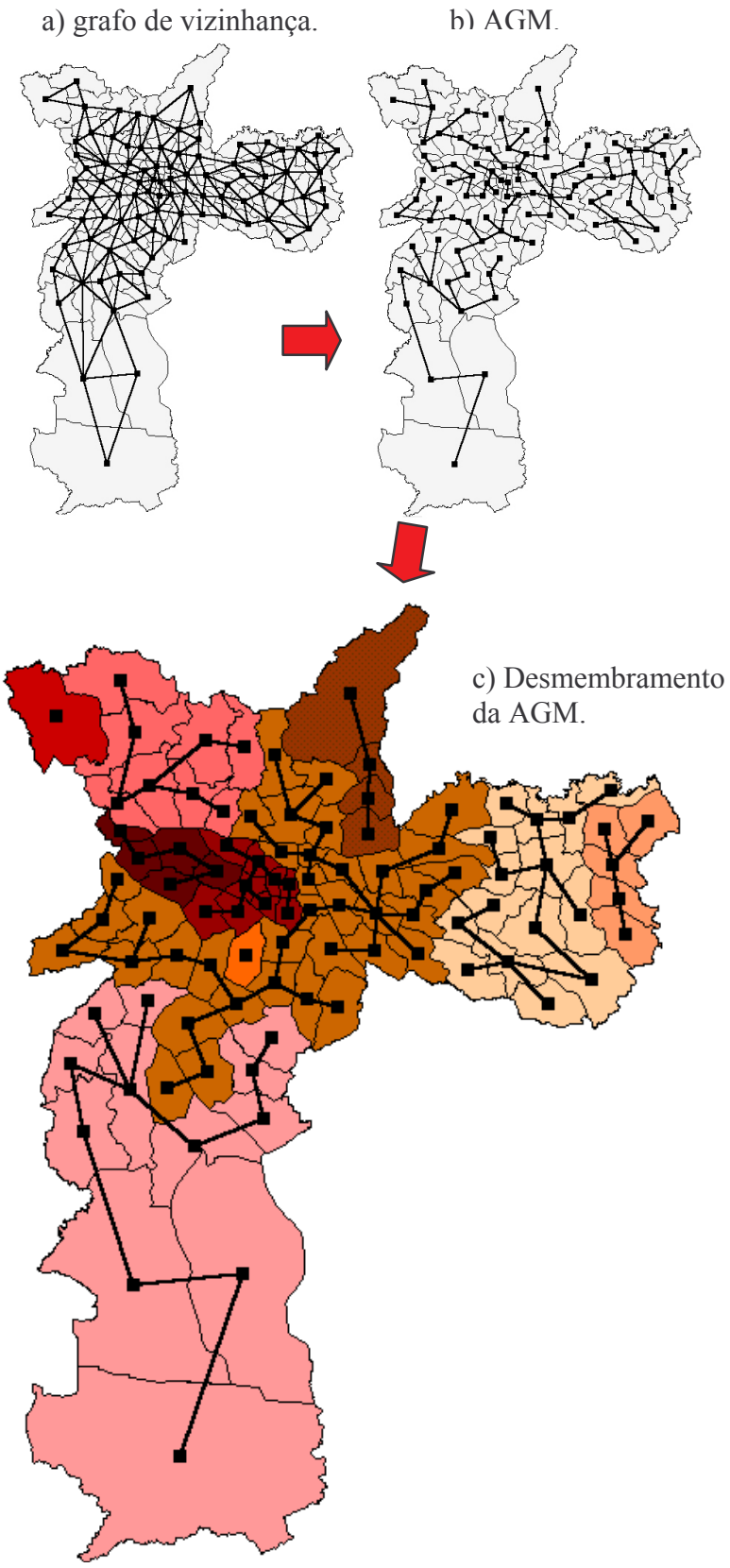


Figura 4.8: Fases do processo: grafo de conectividade, AGM e desmembramento em dez regiões.

4.5 – AVALIAÇÃO DO COMPORTAMENTO DO MÉTODO COM OTIMIZAÇÃO

Para avaliarmos o comportamento do método proposto, realizamos uma série de experimentos, alterando algumas condições de aplicação e utilizando o método original como referência. Nos testes realizados, procuramos variar alguns parâmetros isoladamente, de forma a verificar, com alguma consistência, como cada fator interfere no desempenho do método.

Como o desempenho dos métodos é variável e dependente dos parâmetros e dados utilizados, os resultados obtidos para os testes, aqui apresentados, devem ser vistos como tendências e não como números fixos e verdadeiros para todas as situações. A simples inclusão de um novo atributo na análise, por exemplo, modifica a AGM e, conseqüentemente, todo o processo de regionalização.

Os dados utilizados nos experimentos, que iremos apresentar nas próximas seções deste capítulo, são relativos aos 415 municípios do Estado da Bahia e a um conjunto de três atributos: porcentagem de área do município com pastagem, lavoura e matas. Estes atributos foram gerados a partir de dados do Censo Agropecuário de 1995/1996 do IBGE[2].

4.5.1 – UTILIZAÇÃO DE UM CRITÉRIO DE PARADA ELEVADO

O critério de parada (*CP*) na estratégia de busca significa, como expresso no procedimento apresentados na Seção 4.4, quantas iterações serão realizadas sem que ocorra melhoria no valor da função objetivo. Um *CP* elevado faz com que o mecanismo de exploração do espaço de soluções investigue boa parte das árvores do grafo CAD, procurando pelas melhores soluções, deixando a busca mais robusta à presença de ótimos locais, porém, mais lenta. Portanto, existe uma relação de compromisso entre a qualidade da partição fornecida pelo procedimento e o tempo de sua execução. Como

[2] Os dados do Censo Agropecuário 1995-1996 foram obtidos da Base de Informações Municipais (BIM-IBGE). Eles também estão acessíveis na página: www.ibge.gov.br.

primeiro teste para o método otimizado, utilizamos um critério de parada elevado *(CP = 30)*[3].

A Tabela 4.4 permite verificar o desempenho do método com otimização na fase de poda, utilizando o *CP = 30*, em relação ao método via-AGM com busca exaustiva. São apresentados os resultados obtidos para procedimentos que classificaram os municípios em trinta regiões. Os parâmetros apresentados na tabela são: o índice de qualidade da partição resultante, tempo de execução e número de avaliações. Verifica-se que não houve perda de qualidade na partição, portanto a estratégia de busca identificou as soluções ótimas, nas 29 subdivisões que ocorreram na fase de poda. Porém, os ganhos no desempenho foram relativamente modestos. O procedimento com busca otimizada gastou 58,7% do tempo gasto pelo método original e realizou 56,6% das avaliações necessárias. A queda no tempo de execução não é exatamente na mesma proporção da queda no número de avaliações, porque a estratégia do método otimizado consome um tempo extra para o gerenciamento da estratégia de busca. Estes resultados mostram que para haver ganhos significativos no desempenho precisaremos utilizar valores menores para o *CP*.

Tabela 4.4: Critério de parada elevado (*CP=30*).

dados: 415 municípios do Estado da Bahia em 30 regiões

	Exaustivo	**CP=30**	**%**
tempo	598	351	58,7
número de avals.	4075	2306	56,6
Q(Π)	295.70	295.70	100,0

4.5.2 - INFLUÊNCIA DO CP NO DESEMPENHO DO MÉTODO

Para avaliar a influência do valor adotado para o critério de parada no desempenho do método, foi realizado mais três procedimentos de regionalização, variando os valores de *CP* (*5, 15 e 45*). Os resultados obtidos para estes procedimentos foram reunidos com os

[3] O critério de parada ajustado para 30 iterações é alto em relação aos demais, que foram empregados nos demais experimentos deste capítulo.

resultados obtidos para o experimento anterior e estão apresentados na Tabela 4.5. Para os procedimentos com *CP = 5 e CP = 15*, o índice de qualidade da partição deixou de ser idêntico ao método com busca exaustiva, indicando que o *CP* crítico (em relação à qualidade da partição) está entre 15 e 30, para esta situação. Porém, a queda no índice de qualidade foi pequena se comparada ao ganho no tempo de execução e na queda no número de avaliações executadas.

Tabela 4.5: Efeito da variação do critério de parada.
dados: 415 municípios do Estado da Bahia em 30 regiões.

	CP = 5	***CP = 15***	***CP=30***	***CP=45***	**Exaustivo**
tempo	65	198	351	411	598
	(10,9%)	(33,1%)	(58,7%)	(68,7%)	
número de avals.	720	1597	2306	2621	4075
	(17,7%)	(39,2%)	(56,6%)	(64,3%)	
Q(P)	299.2	295.8	295.7	295.7	295.7
	(101,19%)	(100,03%)	(100%)	(100%)	

A Figura 4.9 mostra em termos relativos, para os quatro valores de *CP*, como é a tendência de variação dos parâmetros (qualidade da partição, tempo de execução e número de avaliações). Analisados simultaneamente, o índice de qualidade da partição praticamente se mantém constante com a variação do *CP*, enquanto que o tempo de execução e o número de avaliações crescem com o aumento do *CP*, ambos tendendo a uma estabilização. A estabilização no número de avaliações ocorre para valores de *CP* exageradamente elevados, para os quais a estratégia de busca investiga muitas soluções, chegando a um número de avaliações próximo, ou mesmo idêntico, ao método exaustivo. Já estabilização para o tempo de execução ocorre em um nível mais elevado que tempo de execução para o método com busca exaustiva, devido ao esforço extra, necessário para o gerenciamento da estratégia de busca.

4.5.3 - VARIAÇÃO DO NÚMERO DE REGIÕES

O experimento anterior mostrou que o desempenho do método é dependente do valor adotado para o *CP*. Para verificar se o desempenho, em relação ao método original, também varia em função do número de regiões pretendidas na análise, foi realizado um

conjunto de classificações, resultando em 10, 30, 50 e 70 regiões, para os dois métodos. A Tabela 4.3 mostra o desempenho dos procedimentos em termos do tempo de execução e do número de avaliações. Para o método com otimização, foram utilizados dois valores de CP (5 e 30).

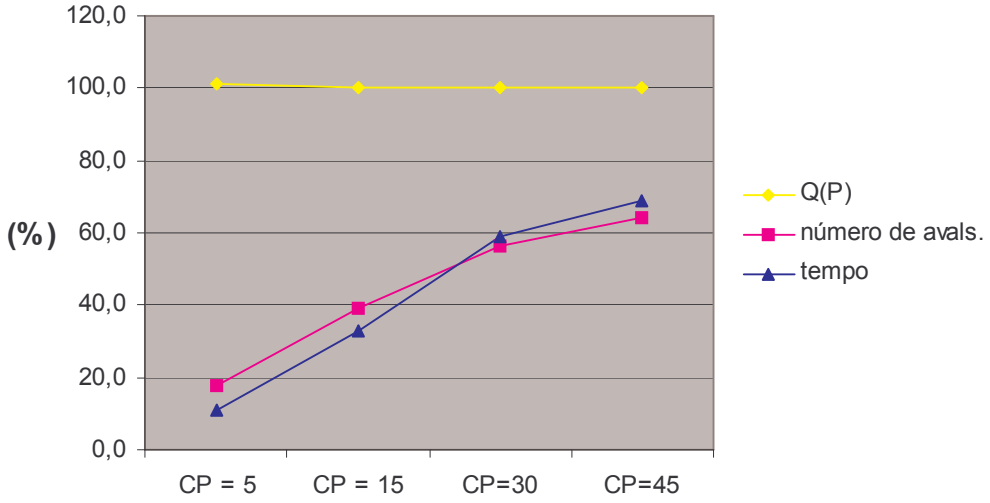


Figura 4.9: Variação de parâmetros em função da variação do *CP*.

Tabela 4.6: Efeito da variação no número de regiões.

dados: 415 municípios do Estado da Bahia; 3 atributos.

regiões	**Otmz(CP:5)**		**Otmz(CP:30)**		**Exaustivo**	
	tempo	num.avals.	tempo	num.avals.	tempo	num.avals
10	46	294	280	1327	571	2964
30	66	720	351	2306	598	4075
50	81	1093	375	2856	617	4625
70	95	1422	393	3293	627	5062

A partir dos resultados da Tabela 4.6 foram gerados duas figuras que permitem comparar graficamente o desempenho em função do aumento do número de regiões. A Figura 4.10 apresenta o desempenho em função do número de avaliações e a Figura 4.11 em função do tempo de execução. Com o aumento do número de regiões a

diferença no número de avaliações continua a aumentar, porém em taxa menores. Isto é explicado pelo processo de desmembramento da AGM em árvores menores, durante o procedimento, que faz diminuir o ganho proporcionado pela estratégia de busca com otimização.

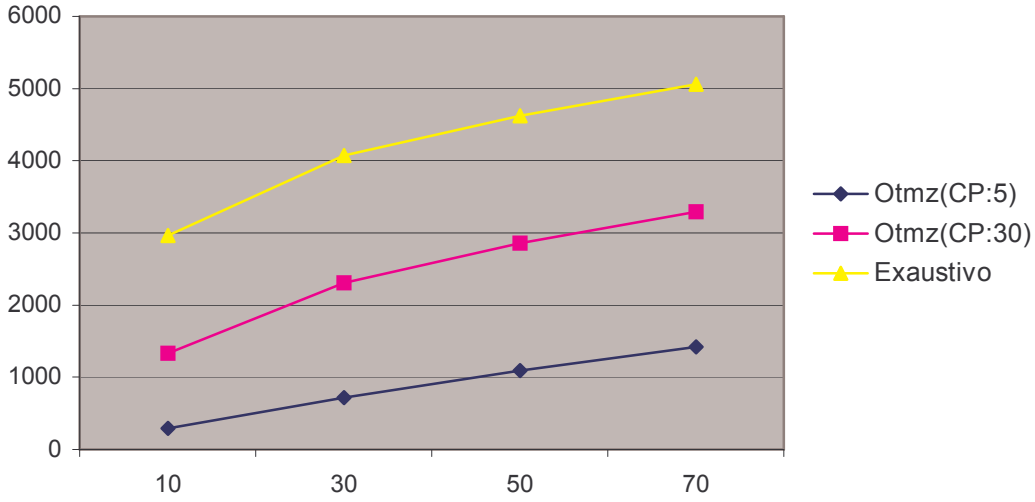


Figura 4.10: Variação do número de avaliações em função do número de regiões.

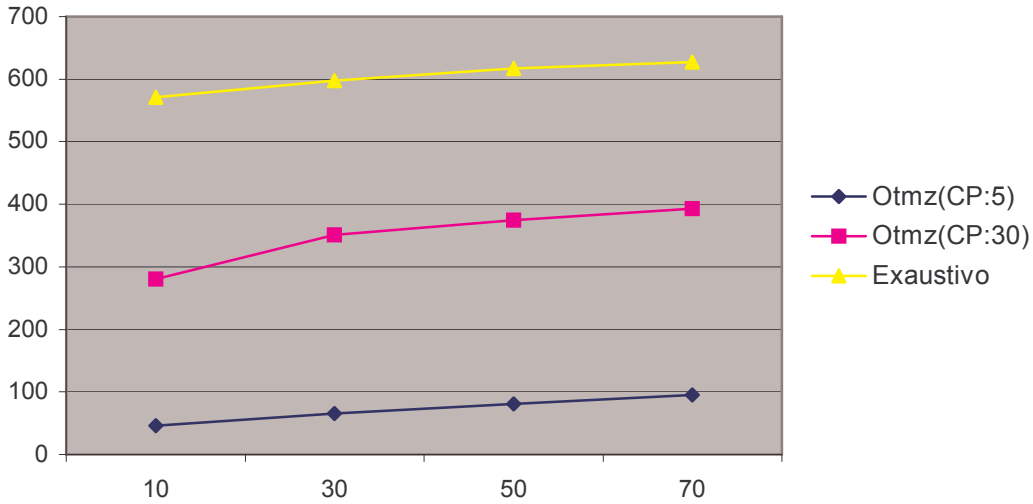


Figura 4.11: Variação do tempo de execução em função do número de regiões.

A diferença quanto ao tempo de execução dos métodos se mantém aproximadamente constante, em função, tanto do desmembramento da árvore, quanto da diminuição do custo em realizar uma avaliação, à medida que as árvores do grafo CAD vão ficando menores. Em termos relativos, o ganho diminui. Para 10 regiões e $CP = 5$, o tempo gasto pelo método com busca otimizada representava cerca de 8% do tempo gasto pelo método com busca exaustiva. Para a classificação em 70 regiões, o tempo relativo foi de cerca de 15%.

4.5.4 - VARIAÇÃO NO NÚMERO DE ATRIBUTOS

Para verificar o efeito do aumento do número de atributos no desempenho do método, outro experimento foi executado. Foram realizadas regionalizações considerando 3, 6 e 9 atributos. Para cada caso, foi medido o tempo de execução. Como a simples inclusão de um atributo na análise modificaria a AGM e, consequentemente, todo o processo de busca, foi usado um artifício para isolar este efeito. Cada um dos três atributos foi replicado duas vezes. Assim, pudemos utilizar nos procedimentos (com 3, 6 e 9 atributos) a mesma AGM para todos os casos.

