# Procedimentos Automáticos e Semi-automáticos de Regionalização por Árvore Geradora Mínima

MARCOS CORRÊA NEVES[1]
GILBERTO CÂMARA[2]
RENATO M. ASSUNÇÃO[3]
CORINA DA COSTA FREITAS[2]

[1]EMBRAPA Meio Ambiente, Caixa Postal 69, 13820-000, Jaguariúna, SP, Brasil
marcos@sigmanet.com.br
[2]Instituto Nacional de Pesquisas Espaciais - INPE, Caixa Postal 515, 12201-027, São José dos Campos, SP, Brasil
gilberto@dpi.inpe.br; corina@dpi.inpe.br
[3]Departamento de Estatística – Universidade Federal de Minas Gerais, Caixa Postal 702, 30123-970, Belo Horizonte, MG, Brasil – assuncao@est.ufmg.br

**Abstract.** Regionalization is a aggregation procedure of contiguous areas in homogeneous regions. This paper describes two proposals for the regionalization process. The first proposal uses optimization techniques to increase the efficiency of the process. The second, proposes an approach of regionalization guided by the analyst. The paper presents also some results of initial experiments and possible contributions of these two proposals.

## 1 Introdução

Regionalização é o procedimento de agrupamento de objetos-área em regiões homogêneas e contíguas no espaço. A regionalização busca uma nova repartição do espaço de estudo em um número menor de objetos e resultando em novas áreas (regiões) com dimensões geográficas mais abrangentes.

Alguns motivos para se agrupar unidades espaciais básicas, como setores censitários, em regiões maiores são: aumento da representatividade dos valores dos atributos e taxas associadas às unidades de área; redução dos efeitos da imprecisão nos valores das variáveis; redução dos erros associados ao posicionamento geográfico de eventos; e redução no custo de análise dos dados (Wise et al., 1997), (Openshaw et al., 1995).

No escopo deste trabalho, são avaliadas diferentes abordagens para o processo de regionalização e algumas técnicas de *Análise de Cluster*. Estas técnicas são amplamente utilizadas como procedimentos de classificação, permitindo extrair estruturas existentes no conjunto de dados sem nenhum conhecimento prévio. Este problema é freqüentemente definido como a procura por *grupos naturais*, ou seja, agrupar objetos de modo que o grau de associação (*similaridade*) seja alto entre objetos de um mesmo grupo e baixo entre membros de grupos diferentes (Andeberg, 1973), (Gordon, 1981). São técnicas aplicáveis também a dados espaciais, sendo utilizadas em processos de mineração de dados (*Spatial Data Mining*) e análise exploratória de dados *(ESDA)* (Goebel et al., 1999), (Ng et al., 1994), (Koperski et al., 1997), (Zhang et al., 2001).

Temos dois objetivos com este trabalho: primeiro, propor um método de regionalização eficiente utilizando técnicas de otimização; e segundo, propor um sistema de apoio ao processo de regionalização que contemple: a escolha entre o procedimento automático ou dirigido pelo analista; a extração de várias informações estatísticas e gráficas referentes aos objetos ou a um subconjunto deles; avaliações quantitativas sobre os agrupamentos obtidos com a regionalização; e modificações durante o processo de regionalização através da interferência do analista.

Este trabalho está organizado da seguinte forma: Na seção 2 são apresentadas as diferentes abordagens utilizadas em procedimentos de regionalização. Na seção 3 é mostrado um processo de regionalização via árvore geradora mínima. Na seção 4 é descrita a proposta de regionalização utilizando técnicas de otimização. A justificativa