A Tabela 4.7 mostra os dados referentes ao tempo de execução para os três experimentos, usando os três conjuntos de atributos. Os resultados indicam que existe uma leve tendência de aumento da eficiência do método proposto com aumento do número de atributos na análise. Isto é explicado pelo maior número de avaliações realizadas pelo método exaustivo, que assim, sofre maior impacto do custo mais elevado de cada avaliação executada. O aumento é pequeno porque existem outros custos na avaliação de uma solução que são significativos, como a verificação dos objetos pertencentes a cada subárvore, que não são influenciados pelo número de atributos. A Figura 4.12 mostra os resultados da tabela em uma forma gráfica.

Tabela 4.7: Efeito da variação do número de atributos no tempo de execução.
dados: 415 municípios do Estado da Bahia.

Atributos	Otmz(CP:5)	Otmz(CP:30)	Exaustivo
três	65	351	598
seis	68	376	680
nove	72	394	717

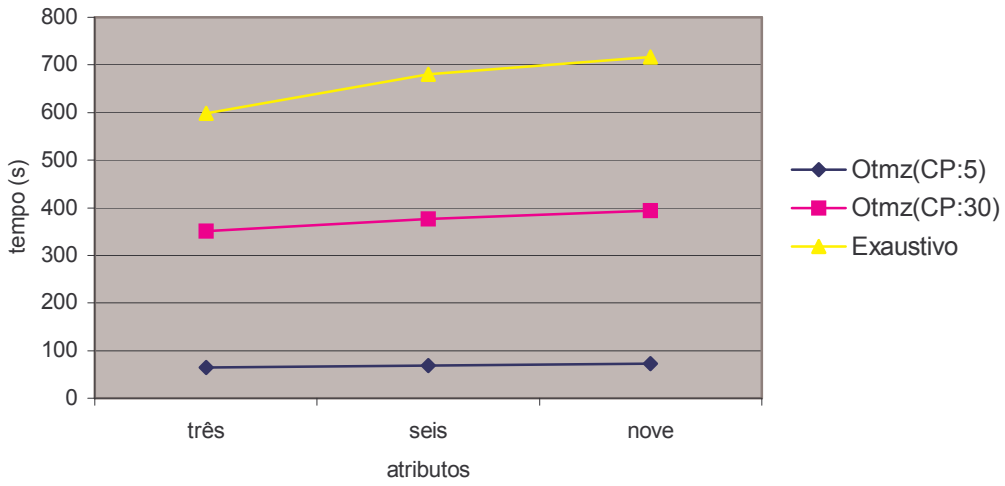


Figura 4.12: Variação do tempo de execução em função do número de atributos.

4.5.5 - VARIAÇÃO NO NÚMERO DE OBJETOS

Não é possível estabelecermos um experimento para avaliarmos diretamente como o aumento no número de objetos influencia no desempenho do método com busca otimizada, pois seriam dois problemas distintos, com duas árvores (AGM) diferentes e, portanto, as comparações seriam inválidas. De qualquer forma, os ganhos de eficiência obtidos para os 415 municípios do Estado da Bahia são mais expressivos que os resultados obtidos para os testes realizados anteriormente com os 96 distritos do município de São Paulo (Seção 4.5). É evidente que em problemas com um maior número de objetos, envolverá a exploração de árvores maiores, e cada solução terá um custo computacional mais elevado, devido ao maior número de objetos. Assim, podemos afirmar que o aumento no número de objetos envolvidos em um procedimento de regionalização tende a aumentar o ganho na eficiência dos métodos que empregam

técnicas de otimização. Esta afirmação também está apoiada nos resultados anteriores, onde percebemos que o método proposto tem ganho de eficiência mais expressivo quando trabalha com árvores maiores, no início do processo de desmembramento da AGM. No Capítulo 6, será mostrado um exemplo envolvendo um grande número de objetos, onde esta afirmação poderá ser verificada.

4.6 – ESCOLHA DO CP

Pelo que vimos na seção anterior, o ponto crítico do procedimento que utiliza a estratégia de busca com otimização é a definição do critério de parada da busca. Um *CP* alto garante que o procedimento execute uma busca ampla pelo espaço de soluções, mas com prejuízos ao desempenho do método. Determinar o valor para o *CP* crítico, de forma a garantir que o resultado seja equivalente ao método exaustivo, executando o menor número possível de avaliações, pode necessitar de várias execuções do procedimento, o que contraria o objetivo do método, que é fornecer uma alternativa rápida para se executar a regionalização. No outro extremo, utilizar um *CP* igual a zero, é equivalente a adotar um *método descendente* (ou *subida da montanha*). Isto faria a estratégia de busca muito sensível a ótimos locais. Por isto, a melhor opção, é a escolha de um *CP* pequeno (não nulo), apenas para garantir que haja alguma exploração adicional do espaço de soluções, após a identificação de uma solução ótima. Com esta escolha, renunciamos a uma busca mais ampla no espaço de soluções, em função de aproveitar melhor a principal característica do método, que é possibilitar uma regionalização mais rápida.

A Figura 4.13 mostra um resultado de um experimento que ajuda a defender o uso de um valor pequeno para o critério de parada. Foi realizada uma série de procedimentos de regionalização, utilizando um $CP = 5$ e sete valores para o número de regiões. Nesta figura, à medida que o número de regiões cresce, os índices de qualidade obtidos pelo método com busca otimizada converge para valores próximos aos obtidos pelo método com busca exaustiva. Isto ocorre porque, mesmo que as regiões geradas não sejam as mesmas identificadas pelo método original, a busca otimizada proporciona uma partição com alguma qualidade. Além disso, à medida que o número de regiões

aumenta, a maioria das regiões são identificadas nos dois métodos, mudando apenas a ordem pela qual são identificadas e alguns objetos isolados que, eventualmente, são classificados de forma diferente, mas que pouco influem no resultado final.

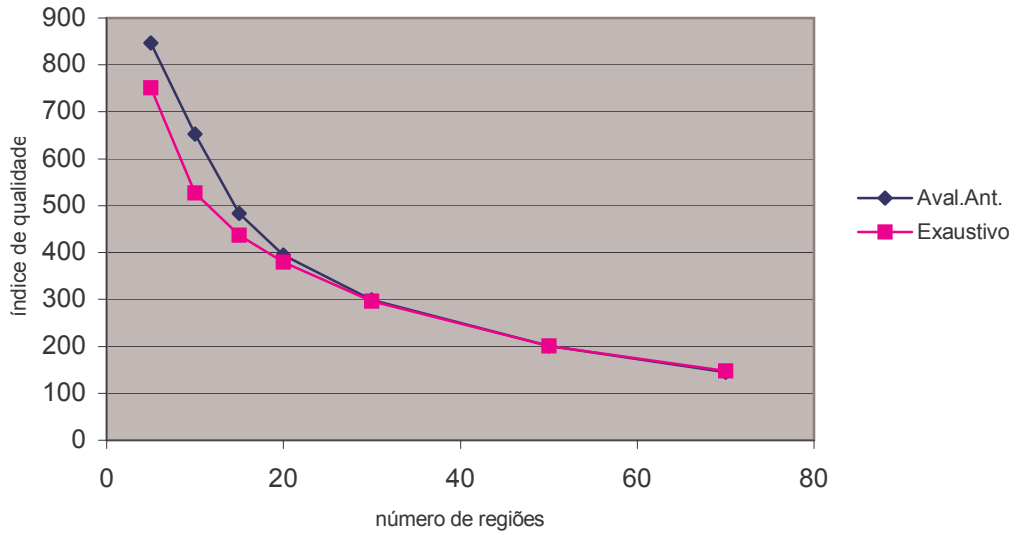


Figura 4.13: Convergência de qualidade com o desmembramento da AGM.

4.7 – CONCLUSÕES DO CAPÍTULO

Com a adoção de técnicas de otimização é possível reduzir significativamente o número de avaliações necessárias para executar a regionalização. Esta redução não provoca necessariamente um piora significativa na qualidade das partições.

De uma forma geral, à medida que o problema de regionalização cresce em complexidade, aumentando o volume de dados (número objetos e atributos) a eficiência do método otimizado torna-se mais significativa. Analisando os diversos fatores isoladamente, temos: i) A utilização de um maior número de atributos, tende a aumentar o ganho relativo (tempo de execução) do método com técnicas de otimização, devido ao aumento no custo das avaliações. Como no método com busca exaustiva o número de avaliações é maior, ele sofre maior impacto; ii) Com o aumento no número de regiões, o ganho relativo do método com otimização tende a diminuir, já que a medida que a AGM original, vai sendo desmembrada em árvores menores, o ganho do método com

busca otimizada é menor. iii) É impossível avaliar precisamente como o crescimento do número de objetos afeta o desempenho, mas certamente, haverá um aumento significativo da eficiência em problemas envolvendo mais objetos, como pudemos observar, comparando os resultados obtidos para os dois conjuntos de dados utilizados nos testes deste capítulo. Isto é explicado pelo maior número de avaliações (provocado pelo aumento no número de arestas) e pelo aumento no custo da avaliação (provocado pelo aumento no número de objetos).

O uso do procedimento de regionalização com estratégia de busca otimizada pode, até mesmo, reproduzir o resultado do método com busca exaustiva, se o critério de parada utilizado for suficiente elevado para garantir uma ampla exploração do espaço de soluções. Porém, se aumentarmos o valor de *CP*, demasiadamente, deixará de haver ganhos significativos na eficiência e, no limite, todas as soluções serão investigadas. Por outro lado, o uso de um *CP* pequeno, leva a um grande ganho na rapidez do método, ainda que possa ocorrer perda na qualidade da partição, mas, como vimos nos experimentos, ela tende a ser relativamente pequena.

CAPÍTULO 5
REGIONALIZAÇÃO DIRIGIDA PELO USUÁRIO

Nos capítulos anteriores, foram mostrados diversos procedimentos automáticos de regionalização. Neste capítulo, é apresentada uma alternativa ao processo automático, onde a regionalização passa a ser vista como um processo interativo, e o analista tem a possibilidade de interferir na sua condução, não somente no início do procedimento, definindo parâmetros de entrada do algoritmo, mas direcionando a sua execução, analisando e comparando regiões, interrompendo o processamento, retroagindo passos e reagrupando os objetos em função dos objetivos de um problema específico. O ambiente de experimentação foi utilizado para testar algumas funcionalidades desta abordagem interativa, como execução *passo-a-passo*, edição de partição e extração de informação sobre as regiões e objetos.

O capítulo está dividido da seguinte forma: na primeira seção, são apresentadas as justificativas para a adoção de uma abordagem de regionalização dirigida pelo analista; a seção seguinte (5.2) apresenta os requisitos básicos que um sistema de regionalização interativo deveria oferecer. A Seção 5.3 apresenta o papel que a AGM desempenha na abordagem dirigida pelo analista, aqui proposta. Na seção seguinte (5.4), é apresentado um exemplo do uso da abordagem interativa, utilizando algumas funcionalidades implementadas no ambiente de experimentação.

5.1 – MOTIVAÇÃO PARA UM PROCEDIMENTO INTERATIVO

Os procedimentos automáticos, como os métodos de regionalização apresentados nos capítulos anteriores deste trabalho, sem dúvida, ampliam o poder de análise, permitindo

classificar objetos espaciais em regiões de uma forma rápida. Porém, os procedimentos automáticos atuam como sistemas fechados, onde o usuário tem acesso apenas aos parâmetros de entrada (escolha de atributos, medida de similaridade, número de classes, etc.) e ao resultado final, a classificação propriamente dita. Esta rigidez dos sistemas automáticos, em certos casos, pode limitar os resultados da classificação, produzindo partições indesejáveis e inadequadas a problemas do mundo real. O acesso restrito aos extremos do processo de regionalização limita também a capacidade de extração de informação do procedimento, por parte do analista.

Alguns problemas reais não possuem uma única restrição, mas um conjunto delas. Este conjunto de restrições pode ser de difícil representação ou grande o suficiente para inviabilizar o processamento dos métodos automáticos. Suponhamos que um problema de regionalização, envolvendo um grande número de objetos, tenha produzido uma partição desequilibrada em termos da população das regiões. O analista resolve então, criar uma restrição adicional, de forma que todas as regiões tenham uma população mínima. Novo procedimento é executado, desde o seu início. Apesar da restrição sobre a população, alguns objetos com população mais adensada, formaram regiões com apenas um membro. Como isto, também não é desejado, o analista resolve adicionar outra restrição, para que nenhuma região tenha menos que um certo número de membros. Ao executar novamente o procedimento, o analista percebe que algumas regiões possuem, agora, uma grande extensão geográfica, e isto também é indesejável, e assim, nova restrição e execução se farão necessárias. Este ciclo (Figura 5.1) envolvendo: execução do procedimento, análise dos resultados, alteração de parâmetros de entrada ou adição de restrições, e nova execução; podem ser demorados e cansativos em função da dimensão dos dados utilizados.

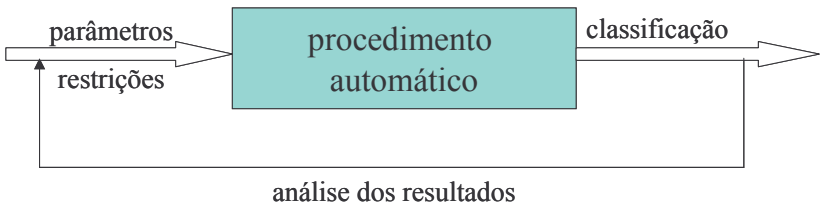


Figura 5.1: Ciclo de execuções para o procedimento automático.

Como alternativa, mas ainda utilizando um procedimento automático, o analista poderia estabelecer um conjunto amplo de restrições, já no início do processo, antes mesmo da primeira execução. Ainda que o analista consiga prever as várias possibilidades de restrições, isto poderia levar a criação de um conjunto exagerado de restrições, acarretando um custo computacional desnecessário ao processo. Existe um outro problema em estabelecer um conjunto amplo de restrições a priori. O analista não tem a percepção dos efeitos que as restrições provocam no resultado da regionalização e isto limita a extração de conhecimento do processo.

Um procedimento de regionalização interativo, dirigido pelo analista, traria uma maior flexibilidade ao processo e permitiria incorporar ao procedimento algumas características humanas, como: capacidade de decisão, consideração simultânea de múltiplos fatores e conhecimento do problema real. No sistema interativo, o analista pode interferir no processo, forçando, por exemplo, a divisão de regiões extensas e fusão de regiões com pouca população. Restrições podem ser estabelecidas durante a execução do procedimento sem a necessidade de recomeçar o processo do seu ponto inicial. A abordagem interativa permite ao analista acesso a informações intermediárias, de forma a auxiliá-lo na compreensão e condução do processo.

A extração de informação é um aspecto relevante a ser considerando, já que muitas vezes o procedimento de classificação, através da Análise de Cluster, é utilizado para extrair estruturas em um conjunto complexo de dados. Nestes casos, o analista executa o procedimento, várias vezes, trocando parâmetros de entrada, com o objetivo de verificar múltiplos aspectos do conjunto de dados. Porém, esta extração de informação poderia ser ampliada se o analista tivesse acesso a algumas informações extras, como a ordem de desmembramento e a homogeneidade de grupos de objetos em diversos níveis de agregação. A própria AGM expressa as relações de similaridade interna entre objetos, criando ramos de similaridade, informação que não está disponível nos mapas com as classes, resultante do processo automático.

5.2 – PROPOSTA DO PROCEDIMENTO DIRIGIDO PELO ANALISTA

Um procedimento de regionalização guiado pela analista pode possibilitar a combinação entre as melhores características humanas e as oferecidas pelos procedimentos automáticos. A abordagem interativa deve oferecer a possibilidade de diversos níveis de interferência do analista, desde uma simples verificação e adequação das classes até uma condução de todo o processo.

5.2.1 – REQUISITOS DO PROCEDIMENTO GUIADO PELO ANALISTA

A seguir listamos alguns requisitos que um sistema voltado para a regionalização dirigida pelo usuário deveria contemplar:

i) **Possibilitar a continuação do processamento a partir de uma partição qualquer:** Esta capacidade permitiria ao usuário escolher um número menor de regiões, analisá-las e verificar se existe a necessidade de criação de novas regiões. Isto traria a possibilidade de uma melhor adequação do resultado ao objetivo da regionalização, sem a necessidade de recomeçar o processo, deste o seu início.

ii) **Permitir a regionalização nos dois sentidos:** A partir de uma partição qualquer, o analista poderia escolher se quer criar novas regiões ou, no sentido inverso, fundir regiões, diminuindo o número final de classes. Isto ampliaria a flexibilidade, já que o processo pode ter gerado, por exemplo, um número exagerado de classes com regiões pouco representativas. Caminhando no sentido oposto, regiões seriam fundidas e este problema corrigido com o aproveitamento do processamento anterior.

iii) Processamento controlado ou *passo-a-passo*: Esta característica permitiria ao analista definir o número de regiões que serão criadas ou fundidas até a próxima parada do sistema e o retorno do controle do processo ao analista. Além do número de classes, outros critérios de parada de execução poderia ser oferecido, como por exemplo: parar o processamento, quando for criada uma região com menos de 10.000 habitantes! Assim o usuário poderia manter o controle sobre o processo de regionalização, aproveitando a

velocidade proporcionada pelo processamento automático, deixando o sistema evoluir, de forma independente, até uma situação pré-determinada.

iv) Interferir na escolha das regiões a serem divididas ou fundidas: Esta funcionalidade permitiria ao analista adequar a regionalização a um problema específico. Ele poderia modificar regiões extensas ou com muitos objetos, forçando sua divisão, ou ainda, escolher fundir regiões pouco representativas de acordo com o objetivo do procedimento.

v) Extração de informação do processo: Para apoiar o analista em sua interferência no processo, o sistema deveria fornecer um conjunto de informações (gráficas e estatísticas) sobre as regiões da partição corrente, ou um conjunto qualquer de objetos.

vi) Armazenar e recuperar partições em disco: Esta característica traria a possibilidade de realizar uma regionalização em diversas sessões de trabalho. Isto seria útil em trabalhos envolvendo um grande volume de dados e também possibilitaria o armazenamento de vários resultados alternativos dentro de uma mesma análise.

5.2.2 – O PAPEL DA AGM NA ABORDAGEM INTERATIVA

Na proposta de regionalização dirigida pelo usuário aqui apresentada, a árvore geradora mínima, ou o conjunto de árvores desconexas (derivado do desmembramento da AGM) tem um papel fundamental, suportando vários dos requisitos listados na seção anterior.

a) Representação do status da regionalização

O grafo CAD e a lista de regiões continuam a representar o status da regionalização, exatamente como no método automático apresentado no Capítulo 4. Agora, porém, o grafo CAD deixa de ser um elemento interno ao procedimento e invisível ao analista e passa a indicar para o usuário o estágio atual da regionalização, de uma forma explícita. Sua estrutura é ainda aproveitada como representação gráfica do processo, podendo o analista sobrepor o conjunto de árvores do grafo CAD sobre o mapa com os objetos, a qualquer momento, ajudando a visualizar as relações de similaridade entre os objetos.

O procedimento de regionalização, conforme mostrado na Seção 4.1, é alterado para iniciar a partir de qualquer conjunto de árvores desconexas (e não somente da AGM, que é uma configuração especifica do grafo CAD) e, assim, atender a possibilidade de execução *passo-a-passo* ou controlada. O grafo CAD, como representação do estágio da regionalização, permite ainda, que sua estrutura seja gravada em disco e recuperada, podendo o processo de regionalização ser conduzido em seções isoladas de trabalho ou produzir e armazenar vários resultados alternativos (um dos requisitos listados anteriormente).

b) Histórico do desmembramento

A AGM original é armazenada durante o processamento para garantir o recomeço do processo e a verificação das arestas "legais", em caso de fusão de regiões. Junto com a AGM, uma lista de arestas eliminadas (ou eventualmente criadas) permite a construção de um histórico do procedimento de regionalização, e isto possibilita ao analista retroagir rapidamente a qualquer ponto do processo, e a qualquer momento. Com estas estruturas, uma função "*desfaz"* (*undo*) em diversos níveis, pode ser facilmente implementada.

Uma outra possibilidade é a geração de um diagrama de divisões das regiões (*dendograma*), que pode ser obtido a partir destas estruturas (AGM e lista de arestas eliminadas), fornecendo estatísticas para vários estágios do desmembramento, funcionando, por exemplo, como uma ferramenta para auxiliar na escolha de um número apropriado de regiões. A Figura 5.2 mostra uma curva com vários valores para o índice de qualidade da partição em função do desmembramento de uma AGM (dados: 96 distritos do município de São Paulo). Em um caso geral, como número de classes vai de a *1* a *n* (sendo *n* o número de objetos), o analista pode escolher o número final de regiões em função do nível de qualidade esperada para a partição, ou em função de um salto significativo na qualidade da partição, ou ainda, em relação a algum outro parâmetro armazenado no processo de desmembramento da árvore.

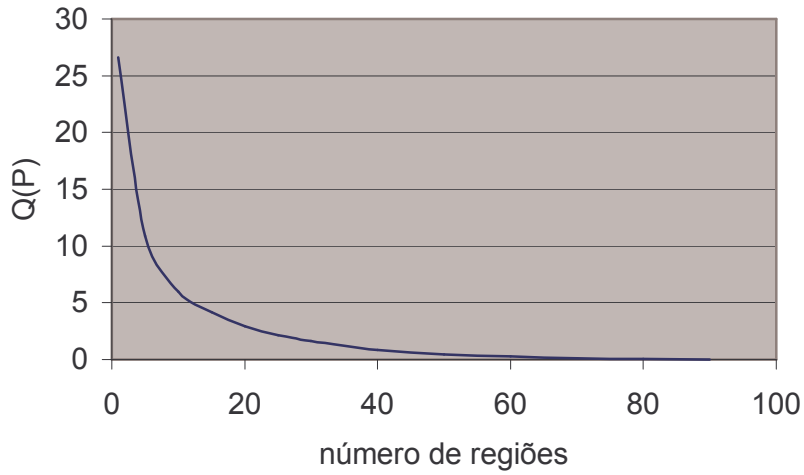


Figura 5.2: Curva do índice de qualidade da partição, durante o desmembramento da AGM.

c) Edição da partição

Um dos papeis fundamentais que o grafo CAD desempenha é suportar a edição da partição. As árvores do grafo CAD são sobrepostas ao mapa de objetos, fornecendo a informação das arestas existentes, permitindo o analista escolher ou experimentar divisões ou fusões de regiões. A AGM é também usada para verificar se uma aresta a ser criada, é uma aresta que existia na árvore original e, portanto, é uma aresta válida.

d) Extração de informação

A AGM executa um outro papel importante, auxiliar na extração de informações, pois ela explicita relacionamentos de similaridade existentes entre os objetos. Na árvore, cada objeto aparece ligado a objetos vizinhos com os quais possui maior similaridade. Assim, a árvore mostra estruturas de similaridade existentes no conjunto dos dados. Estas estruturas podem ser identificadas em vários níveis, deste grandes ramos, contento muitos objetos, a pequenos ramos, com poucos elementos, constituindo assim, uma hierarquia de similaridades.

Além das relações de similaridades entre os objetos, outras informações podem ser extraídas da árvore. Por exemplo: objetos representados por vértices com apenas uma aresta incidente (vértices-folha) tendem a corresponder a objetos com valores de atributos extremados. Por outro lado, objetos representados por vértices posicionados mais ao centro das árvores, com mais de uma aresta são objetos que tendem a ter valores de atributos centrais. A AGM fornece, portanto, informações que não estão visíveis nos mapas em cores, que é a forma tradicional de visualização dos resultados de procedimentos de regionalização. Estas informações adicionais demonstram que a AGM pode ser utilizada como uma ferramenta auxiliar para análise exploratória dos dados e não somente como um passo intermediário e escondido dentro de um procedimento automático.

A chave para que regionalização dirigida pelo usuário forneça resultados atrativos e seja, de fato, uma boa alternativa é disponibilizar o máximo de informações sobre objetos, grupos de objetos e os seus atributos. Neste sentido, a AGM pode ser utilizada para facilitar a escolha de grupos de objetos similares, através da seleção de ramos da árvore. A Figura 5.3 mostra uma seleção de um grupo de objetos (em verde), realizada pela seleção de uma árvore do CAD. Alguns resultados estatísticos, referentes ao grupo de objetos selecionados, são mostrados em uma janela do ambiente de experimentação, exemplificando a extração de informações através da AGM.

5.2.3 – FERRAMENTAS AUXILIARES

Para auxiliar ao analista na tomada de decisões sobre a condução do processo de regionalização, o sistema deve fornecer um amplo conjunto de dados. Nesta seção são sugeridas algumas medidas entre conjuntos de objetos, estatísticas e gráficos básicos que podem atuar como ferramentas de apoio.

Medidas de "distância" (índices de similaridade) entre objetos vizinhos, objeto e regiões vizinhas, e ainda, entre regiões podem ajudar o usuário a decidir sobre fusões de objetos e regiões. Outras medidas como o *SQD*, *SQD-médio* e o número de objetos membros de uma região, podem indicar se uma região é pouco homogênea e deva ser subdividida. Valores referentes a atributos de objetos e regiões (mesmos atributos não

diretamente utilizados na classificação) podem, também, trazer informações importantes para auxiliar o analista e possibilitar comparações entre regiões. Soma de atributos por região (população total de uma região, área total, total de residências), valores médios, e dispersão são fundamentais para a condução do processo e na sua adequação aos objetivos da regionalização.

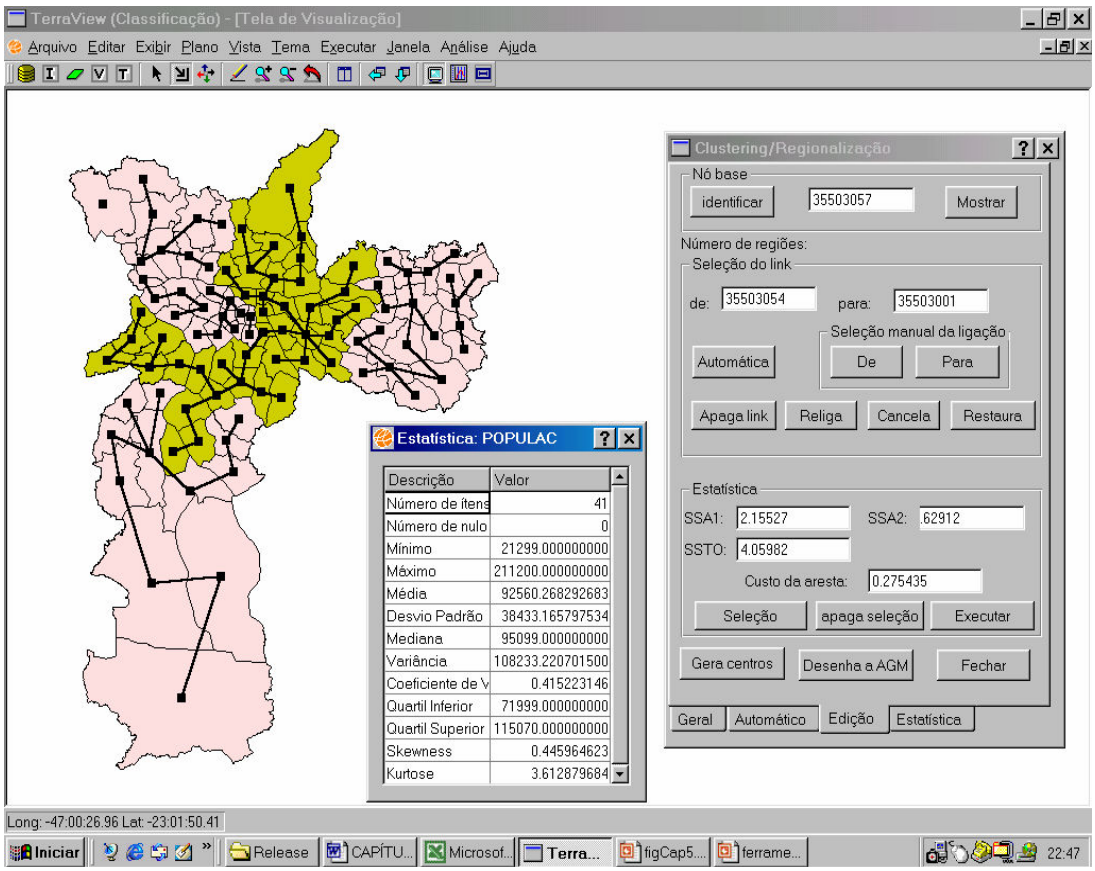


Figura 5.3: Seleção de um grupo de objeto, utilizando uma árvore do grafo CAD.

Gráficos de espalhamento podem indicar se a partição obtida é de boa qualidade e se os objetos membros de regiões estão agrupados no espaço de atributos. Outros recursos gráficos podem auxiliar a extração de informação. A Figura 5.4 mostra um gráfico de "*caixas e bigodes*" onde são mostrados vários parâmetros por atributo, simultaneamente: valores máximos e mínimos, primeiro e terceiro quartil e mediana.

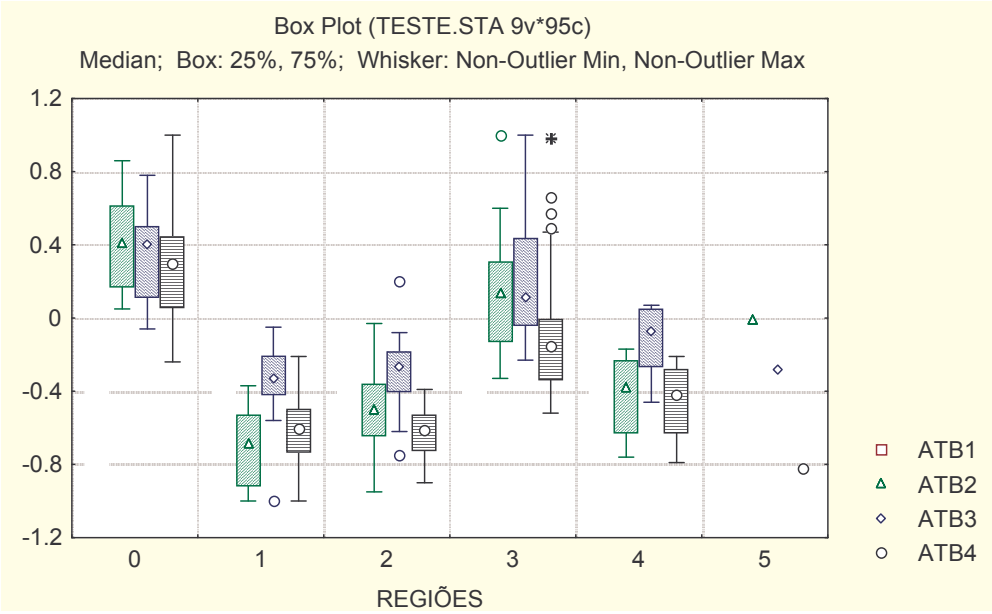


Figura 5.4: Analisando regiões com gráfico *caixas e bigodes*.

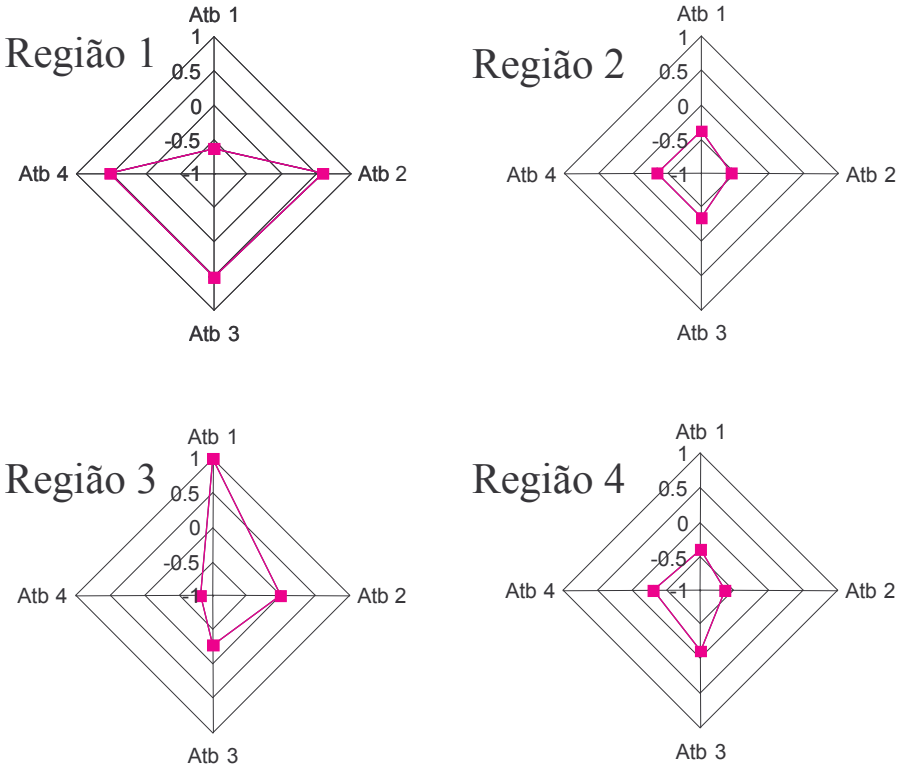


Figura 5.5: Comparando regiões (gráficos *estrela*).

Neste gráfico, os parâmetros para quatro atributos, relativos a seis regiões, são apresentados. Este tipo de representação permite uma análise e comparação entre diferentes regiões, simultaneamente. Neste mesmo sentido, um outro exemplo de representação multivariada é apresentado na Figura 5.5, onde quatro regiões são representadas em um gráfico "*estrela*".

5.3 – EXEMPLO DE EDIÇÃO DA PARTIÇÃO

Para ilustrar a idéia da edição de partição, e do controle do processo de regionalização, é apresentado um pequeno exemplo nesta seção. Utilizando os dados referentes aos 96 distritos de São Paulo e os mesmos atributos utilizados anteriormente (índice de desenvolvimento humano e índice de qualidade de vida). Foi gerado uma classificação em oito regiões. O resultado é apresentado na Figura 5.6, onde alguns rótulos são atribuídos às regiões criadas. Os resultados referentes a esta partição estão expressos na Tabela 5.1. Algumas características negativas nesta partição poderiam se apontadas, devido a discrepância entre algumas regiões criadas. Enquanto que a região *noroeste* possui apenas um objeto membro e este objeto possui uma pequena população de 12.408 pessoas, a região *maior* contem mais de 41 objetos e uma população de quase 3,8 milhões de pessoas, representando quase 40% da população total.

Tabela 5.1: Parâmetros da partição.
dados: 96 distritos do município de São Paulo; Atributos: IEX_EQID e IEX_DSHU.

	membros	SQD	QQD-médio	População
leste	5	0.046	0.009	578030
Meio-leste	15	0.718	0.048	1725189
maior	41	4.06	0.099	3794971
norte	8	0.53	0.066	932714
sul	12	0.838	0.070	1733507
noroeste	1	0	0.000	12408
centro	9	0.591	0.066	583231
centro-oeste	5	0.43	0.086	286135
	Q(P):	7.213	**Pop.Tot.**	9646185

Para ilustrar o processo de edição da partição, vamos mostrar algumas operações realizadas no ambiente de experimentação. Primeiro, as árvores foram sobrepostas às

regiões. Devido a pouca população da região noroeste ela foi fundida com a região norte, através da criação da aresta ligando os objetos 35503003 e 35503045 (Figura 5.7). Devido ao número elevado de objetos e sua grande população, a região *maior* foi escolhida para ser subdividida. Uma busca elementar (automática) na árvore correspondente foi realizada, porém a aresta de maior custo separava a região em duas subárvores contento, uma 40 objetos e outra apenas um elemento (Figura 5.8). Outras divisões foram investigadas e foi escolhida a aresta entre os objetos 35503054 e 35503001 (Figura 5.9) que melhor dividia a região, em relação ao número de objetos, apesar do custo da aresta inferior ao encontrado pela busca elementar (o que resulta em uma piora na qualidade geral da partição, em relação a homogeneidade interna dos grupos). A partição definitiva é mostrada na Figura 5.10 e a Tabela 5.2 apresenta os parâmetros para as regiões resultantes. O processamento interativo abre a possibilidade que a decisão sobre a divisão de regiões considere outros fatores, em função dos objetivos do analista em problemas específicos, adequando o resultado do procedimento.

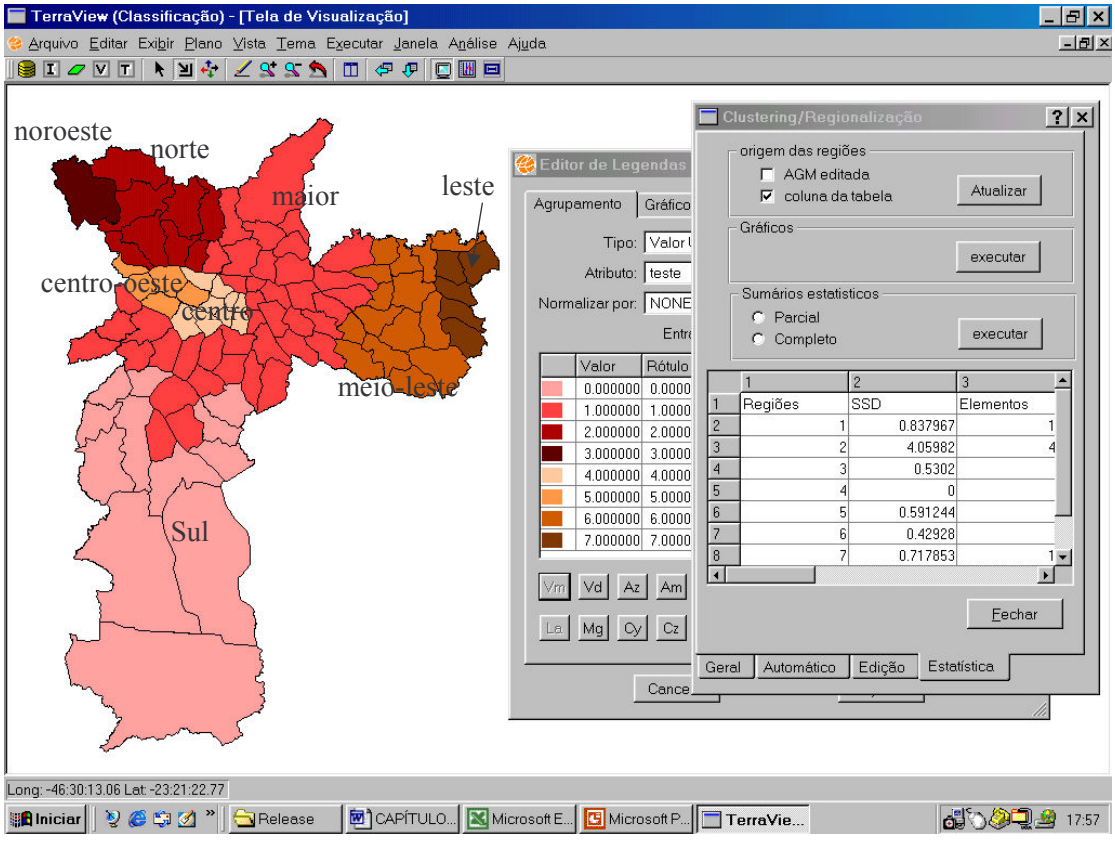


Figura 5.6: Classificação em oito regiões.

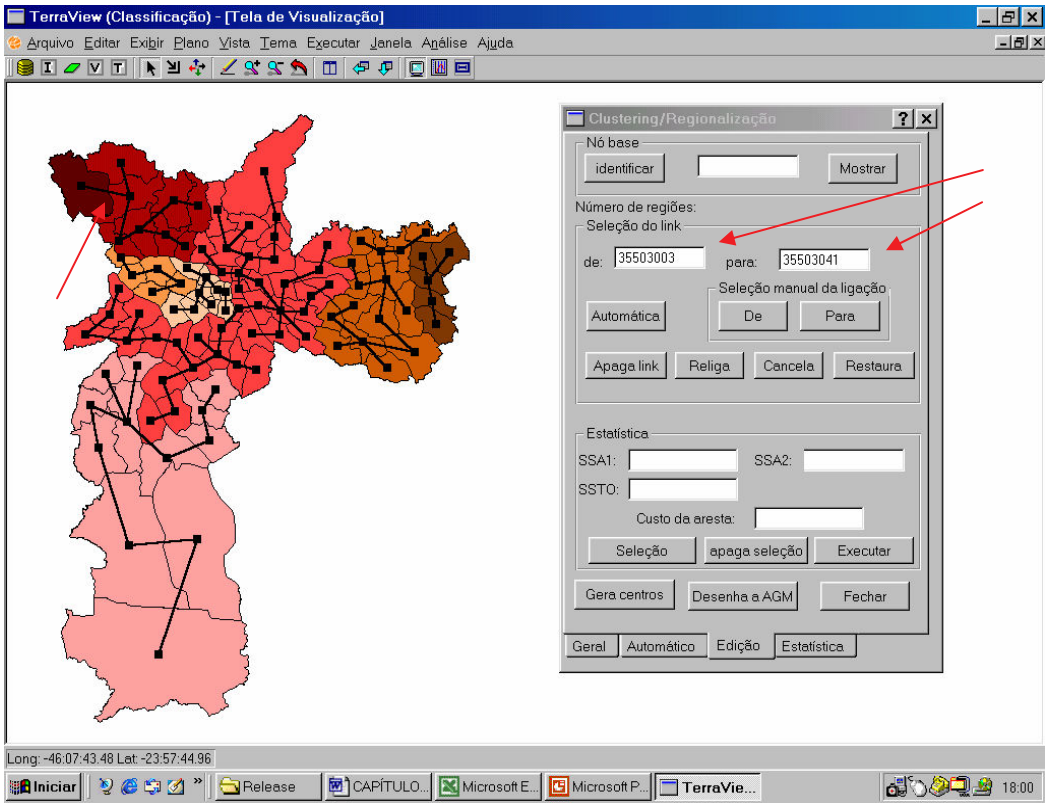


Figura 5.7: Edição de uma partição - fusão de regiões.

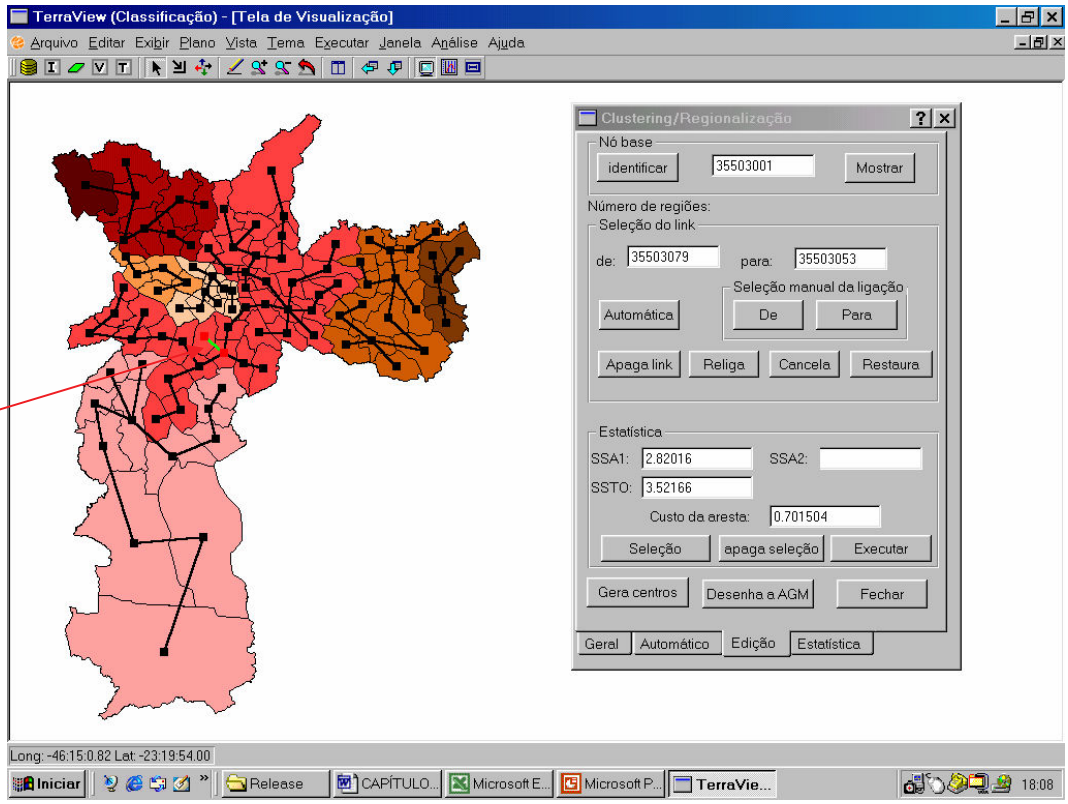


Figura 5.8: Edição de uma partição – busca elementar em uma região.

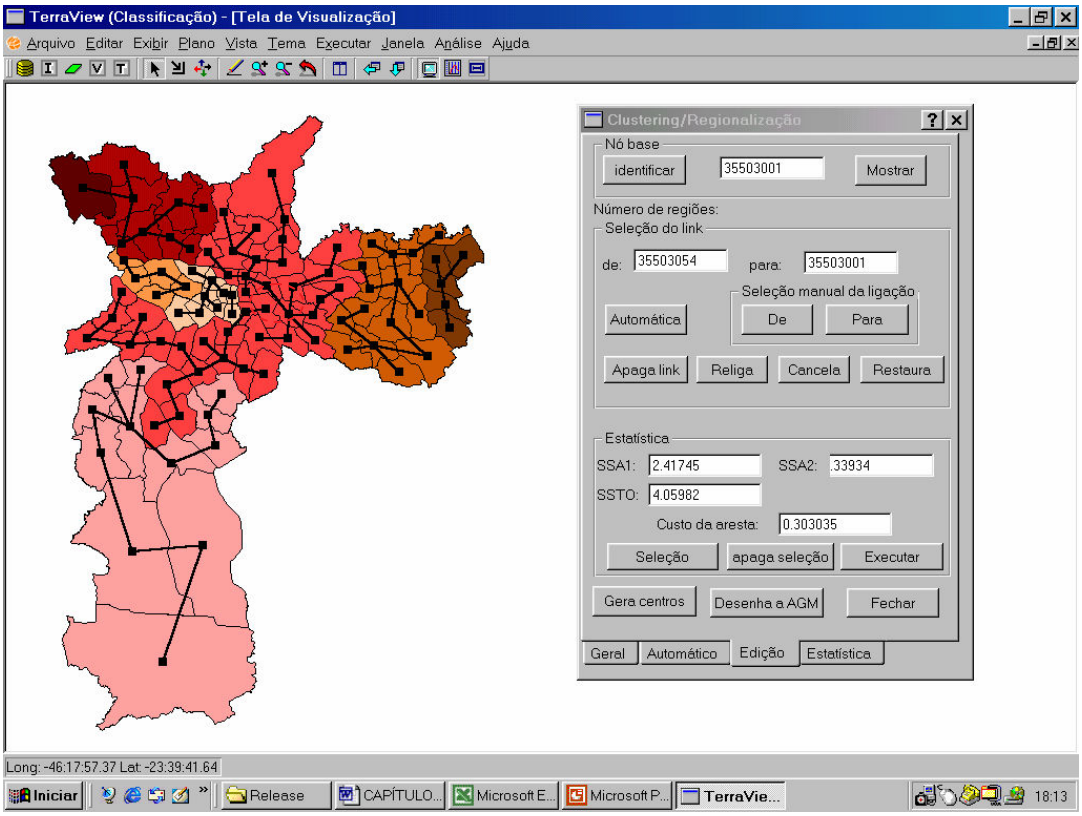


Figura 5.9: Edição de uma partição – desmembramento "forçado".

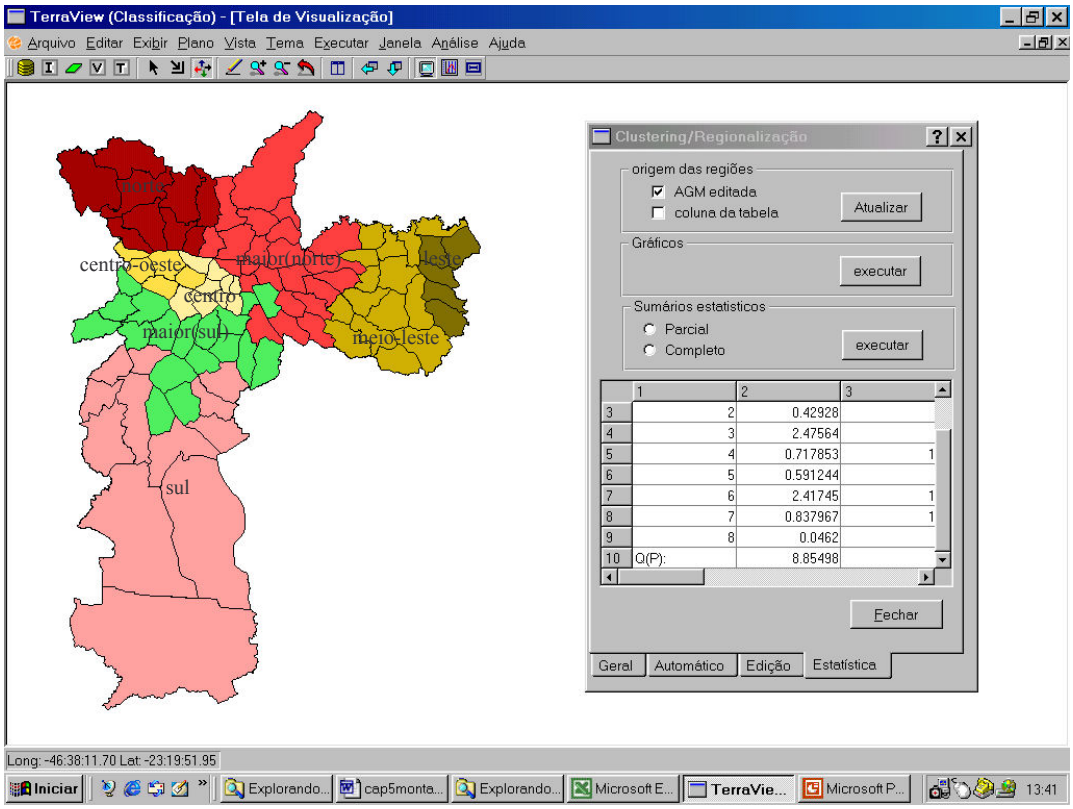


Figura 5.10: Edição de uma partição – resultado final.

Tabela 5.2: Parâmetros da partição editada.
dados: 96 distritos do município de São Paulo; Atributos: IEX_EQID e IEX_DSHU.

	membros	SQD	QQD-médio	População
leste	5	0.046	0.009	578030
meioLeste	15	0.718	0.048	1725189
maior(norte)	23	1.339	0.058	850161
maior(sul)	18	2.417	0.134	2944810
norte	9	2.475	0.275	945122
sul	12	0.838	0.070	1733507
centro	9	0.591	0.066	583231
centro-oeste	5	0.43	0.086	286135
	Q(P):	8.854	**Pop.Tot.:**	9646185

5.4 – CONCLUSÃO DO CAPÍTULO

Neste capítulo foi apresentada uma proposta para a condução do processo de regionalização de forma interativa, onde a classificação ocorre com a interferência do analista. A abordagem interativa visa dotar o procedimento de regionalização de maior flexibilidade, possibilitando que ele atenda a problemas específicos. Foi apresentado um conjunto de requisitos mínimos para que um procedimento interativo tenha atrativos e forneça as ferramentas básicas para realizar a adequação de resultados e extração de informação do processo.

A abordagem interativa, aqui proposta, utiliza a AGM como elemento fundamental, e a ela são conferidos múltiplos papéis. O desmembramento da AGM representa o processo de regionalização, sendo utilizada na construção de um histórico do processo. A AGM contêm informações importantes, não presentes em outros resultados obtidos por procedimentos de classificação, merecendo ser melhor aproveitada dentro de um processo de análise exploratória dos dados. Ela atua como um elemento "estruturante", apresentando uma hierarquia de similaridades entre os objetos e permite ao usuário selecionar e comparar grupos de objetos similares e editar partições, atendendo assim, a casos específicos e a critérios subjetivos, difíceis de serem considerados em procedimentos totalmente automáticos.

CAPÍTULO 6
EXEMPLOS DE APLICAÇÃO

O objetivo deste capítulo é aplicar o procedimento de regionalização em casos próximos a problemas reais, explorando vários aspectos das propostas apresentadas nos dois capítulos anteriores (método com busca otimizada e abordagem interativa). Os dois exemplos possuem características distintas, envolvendo objetos espaciais diferentes. O primeiro exemplo, apresentado na Seção 6.2, é realizado em etapas. Na primeira fase, são mostrados todos os passos de um processo de regionalização, deste a escolha dos objetos de estudo até o resultado do procedimento automático. A partir do resultado inicial, é realizada uma análise das regiões e o processo de desmembramento é continuado, em etapas, explorando alguns aspectos da abordagem interativa. Na Seção 6.3, é apresentado um segundo exemplo, envolvendo um número elevado de objetos espaciais. Além demonstrar o desempenho alcançado pelo método proposto no Capítulo 4, neste exemplo é explorado a utilização de uma restrição adicional.

6.1 – CONSIDERAÇÕES INICIAIS

Nos dois exemplos apresentados neste capítulo serão apresentados os principais passos existentes em um trabalho de classificação. Muitas das escolhas envolvidas em um processo de classificação são dependentes do objetivo específico do analista. Assim, logo no início de cada um dos exemplos, é definido qual é objetivo do procedimento de regionalização. Os dois casos escolhidos possuem características distintas. No primeiro exemplo, os objetos espaciais utilizados são os municípios do Estado da Bahia, os quais são agrupados em um número restrito de regiões. Este estudo pretende caracterizar a agricultura do estado de forma rápida. No segundo exemplo, é utilizado um grande

volume de dados, onde os setores censitários do município de Belo Horizonte são as unidades básicas de área e são agrupados em um número elevado de classes, formando muitas regiões homogêneas.

6.2 – Exemplo 1: Regiões agrícolas do Estado da Bahia

6.2.1 Objetivo da análise

O objetivo definido para este exemplo, do ponto de vista do analista, é desenvolver uma rápida percepção sobre alguns aspectos da agricultura da Bahia, buscando identificar e caracterizar as grandes regiões agrícolas do estado. As características utilizadas para a formação das regiões são: o tipo de exploração (pecuária, lavoura permanente, etc.); o emprego de técnicas de produção (irrigação, adubação, etc.); e a estrutura fundiária (dimensão dos estabelecimentos agropecuários, pessoas ocupadas, etc.).

6.2.2 – Escolhas dos elementos da análise

Como relatado no Capítulo 2, existe uma série de escolhas e definições a serem feitas pelo analista na condução dos procedimentos de classificação. Estas escolhas interferem diretamente no resultado do procedimento e devem, portanto, estar pautadas no objetivo da análise.

a) Escolha dos objetos: A escolha dos objetos para a análise é naturalmente imposta pelo objetivo do trabalho e pela disponibilidade de informação básica. Uma fonte de informação importante e relacionada com a agricultura é o Censo Agropecuário, realizado periodicamente pelo IBGE. Utilizaremos, neste exemplo, os dados do último levantamento, executado entre 1995 e 1996. Para as unidades da Federação, os dados do Censo Agropecuário são apresentados em três níveis de detalhamento, correspondendo às seguintes unidades de área: municípios, microrregiões e mesorregiões. Utilizaremos como unidade básica *o município*, pois os dados possuem um nível menor de agregação que as demais representações. Na época do levantamento, o Estado da Bahia possuía 415 municípios.

b) Escolha dos atributos: No censo agropecuário são listados 133 variáveis. A maioria destas variáveis são expressas em unidades de área, número de estabelecimentos ou valor da produção (ex.: área plantada com a cultura de cana-de-açúcar, área do municípios com pastagem, número total de estabelecimentos agropecuários, valor da produção vegetal, etc.).

Algumas das variáveis do censo precisaram receber um tratamento para serem utilizadas na análise. O número de estabelecimentos agropecuário com irrigação, por exemplo, sofre influência direta da dimensão geográfica do município e do número total de estabelecimentos existentes na localidade. Isto pode provocar uma distorção na análise. Para corrigir este problema, foi criado um conjunto de índices (variáveis derivadas), a partir da combinação de variáveis primárias do censo:

- *Id_Past*: porcentagem da área do município com pastagem.
- *Id_Lav*: porcentagem da área do município com lavouras.
- *Id_Matas*: porcentagem da área do município com matas.
- *Id_Irrig*: porcentagem de estabelecimentos agropecuários que utilizam irrigação.
- *Id_Eletr*: porcentagem de estabelecimentos com energia elétrica.
- *Id_AssTc*: porcentagem de estabelecimentos com assistência técnica.
- *Id_Mn10*: porcentagem de estabelecimentos com menos de 10 ha.
- *EstMedio*: tamanho médio dos estabelecimentos.
- *Id_LPerm*: porcentagem de estabelecimentos com lavouras permanentes.
- *Id_LTemp*: porcentagem de estabelecimento com lavouras temporárias.

As variáveis de censos tendem a ser fortemente correlacionadas. Deve-se evitar o uso simultâneo de variáveis com alto índice de correlação, pois esta redundância de informação, além de encarecer desnecessariamente a análise, pode provocar uma distorção no peso das informações consideradas no processo de classificação. A colinearidade existente entre as variáveis primárias do censo agropecuário refletiu nos índices criados. A Figura 6.1 mostra, como exemplo, os gráficos de espalhamento, envolvendo três atributos (*Id_Past*, *Id_Matas* e *Id_Lav*), mostradas duas a duas. Há uma correlação negativa clara entre estas variáveis, o que é esperado, pois as áreas dos municípios ocupadas por pastagem, lavoura e matas concorrem pelo mesmo espaço

físico nos municípios. A Tabela 6.1 mostra os valores dos índices de correlação para este conjunto de variáveis. Entre estas três variáveis foi escolhida para fazer parte da classificação apenas o índice de pastagem, já que ele possui, simultaneamente, uma forte correlação com os índices de lavoura e matas.

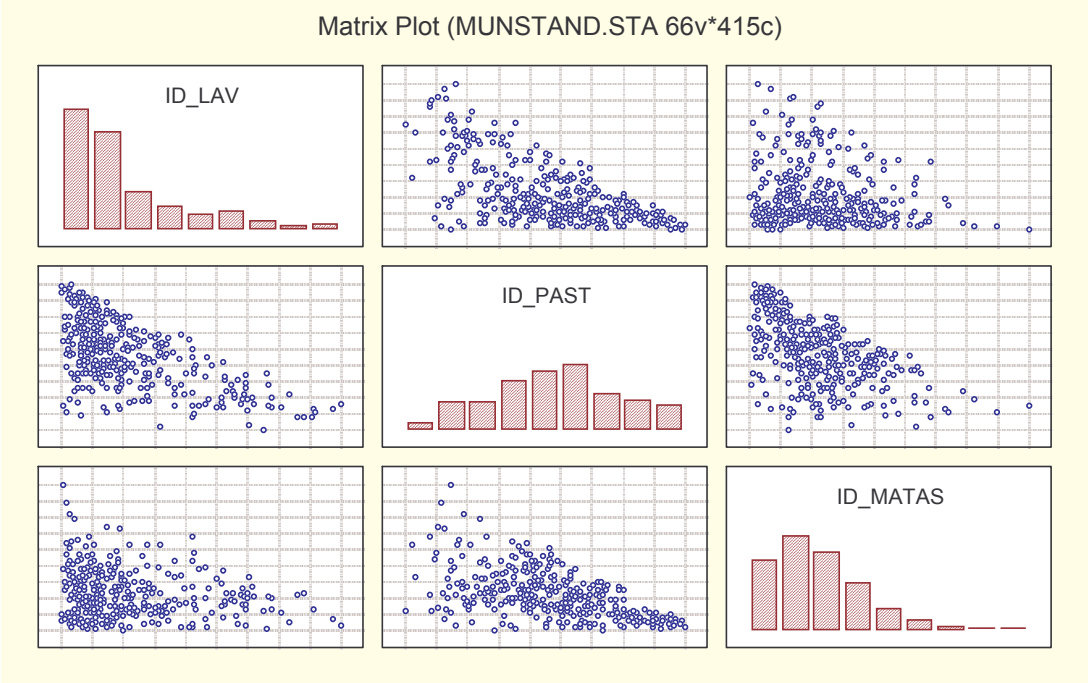


Figura 6.1: Gráficos de espalhamento entre as variáveis *Id_Past*, *Id_Matas* e *Id_Lav*.

Tabela 6.1: Correlação entre variáveis
dados: 415 municípios do Estado da Bahia

	ID_LAV	ID_PAST	ID_MATAS
ID_LAV	1.00	-0.61	-0.16
ID_PAST	-0.61	1.00	-0.57
ID_MATAS	-0.16	-0.57	1.00

Para a escolha do conjunto final de variáveis foram considerados, além da correlação, os seguintes aspectos: a importância do atributo para a análise, a distribuição e a dependência espacial dos atributos. A dependência espacial tem uma importância significativa na classificação com restrição de contigüidade. Se utilizarmos somente

variáveis sem dependência espacial, o resultado da regionalização será pobre, pois será determinado apenas pela proximidade física dos membros, ou seja, qualquer partição tenderá a ter a mesma qualidade, do ponto de vista de homogeneidade. Por outro lado, variáveis com dependência espacial tendem a formar grupos de objetos contíguos, resultando em regiões distintas e internamente mais homogêneas. A dependência espacial das variáveis foi avaliada utilizando *o Índice Global de Moran* (Bailey & Gatrell, 1995).

A escolha da variável *Id_Lperm*, por exemplo, foi definida com o objetivo de diferenciar municípios com predomínio de lavouras permanentes dos municípios com culturas temporárias. Assim, as regiões poderiam mostrar, por exemplo, áreas com predomínio da lavoura de cacau e cana-de-açúcar, em contraponto a regiões com prevalências das culturas de soja, milho ou feijão. O conjunto final de atributos considerados na análise foi: *Id_Past*, *Id_Irrig*, *EstMedio* e *Id_Lperm*. A Tabela 6.2 mostra os índices de correlação entre as variáveis selecionadas.

Tabela 6.2: Correlação entre as variáveis escolhidas.
dados: 415 municípios do Estado da Bahia

	Id_Pasto	**Id_Irrig**	**Id_Mn10**	**Id_LPerm**
Id_Pasto	1.00	-0.09	-0.17	-0.31
Id_Irrig	-0.09	1.00	0.04	-0.13
Id_Mn10	-0.17	0.04	1.00	-0.01
Id_LPerm	-0.31	-0.13	-0.01	1.00

***c) Homogeneização das variáveis**:* das quatro variáveis escolhidas para a análise, três são percentagens (com variação entre zero e 1) e uma variável representa um valor médio, expresso em unidade de área. Para que houvesse equilíbrio entre as quatro variáveis no processo de classificação, optou-se por padronizá-las, utilizando-se:

$$y' = \frac{y - \bar{y}}{sd}$$

onde:

- y é o valor de índice;

- *y'* o valor padronizado para o índice;
- $\bar{y}$ é o valor médio e;
- *sd* é o desvio padrão do índice *y*.

d) ***Outras escolhas:*** As demais escolhas acompanharam as definições utilizadas durante os experimentos apresentados nos capítulos anteriores. A *medida de similaridade* utilizada nos exemplos deste capítulo foi a distância euclidiana. O *método* utilizado foi, evidentemente, o procedimento de regionalização baseado no uso da AGM com otimização na fase de poda (CP = 10). O *critério de agrupamento* utilizado foi a homogeneidade interna das classes, conforme a Expressão 2.1 Quanto ao *número de regiões*, em função do objetivo proposto para o exemplo (determinar as grandes regiões agrícolas do estado) escolhemos trabalhar inicialmente com um número pequeno de classes e, se necessário, aumentar o número de regiões posteriormente. Esta estratégia pode ser utilizada, mesmo em estudos envolvendo um maior número de classes, dentro de uma abordagem interativa, que vai do menor para maior nível de detalhamento das regiões. O número inicial utilizado foi de seis regiões, e depois foram gerados outros resultados com um número maior de classes.

6.2.3 – EXECUÇÃO DO PROCEDIMENTO

A Figura 6.2 mostra o resultado da regionalização onde os 415 municípios aparecem agrupados em 6 regiões. A Tabela 6.3 mostra os valores para as *SQD*s, o número de elementos membros e o valores médios dos atributos das regiões. A Figura 6.3 mostra um diagrama onde os valores médios dos atributos são apresentados, por região. Os valores referentes a uma mesma região aparecem separados por cores e ligados por um segmento de reta. Nesta figura pode-se perceber que algumas variáveis são determinantes para a formação das regiões e que não existe, neste caso, duas regiões que possuam valores médios semelhantes, para os quatro atributos simultaneamente. Mas, pelos valores das *SQD*s e número de objetos, podemos perceber diferenças entre as regiões quanto a homogeneidade dos objetos membros. A Região 3, por exemplo,

apresenta mais da terça parte dos municípios do estado e apresenta uma alta dispersão (segundo maior valor de *SQD-médio*), enquanto que a Região 6 foi mais homogênea, agrupando apenas 8 municípios.

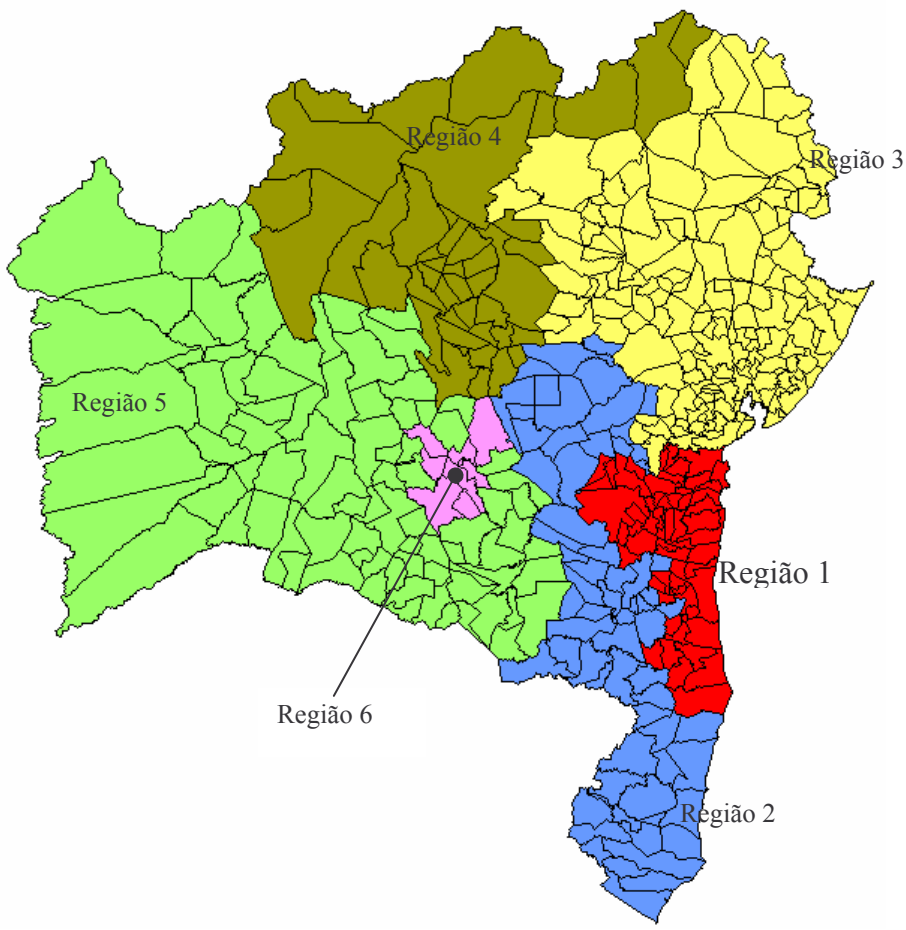


Figura 6.2: Classificação em seis regiões.

Tabela 6.3: Valores médios dos atributos por região.
dados: 415 municípios do Estado da Bahia.

	Id_Pasto	Id_Irrig	Id_Mn10	Id_Lperm	SQD	Num.Elems.	SQD-médio
Região 1	-0.85	-0.24	-0.35	1.94	117.23	57	2.06
Região 2	0.97	-0.21	-1.13	-0.17	102.50	65	1.58
Região 3	0.35	-0.19	0.68	-0.16	397.14	158	2.51
Região 4	-1.11	0.51	0.47	-0.47	142.30	41	3.47
Região 5	-0.26	0.13	-0.43	-0.61	146.50	86	1.70
Região 6	-0.26	3.17	0.56	-0.43	12.71	8	1.59

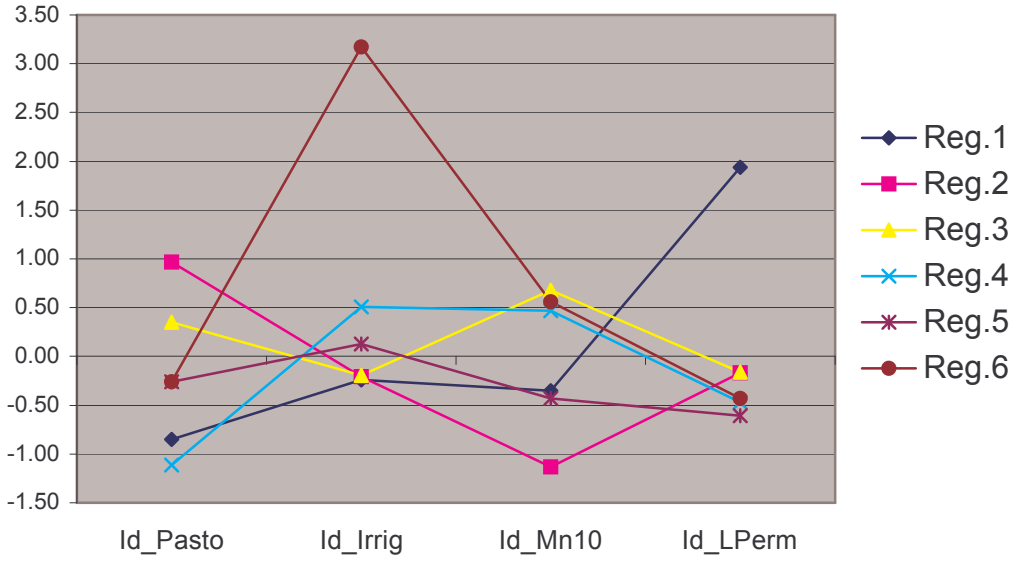


Figura 6.3: Médias dos atributos por região.

À medida que formos apresentando a análise dos resultados da Figura 6.3 e Tabela 6.3, vamos atribuir rótulos às regiões como forma de caracterizá-las. A Região 1 apresenta como maior destaque um alto valor médio para o índice de lavoura permanente, chegando a quase dois desvios padrão em relação à média geral. Em relação aos demais atributos, esta região apresenta pouca área com pastagem, pouca irrigação e pouca presença de pequenos estabelecimentos. O rótulo escolhido para a Região 1 foi: *região com estabelecimentos extensos e com lavouras permanentes*. Tratando-se da Bahia, o mais importante produtor de cacau do Brasil, e pela localização da região, é fácil imaginar que a Região 1 corresponde à região cacaueira do estado. A Bahia é responsável por cerca de 91% do total da área colhida no país. A Região 1 concentra, por sua vez, cerca de 90% da área colhida no estado.

A Região 2 apresenta a maior porcentagem de área com pastagem, o menor índice de pequenas propriedades, baixo índice de irrigação e o segundo maior índice de lavoura permanente. Com estas características, denominaremos esta região como sendo: *região com predominância de criação extensiva de gado*. Diferente da região cacaueira,

as pastagens aparecem mais distribuídas pelo estado, ainda assim, este parece ser um rótulo apropriado para esta região.

A Região 3, posicionada no nordeste do estado, compreende deste a região metropolitana de Salvador até a divisa norte do estado, agrupando 158 municípios. Ela apresenta o maior índice de pequenas propriedades entre as seis regiões, o segundo maior índice de pastagem e valores pequenos para os índices de irrigação e lavoura permanente. O valor da *SQD* da região demonstra que ela é muito diversificada, possuindo o segundo maior valor de *SQD-médio* das seis regiões geradas neste experimento. O maior índice de pequenas propriedades pode estar relacionando à maior densidade populacional da região e por ela compreender a região metropolitana de Salvador. Por ser pouco homogênea esta região deve sofrer algumas subdivisões. Isto certamente permitirá que mais informações sejam extraídas e que a região seja melhor caracterizada. Um rótulo, preliminarmente estabelecido, para esta grande região foi: *região diversificada com pequenas propriedades.*

Ao norte do estado está localizada a Região 4. Sua fronteira norte coincide com o curso do rio São Francisco, o qual fornece as águas que suprem vários projetos de agricultura irrigada. Esta região, formada por 41 municípios, é a região com menor média de áreas de pastagem. Pela correlação negativa observada na Tabela 6.1, podemos supor que esta região possua grande concentração de lavouras. Além disso, apresenta o segundo maior valor para o índice de irrigação, alto índice de pequenas propriedades e pouca lavoura permanente. Esta região ainda apresenta um alto valor para o parâmetro *SQD-médio*, indicando que ela não é homogênea. Como rótulo para a Região 4 foi estabelecido: *região com estabelecimentos pequenos, com destaque de lavouras temporárias e irrigadas*.

A Região 5, situada no oeste do estado, possui pouca área com pastagem (abaixo da média), estabelecimentos com irrigação acima da média, poucas propriedades de pequeno porte pequenas e prevalência de lavoura temporárias. É a segunda maior região em número de membros (84), mas é bem mais homogênea que a Região 2. O rótulo escolhido foi: *Região com grandes propriedades, com lavouras temporárias e com presença de áreas irrigadas.*

A última região analisada apresentou um comportamento singular. A Região 6 é central, formada por apenas 8 municípios. Ela apresentou um valor para a *SQD* de apenas 13,05, que considerando o número de municípios, resulta em um valor médio de 1,63, o menor entre todas as seis regiões. O principal destaque desta região é o alto valor para o índice de estabelecimentos com irrigação, que alcança a três desvios padrão em relação à média geral. As demais características da região são: poucas áreas com pasto, alto índice de pequenas propriedades e poucos estabelecimentos com lavoura permanente. O que mais diferencia esta região da Região 4 é a intensidade de estabelecimentos que utilizam irrigação, demonstrando-se tratar de um pólo de irrigação no estado. Denominamos esta classe como *Região com forte predominância de estabelecimentos irrigados, pequenos e com lavoura temporárias.*

A utilização de um número pequeno de regiões facilita a análise, mas pode esconder algumas informações importantes e produzir regiões com muitos membros e baixa homogeneidade, como a Região 3. Ao realizarmos um novo procedimento, agora classificando os municípios em 7 regiões, a Região 4 foi desmembrada, determinando um pequeno grupo de 4 municípios. A Figura 6.4 mostra o resultado para a classificação em 7 regiões e o gráfico da Figura 6.5 compara os resultados das médias dos atributos das duas novas regiões (4a e 4b) com a Região 6. Analisando os dados para a região 4.a, verifica-se que ela concentra fortemente os estabelecimentos com irrigação, que estavam presentes na antiga região norte (Região 4), e o novo grupo de quatro municípios possui valores médios para os atributos mais próximos aos da Região 6 (pólo de irrigação). Como as Regiões 4 e 6 estão fisicamente distantes, sem conectividade, elas foram separadas pelo método de regionalização em regiões distintas. O procedimento de classificação em sete regiões destacou o pólo de irrigação existente no entorno de Petrolina (PE) e Juazeiro (Ba), que passa a ser, agora, a região com o maior valor médio para o índice de irrigação.

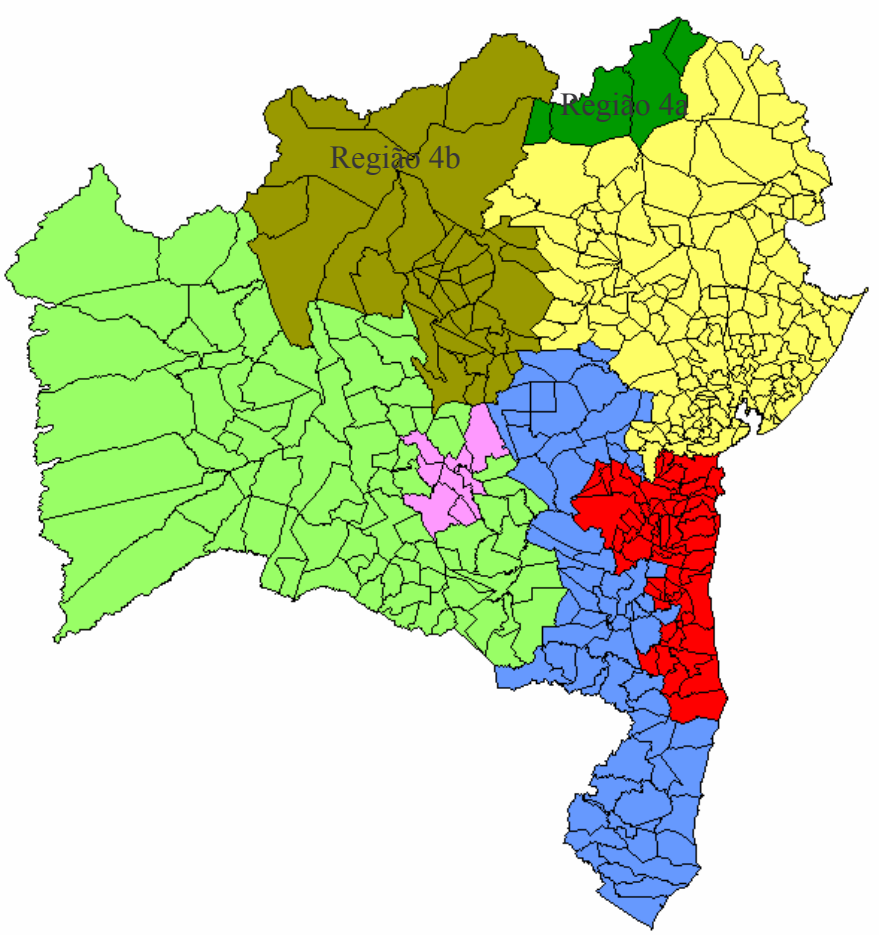


Figura 6.4: Classificação em sete regiões.

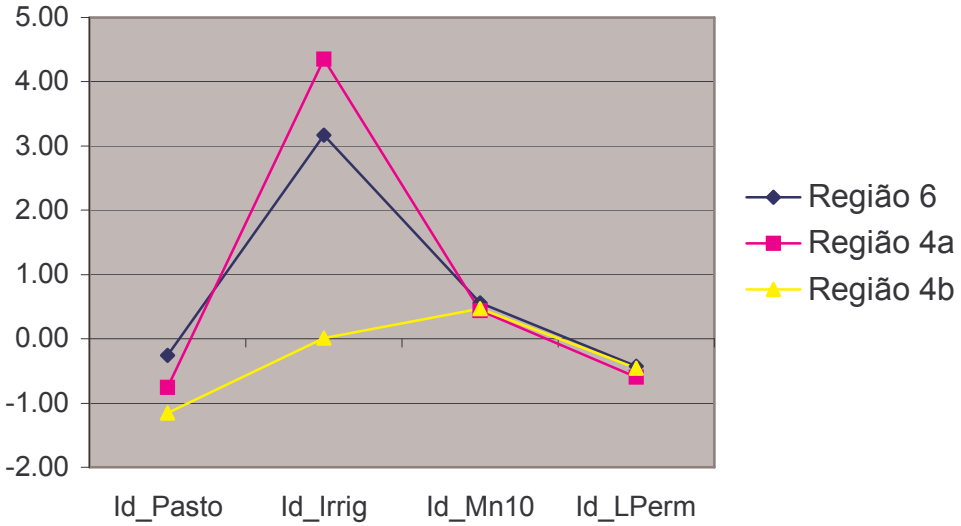


Figura 6.5: Médias dos atributos para a nova região (pólo de irrigação de Juazeiro).

Por fim, como a Região 3 continha um grande número de objetos (158 municípios) e era pouco homogênea, escolhemos realizar um desmembramento deste grupo, em quatro novas regiões, totalizando dez regiões para o estado (Figura 6.6). Uma das novas regiões, com apenas um membro (Região 3d), possui muitos estabelecimentos com irrigação, e também está situada a margem do rio São Francisco. Por não estar contíguo, por pouco, ao pólo de irrigação de Juazeiro, esta região aparece como um objeto solitário. Este município é um caso especial, que combina um alto índice de irrigação e com alto índice de pastagem. As demais divisões criadas subdividem melhor os municípios, modificando ligeiramente as médias dos atributos. Como destaque, aparece uma região (3c) com forte presença de pastagens, mas com presença maior de pequenas propriedades em relação à Região 2. A Figura 6.7 mostra os valores médios dos atributos para as novas regiões criadas com o desmembramento da região nordeste.

Com este primeiro exemplo vimos que a regionalização pode ser utilizada para uma rápida identificação das regiões e a compreensão de suas características. Também verificamos que pode haver a necessidade de alterar o número de regiões em função da análise dos resultados da partição e que medidas e gráficos podem auxiliar ao analista a conduzir o processo de classificação. O aumento gradual de regiões pode ser utilizado para um estudo em diversos níveis de detalhamento. A abordagem dirigida pelo usuário permite investigar várias partições diferentes sem a necessidade de executar o procedimento deste o seu início. Neste caso, o número de regiões foi ampliado a partir de um resultado inicial (seis regiões) e, posteriormente, duas regiões foram desmembradas. Novas modificações poderiam ser realizadas, em função do conhecimento do analista e de objetivos específicos.

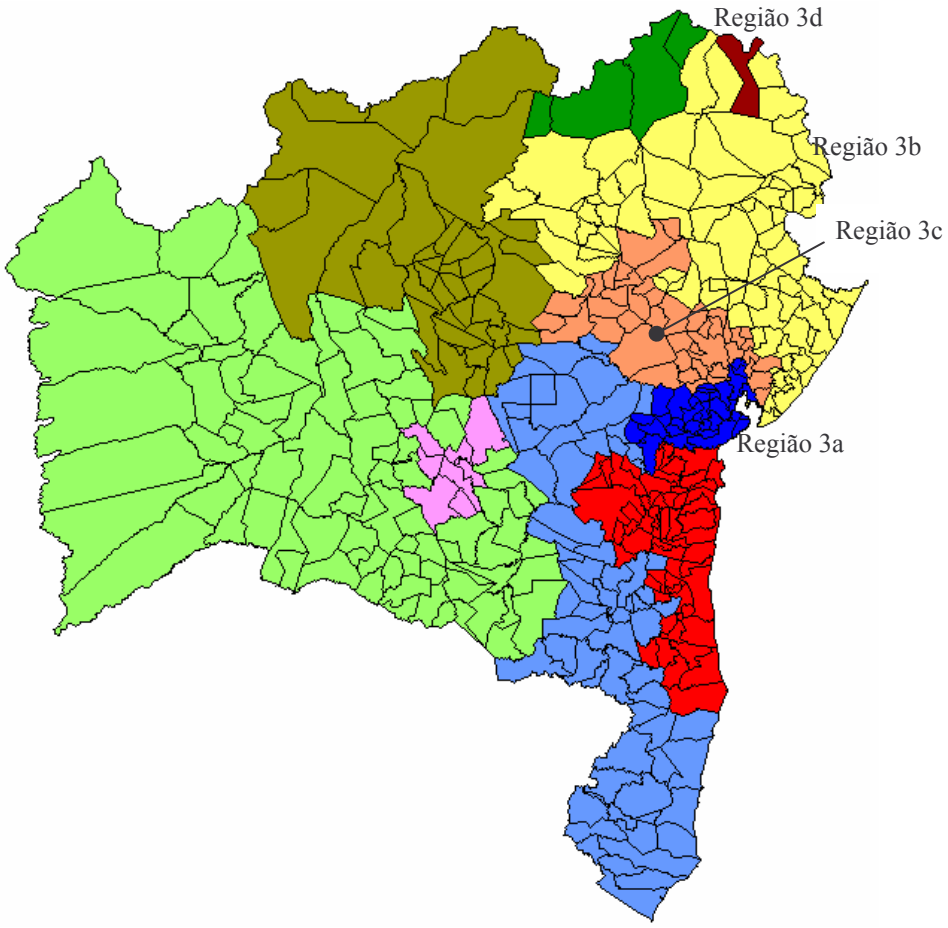


Figura 6.6: Classificação em dez regiões.

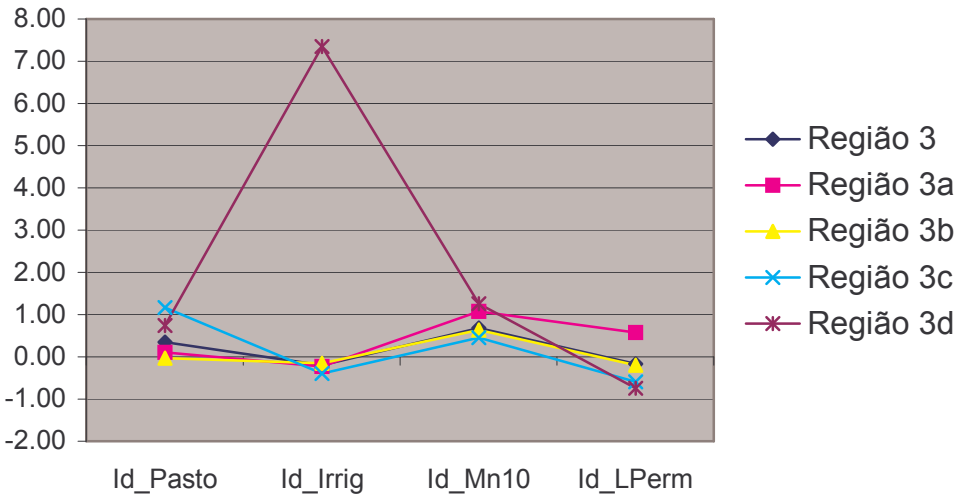


Figura 6.7: Médias dos atributos para as novas regiões.

6.3 - Exemplo 2: Regionalização dos setores censitários de BH

Em algumas aplicações há necessidade de adicionar outras restrições ao procedimento de regionalização além da contigüidade. Em estudos envolvendo a análise de eventos com baixa freqüência de ocorrência, por exemplo, pode ser necessário estabelecer valores mínimos para a população de uma região, de forma que as taxas derivadas sejam representativas. Em outros casos, pode-se desejar que as regiões resultantes não tenham dimensões geográficas com muita variação, ou ainda, que o número de objetos membros das regiões sejam semelhantes.

Uma forma simples de adicionar restrições ao método baseado na AGM é utilizar limites para determinados atributos de uma região. Se uma solução ultrapassar algum dos limites estabelecidos ela será considerada uma solução inválida. Assim, a estratégia de busca passa a procurar, não a melhor solução exclusivamente do ponto de vista da homogeneidade das regiões, mas sim, a melhor solução válida. Este mecanismo foi utilizado neste segundo exemplo.

Supomos que a prefeitura de Belo Horizonte pretenda identificar áreas prioritárias do município para estabelecer políticas publicas como saneamento básico ou programas de distribuição de renda. Para isto, ela precisa classificar grandes grupos homogêneos de domicílios, priorizar e atuar especificadamente em cada região. A regionalização pode ser uma ferramenta útil neste caso. Porém, regiões com poucos domicílios podem não justificar, isoladamente, o desenvolvimento de uma política específica por parte da prefeitura. Neste caso, a adoção de uma restrição adicional, pode evitar que regiões pouco representativas sejam identificadas pelo método automático.

Neste exemplo, experimentamos o procedimento em um grande volume de dados e aproveitamos para explorar o uso de uma restrição adicional, estabelecendo um número mínimo de domicílios por região.

6.3.1 – OBJETIVO DA ANÁLISE

Agrupar os setores censitários, definidos pelo IBGE para o município, em 90 regiões homogêneas considerando aspectos sócio-econômicos e condições dos domicílios, de forma que as regiões resultantes contenham um número mínimo de 1000 domicílios.

6.3.2 – ESCOLHAS DOS ELEMENTOS DA ANÁLISE

a) Escolha dos objetos: Os objetos utilizados neste exemplo são os 1999 setores censitários do município de Belo Horizonte, capital o Estado de Minas Gerais, relativos ao Censo Demográfico realizado pelo IBGE, em 1991.

b) Escolha das variáveis: As variáveis do Censo Demográfico referem-se ao nível de educação da população (anos de estudo do chefe da família, número de analfabetos, etc.), condições econômicas (renda familiar, faixas de rendimento do chefe da família, renda média, etc.) e condições de moradia (domicílio com canalização interna de água, com esgoto ligado a rede geral, existência de sanitário, etc.). Da mesma forma que as variáveis do censo Agropecuário, utilizadas no primeiro exemplo deste capítulo, as variáveis do Censo Demográfico também são fortemente correlacionadas. Foi escolhido um conjunto preliminar de variáveis e, com elas, foram criadas algumas variáveis derivadas, formando índices, e assim, corrigindo a variação do número da população e domicílios por setor censitário. As variáveis utilizadas são:

- ***IdSal5a10***: porcentagem de salários entre 5 a 10 salários mínimos:
- ***RendMedChef***: renda média dos chefes de famílias:
- ***taxAlfa***: taxa de alfabetização:
- ***IdEst8a10***: porcentagem de população com 8 a 10 anos de estudo.
- ***IdEstMais15***: porcentagem de população com mais de 15 anos de estudo
- ***IdDomAg***: porcentagem de domicílios com canalização interna de água.
- ***IdSanitRG***: porcentagem de domicílios com vasos sanitários ligados à rede geral
- ***IdColLixo***: porcentagem de domicílios atendidos pela coleta de lixo.

A Tabela 6.4 mostra as correlações entre todos os atributos listados acima. Destaca-se o índice de correlação existente entre as variáveis ***IdEstMais15 e RendMedChef*** que foi de 90%, havendo, praticamente, uma duplicidade de informação. A escolha final dos atributos recaiu sobre um subconjunto de três variáveis: *RendMedChef, IdEst8a10 e IdDomA*.

Tabela 6.4: Indice de correlação entre as variáveis.
dados: 1999 setores censitários de Belo Horizonte (MG)

	RenMedChef	IdSal5a10	TaxaAlfa	IdEst8a10	IdEstMais15	IdDomAg	IdColLixo	IdSanRG
RenMedChef	1.00	0.46	0.59	-0.11	0.90	0.25	0.34	0.41
IdSal5a10	0.46	1.00	0.69	0.47	0.64	0.36	0.48	0.56
TaxaAlfa	0.59	0.69	1.00	0.45	0.61	0.55	0.61	0.67
IdEst8a10	-0.11	0.47	0.45	1.00	-0.07	0.36	0.40	0.40
IdEstMais15	0.90	0.64	0.61	-0.07	1.00	0.24	0.34	0.43
IdDomAg	0.25	0.36	0.55	0.36	0.24	1.00	0.56	0.64
IdColLixo	0.34	0.48	0.61	0.40	0.34	0.56	1.00	0.76
IdSanRG	0.41	0.56	0.67	0.40	0.43	0.64	0.76	1.00

c) Homogeneização das variáveis*:* As variáveis escolhidas foram padronizadas da mesma forma exemplo anterior, utilizando a Expressão 6.1.

d) ***Outras escolhas:*** A *medida de similaridade* e o *critério de agrupamento* também foram os mesmos utilizados anteriormente. Quanto ao *número de regiões*, o objetivo proposto para o exemplo e identificar grandes grupos homogêneos de domicílios. Em função deste objetivo, estabelecemos um número de 90 regiões para a análise. Foi utilizado o método proposto no capítulo 4 (CP = 5).

6.3.3 – EXECUÇÃO DO PROCEDIMENTO

Inicialmente, foi executado um procedimento de regionalização sem a consideração de um limite mínimo de domicílios. O resultado é mostrado na Figura 6.8. Existe uma variabilidade acentuada em relação à quantidade de setores censitários por região. Há regiões compostas por dezenas e até centenas de setores censitários e, por outro lado, algumas regiões formadas por poucos setores. Foram criadas quatorze regiões unitárias. Muitas destas regiões possuíam pouca população e domicílios. A Figura 6.9 mostra uma parte do município, em uma escala maior, onde aparecem algumas regiões e as suas

respectivas árvores sobrepostas. Duas regiões (destacadas na Figura 6.10) foram selecionadas para mostrar a discrepância existente entre a representatividade das regiões. Os dados apresentados na Tabela 6.5 são referentes aos valores médios dos atributos (padronizados), número de objetos, população total e número de domicílios das duas regiões destacadas.

Buscando a eliminação de regiões pouco representativas, um segundo procedimento de regionalização foi executado, desta vez, utilizando uma restrição adicional: o número mínimo de domicílios em uma região tem que ser superior a mil unidades. A Figura 6.11 mostra a mesma área da Figura 6.9 com a nova classificação. As regiões com poucos membros desapareceram, sendo incorporadas às regiões maiores ou dando origem a novas regiões. Com esta nova classificação não ocorreu nenhuma região unitária, sendo a menor região identificada, formada por três setores com 1125 domicílios.

Tabela 6.5: Discrepâncias entre regiões.

Dados: Setores censitários de Belo Horizonte.

	Região 1	Região 2
membros	192	1
SSD	277	0
idRenStd	0.257	-0.164
Id8a10Std	0.314	0.031
IdDomAgStd	0.251	0.5
População	169729	849
Domicílios	47451	210

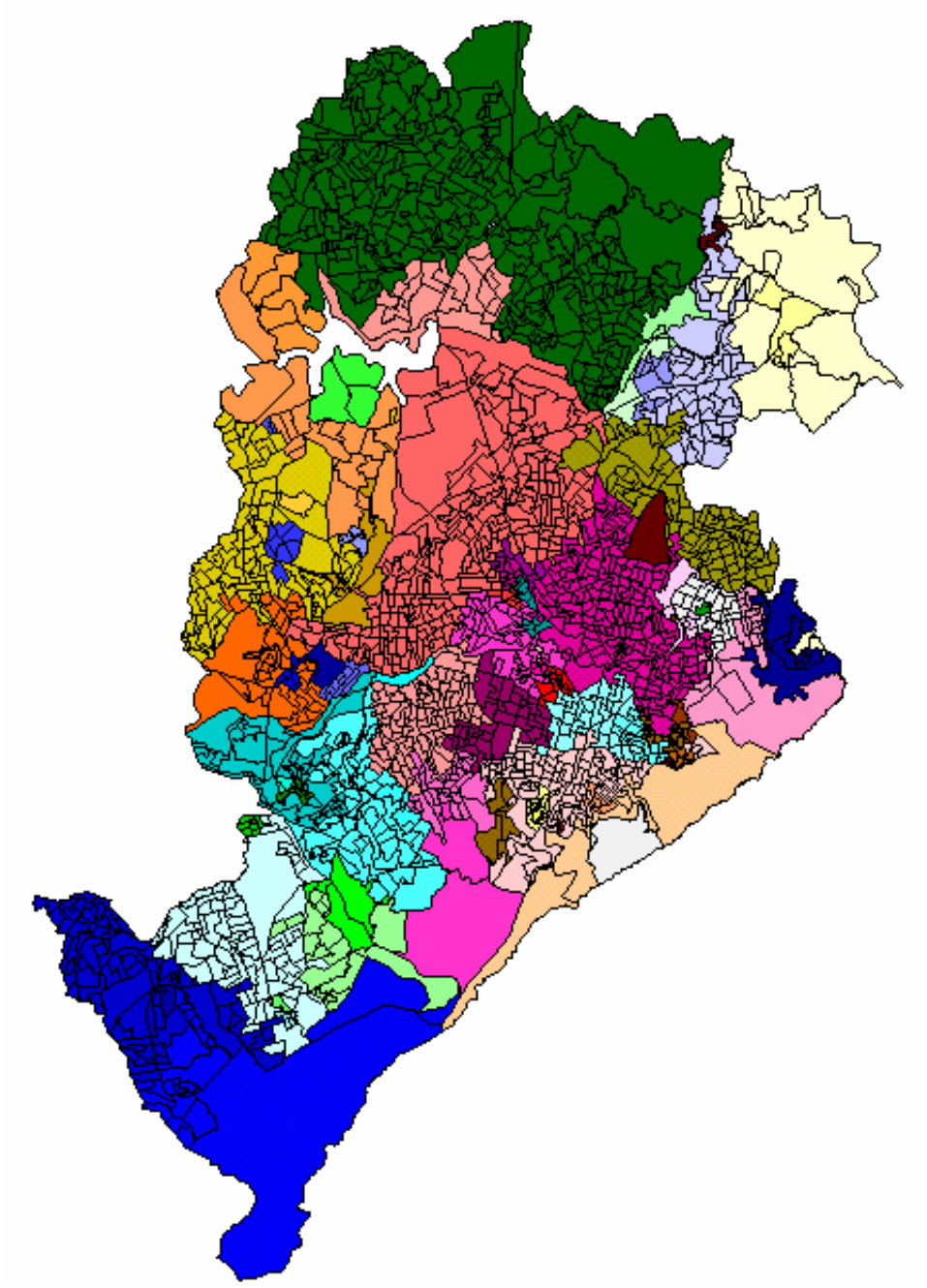


Figura 6.8: Setores censitários classificados em 90 regiões.

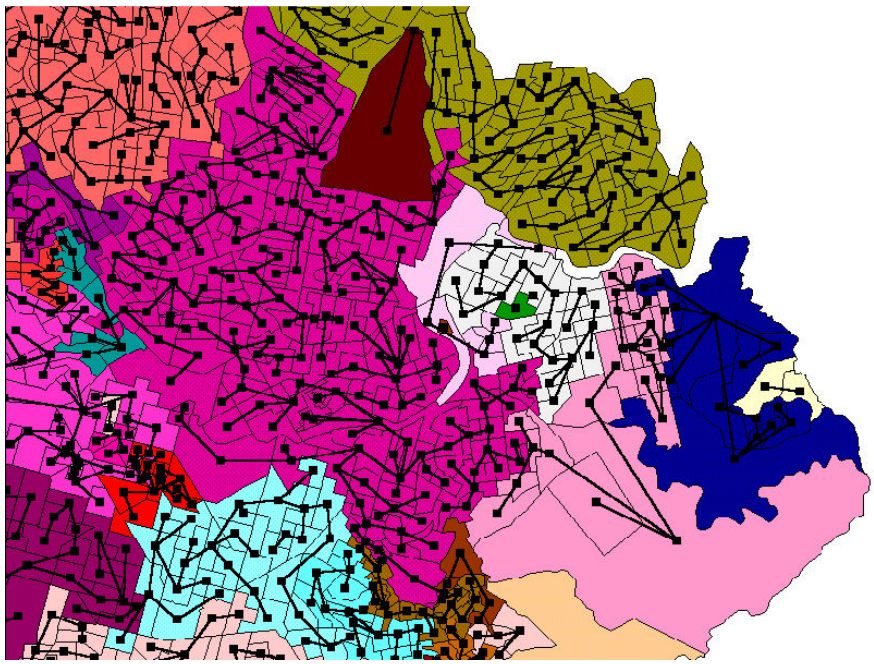


Figura 6.9: Detalhe de regiões e árvores sobrepostas.

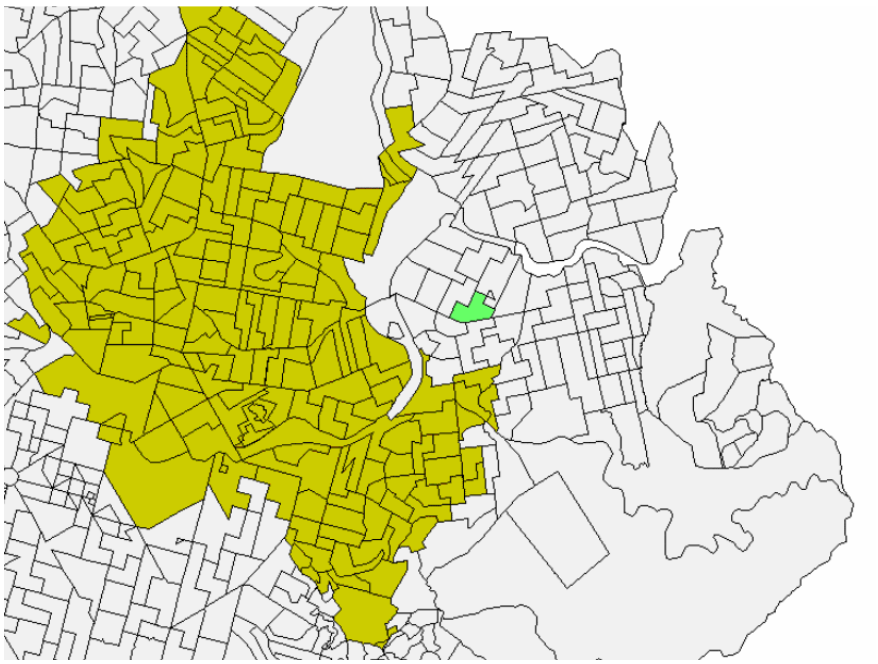


Figura 6.10: Regiões selecionadas.

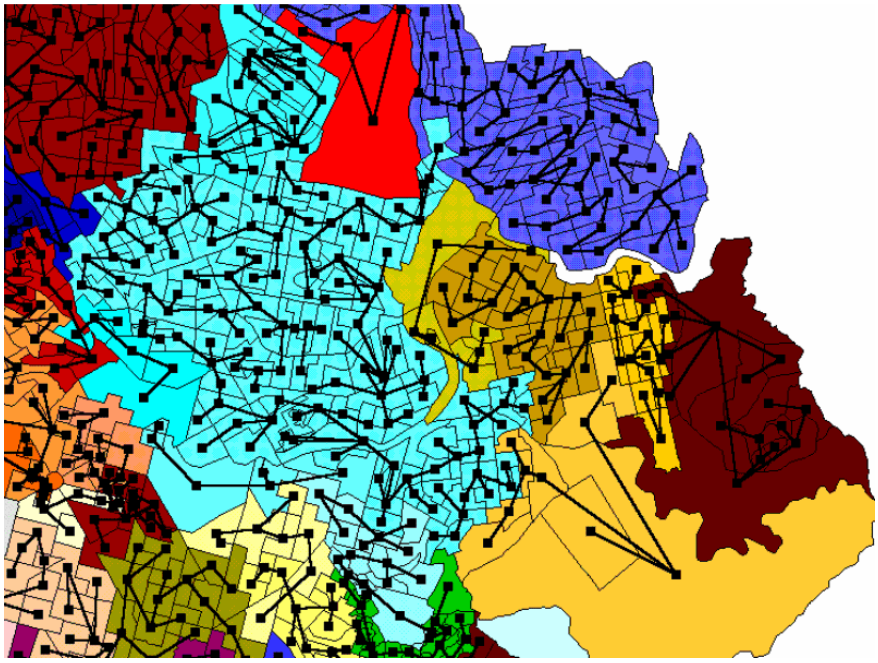


Figura 6.11: Regionalização com restrição adicional.

Neste segundo exemplo foi utilizado um CP muito pequeno, principalmente considerando o número de objetos envolvidos. Isto privilegiou a rapidez, facilitando o experimento com a restrição adicional, mas ocasionando um prejuízo na qualidade da partição. Aproveitamos este exemplo para avaliarmos o desempenho do método em um problema com muitos objetos. Experimentamos outros valores de CP, e os resultados obtidos para o tempo de execução e para qualidade da partição são apresentados na Tabela 6.6, em comparação com o método exaustivo. O tempo de execução para o método com busca exaustiva (quase dois dias e meio) mostra claramente que à medida que o número de objetos cresce o tempo de processamento tende a se elevar fortemente, inviabilizando sua aplicação. O valor da degradação para p CP =5, mostrou que outras escolhas teriam uma melhor relação de compromisso entre o ganho no tempo de execução e a degradação da qualidade da partição.

Tabela 6.6: Comparação entre o ganho no tempo e a degradação da qualidade.
dados: 1999 setores sensitários de BH.

	CP = 5	CP = 10	CP = 20	CP=30	Exaustivo
tempo(s)	3389	4892	9193	13585	209649
tempo(h)	0,9	1,4	2,6	3,8	58,2
tempo(%)	1,6	2,3	4,4	6,5	
Q(P)	2125,7	1992,5	1715,4	1675,0	1612,6
degradação	31,8%	23,5%	6,4%	3,9%	

6.4 – Conclusão do capítulo

Neste capítulo foram apresentados dois exemplos de aplicação utilizando o procedimento de regionalização proposto. No primeiro exemplo, a regionalização foi empregada como forma de caracterização de grandes regiões agrícolas no Estado da Bahia, sendo exploradas algumas características da abordagem interativa, onde o processo de regionalização foi realizado em etapas, inicialmente estabelecendo um pequeno número de regiões, que após uma análise, foram sendo subdivididas. No segundo exemplo, o método de regionalização foi aplicado a um caso envolvendo um grande número de objetos, sendo explorada também, a possibilidade de utilização de restrições adicionais.

CAPÍTULO 7
CONCLUSÕES DO TRABALHO

Este trabalho apresentou duas propostas relacionadas à condução do procedimento de regionalização. A primeira refere-se a um método alternativo para a regionalização, baseado no uso da *árvore geradora mínima* associado a técnicas de otimização, como forma de viabilizar a sua aplicação a problemas com grandes volumes de dados espaciais. A segunda proposta apresenta a possibilidade de tratar o problema de regionalização como um processo interativo, onde o analista interfere na condução da classificação, aumentando a flexibilidade do procedimento. Neste último capítulo, são apresentadas as principais conclusões que o trabalho forneceu, agrupadas em três seções: conclusões relacionadas ao estudo comparativo entre abordagens de regionalização (Seção 7.1); referentes ao método de regionalização com emprego de técnicas de otimização (Seção 7.2); e relacionadas à proposta de regionalização interativa (Seção 7.3).

7.1 - CONCLUSÕES REFERENTES ÀS ABORDAGENS DE REGIONALIZAÇÃO

Dentro das abordagens utilizadas para a regionalização, o procedimento realizado em duas etapas (classificação e geração de regiões) não fornece qualquer controle sobre o processo e proporciona resultados pobres em termos de qualidade da partição, já que para restringir sua tendência em gerar um número elevado de regiões ao final do processo, é preciso limitar o número de classes na primeira etapa (classificação).

O método de regionalização por componentes ponderadas, utilizando a técnica de classificação *k-médias*, demonstrou ser muito rápido, o que é bom para problemas com um grande número de objetos. Porém, o procedimento apresentou baixa qualidade para a partição resultante em relação aos métodos AZP e via-AGM. Este método apresenta dois problemas adicionais: determinar o peso ideal para as componentes e a variabilidade da dimensão das áreas dos objetos. Para forçar que todos os objetos fiquem contíguos em toda a área de estudo é necessário aumentar o peso da componente geográfica, o que faz a qualidade da partição cair (em termos de homogeneidade interna das classes).

Os métodos baseados no uso direto da estrutura de vizinhança foram os que apresentaram melhores índices de qualidade para as partições resultantes, porém, foram os mais caros computacionalmente. O método AZP foi o mais dispendioso e se mostrou sensível a ótimos locais. Em problemas envolvendo um número elevado de objetos e que necessitem na análise a realização de vários experimentos sucessivamente, o custo computacional pode inviabilizar o uso do AZP.

O uso da AGM como forma de simplificar o problema de particionamento de grafos se mostrou eficiente, já que o método via-AGM obteve, nos testes realizados, desempenho superior ao método AZP, fornecendo um resultado ligeiramente superior, em termos de qualidade da partição, em um tempo de execução significativamente menor.

7.2 – CONCLUSÕES SOBRE O MÉTODO ALTERNATIVO

O uso de técnicas de otimização permitiu uma diminuição expressiva no número de avaliações necessárias para a execução do procedimento de regionalização. Isto combinado com a simplificação do grafo de conectividade, conseguida por meio da árvore geradora mínima, torna o método eficiente, viabilizando seu emprego em problemas envolvendo um número elevado de objetos.

O ganho na utilização da estratégia de busca para a fase de poda da AGM é variável. Ele é significativamente maior quando a busca é realizada em uma árvore com muitos

vértices. Por isto, o ganho do método aumenta de forma acentuada com o número de objetos na análise. Isto explica também, o maior ganho no início do processo de desmembramento, quando as árvores exploradas são maiores. O aumento do número de atributos na análise, também aumenta o ganho do método proposto, pois como ele demanda um número menor de avaliações, o impacto provocado por um conjunto mais amplo de atributos é menor.

A utilização do *vértice central* de uma árvore como ponto de partida para a estratégia de busca elementar (divisão de uma árvore em duas) é uma solução interessante, pois proporciona, como vantagens, evitar ótimos locais e diminuir a amplitude de exploração do espaço de soluções.

O critério de parada da estratégia é quem determina a amplitude do processo de busca pela melhor divisão de uma árvore. Valores elevados fazem o processo de busca explorar grande parte das árvores, fazendo cair a velocidade da geração de resultados. Valores pequenos, diferentes de zero, permitem à estratégia um certo grau de exploração, fornecendo grande economia no número de avaliações e, consequentemente, no tempo de execução. Nos testes realizados, a qualidade da partição mostrou uma queda pequena em relação ao ganho no tempo de execução do procedimento e, além disso, à medida que a AGM sofre desmembramentos sucessivos, há uma convergência na qualidade dos resultados em relação à busca exaustiva.

7.3 – EM RELAÇÃO À ABORDAGEM INTERATIVA

O objetivo desta proposta não foi oferecer um procedimento completo e pronto para a regionalização guiada pelo usuário, e sim, discutir as vantagem e aplicações para uma abordagem de regionalização interativa, com base na experiência obtida com o desenvolvimento do método de regionalização alternativo e estabelecer um conjunto de requisitos básicos que um processo deste tipo deveria atender.

Na proposta apresentada, a AGM (e o grafo CAD) desempenha vários papeis importantes: ela representa a status da regionalização; é utilizada para a criação de um histórico do processamento e geração do diagrama de desmembramento (*dendrograma*); e ainda na edição de partições e extração de informação. Alguns testes iniciais foram realizados no ambiente de experimentação e a estrutura da AGM se mostrou capaz de suportar a maioria das funcionalidades requeridas.

7.4 – TRABALHOS FUTUROS

Para soluções que consideram a estrutura de vizinhança as abordagens mais promissoras são as que buscam a simplificação do problema, tal como o uso da AGM. Outras formas de simplificação poderiam ser estudadas. Mas, ainda seguindo a linha deste trabalho, dois caminhos podem ser experimentados no sentido de tentar melhorar o desempenho do método: buscar a diminuição no número de avaliações ou diminuir o custo computacional de cada avaliação.

A primeira alternativa seria possível com o desenvolvimento de nova estratégia de busca elementar, utilizando variações da idéia aqui apresentada ou experimentando outras técnicas de otimização. O segundo caminho, procuraria simplificar a avaliação, buscando outros critérios de agrupamento que demandem menos esforço computacional.

Em relação à abordagem interativa, seria interessante que a proposta fosse desenvolvida, dando origem a um sistema de apoio ao projeto de regiões, onde diversas possibilidades seriam oferecidas ao analista, como ferramentas de exploração de dados, escolha de restrições e parâmetros. Além da classificação propriamente dita, esta ferramenta poderia ser utilizada na análise exploratória de dados espaciais, no estudo do efeito de variação da unidade básica de área e, ainda, na detecção de regimes espaciais.

Uma vez que nos experimentos aqui realizados, foi utilizado um critério de agrupamento voltado para a homogeneidade interna das regiões, os desempenhos dos métodos de regionalização poderiam ser experimentados utilizando critérios que

considerassem também a inter-relação entre as regiões, como por exemplo, a separação entre as classes.

7.5 – CONCLUSÃO FINAL

Este trabalho apresentou uma contribuição específica ao procedimento de classificação, aplicado a objetos espaciais com representação poligonal, que visa reagrupar unidades de área em regiões mais abrangentes. Embora este esforço e os seus resultados representem uma pequena contribuição diante dos desafios e possibilidades existentes na integração de técnicas de análise espacial e sistema de informação geográfica, temos a expectativa que ele possa ser útil em situações reais e possa estimular novos desenvolvimentos.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVANIDES, S.; OPENSHAW, S.; REES, P. Designing Your Own Geographies. In: REES, P.; MARTIN, D.; WILLIAMSON, P., Eds. **The Census Data System**. Chichester: Jonh Wiley & Sons, 2002.

ANDEBERG, M.R. **Cluster analysis for applications**. New York: Academic Press, Probability and mathematical statistics. 19. 1973. 359 p.

ANSELIN, L. **Space Stat, version 1.80: User Guide**. 1995.

ASSUNÇÃO, R.M.; LAGE, J.P.; A.REIS, E.; SILVA, P.L.N. *Comunicaçao Pessoal* - Análise de conglomerados espaciais via árvore geradora mínima. 2000.

BAILEY, T.C.; GATRELL, A.C. **Interactive spatial data analysis**. Harlow: Longman, 1995.

BURROUGH, P.A.; MCDONNEL, R.A. **Principles of Geographical Information Systems**. Oxford: University Oxford Press, 1998. 333 p.

BUTTENFIELD, B.; GAHEGAN, M.; MILLER, H.; YUAN, M. **Geospatial Data Mining and Knowledge Discovery**. 2001.

CÂMARA, G.; MEDEIROS, J.S.D. Princípios básicos em geoprocessamento. In: ASSAD, E.D.; SANO, E.E., Eds. **Sistemas de Informações Geográficas: Aplicações na Agricultura**. Brasília: Embrapa, 1998. p. 3-11.

CLIFF, A.D.; HAGGETT, P.; ORD, K.; BASSETT, K.; DAVIES, R. **Elements of Spatial Structure.** Cambridge: CUP, 1975.

FISCHER, M.M.; GETIS, A. Advances in Spatial Analysis. In: FISCHER, M.M.; GETIS, A., Eds. **Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling, and Computational Intelligence**. Berlin: Springer, Advances in Spatial Science, 1997. p. 1-12.

FISCHER, M.M.; SCHOLTEN, H.J.; UNWIN, D. Geographic information systems, spatial data analysis and spatial modelling: an introduction. In: FISCHER, M.M.; SCHOLTEN, H.J.; UNKIN, D., Eds. **Spatial Analytical Perspectives on GIS**. London: Taylor & Francis, Gisdata IV, 1996. p. 3-19.

FOTHERINGHAM, A.S.; BRUNSDON, C.; CHARLTON, M. **Quantitative Geography: Perspectives on Spatial Data Analysis**. London: SAGE, 2000. 270 p.

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. v. 1, n. 1, p. 20-33, 1999.

GOODCHILD, M.F.; STEYAERT, L.T.; PARKS, B.O.; JOHNSTON, C.; MAIDMENT, D.; CRANE, M.; GLENDINNING, S., Eds. **GIS and environment modeling: progress and research Issues**. Fort Collins: GIS Word Books, 1996.

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. **SIGKDD Explorations**. v. 1, n. 1, p. 20-33, 1999.

GORDON, A.D. **Classification - methods for the exploratory analysis of multivariate data**. London: Chapmam and Hall, 1981. 193 p.

GORDON, A.D. A survey of constrained classification. **Computational Statistics & Data Analysis**. v. 21, n. p. 17-29, 1996.

HAINING, R. Designing spatial data analysis modules for geographical information system. In: FOTHERINGHAM, S.; ROGERSON, P., Eds. **Spatial Analysis and GIS**. London: Taylor & Francis, 1995. p. 45-63.

JUNGNICKEL, D., Ed. **Graphs, Networks and Algorithms**. Berlin: Springer, Algorithms and Computation in Mathematics. **5**. 1999. 589 p.

KARYPIS, G.; KUMAR, V. **A fast and high quality multilevel scheme for partitioning irregular graphs**. University of Minnesota: 1998.http://www-users.cs.umn.edu/~karypis/publications/partitioning.html. (10/01/2003).

KAUFMAN, L.; ROUSSEEUW, P.J. **Finding groups in data: an introduction to cluster analysis**. Jonh Wiley & Sons, 1990.

KOPERSKI, K.; ADHIKARY, J.; HAN, J. **Spatial Data Mining: progress and challenges survey paper**. Simon Fraser University, 1997.

LAGE, J.P.; ASSUNÇÃO, R.M.; REIS, E.A. **A Minimal Spanning Tree Algorithm Applied to Spatial Cluster Analysis**. Elsevier Science Publishers, 2001. http://www.elsevier.nl/gej-ng/31/29/24/39/23/94/endm7063.ps.

LAGUNA, M. A guide to Implementing Tabu Search. **Investigación Operativa**. v. 4, n. 1, p. 5-25, 1994.

LONGLEY, P.; BATTY, M. Analysis, modelling, forecasting, and GIS technology. In: LONGLEY, P.; BATTY, M., Eds. **Spatial Analysis: Modelling in a GIS Environment**. New York: John Wiley & Sons, 1996. p. 1-15.

MA, J.; HAINING, R.P.; WISE, S.M. **SAGE user's guide**. University of Sheffield, v. 2001, n. 09/02/2001, 1997.

MARAVALLE, M.; SIMEONE, B. A Spanning Tree Heuristic for Regional Clustering. **Communications in Statistics - Theory Methods**. v. 24, n. 3, p. 625-639, 1995.

MARAVALLE, M.; SIMEONE, B.; NALDINI, R. Clustering on trees. **Computational Statistics & Data Analysis**. v. 24, n. p. 217-234, 1997.

MARSH, W.M.; GROSSA JR, J.M. **Environmental geography, science, land use, and earth systems**. New York: John Wiley & Sons, 1996. 426 p.

MARTIN, D. Optimizing census geography: the separation of collection and output geographies. **International Journal of Geographical Information Science**. v. 12, n. p. 673-685, 1998.

NG, R.T.; HAN, J. Efficient and effective clustering methods for Spatial Data Mining. In: Twentieth International Conference on Very Large Data Base, 1994, Santiago. **Anais.**, p. 144-155.

OPENSHAW, S., Ed. **Census Users Handbook**. Cambridge: Geoinformation International, 1995. 450 p.

OPENSHAW, S.; ALVANIDES, S. Designing zoning systems for representation of socio-economic data. In: FRANK, I.; RAPER, J.; CLEYLAN, J., Eds. **Time and Motion of Socio-Economic Units**. London: Taylor and Francis, GISDATA Series, 2001.

OPENSHAW, S.; ALVANIDES, S.; WHALLEY, S. **Some further experiments with designing output areas for the 2001 UK census**. University of Leeds: 1998.http://www.geog.leeds.ac.uk/pgrads/s.alvanides/zdes3.html. (30/10/2001).

OPENSHAW, S.; RAO, L. **Re-engineering 1991 census geography: serial and parallel algorithms for unconstrained zone design**. University of Leeds: 1995.http://www.geog.leeds.ac.uk/papers/93-3/. (21/01/2003).

OPENSHAW, S.; WYMER, C. Classifying and regionalizing census data. In: OPENSHAW, S., Ed. **Census users' handbook**. Cambridge: GeoInformation International, 1995. p. 460.

RICHARDS, J.A. **Remote Sensing Digital Image Analysis - An Introduction.** Berlin: Springer-Verlag, 1995. 340 p.

SCHOWENGERDT, R.A. **Remote Sensing, models and methods for image processing**. San Diego: Academic Press, 1997.

SPOSATI, A. Mapa de Exclusão/Inclusão da Cidade de São Paulo. São Paulo.: Editora PUC-SP, 1996.

WISE, S.; HAINING, R.; MA, J. Regionalisation Tools for The Exploratory Spatial Analysis od Health Data. In: FISCHER, M.M.; GETIS, A., Eds. **Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling, and Computational Intelligence**. Berlin: Springer, 1997. p. 83-100.

ZHANG, B.; HSU, M.; DAYAL, U. K-harmonic means - A spatial clustering algorithm with bossting. In: RODDICK, J.F.; HORNSBY, K., Eds. **Temporal, Spatial and Spatio-Temporal Data Mining**. Berlin: Springer, Lecture Notes in Artificial Intelligence, 2001. p. 31-45.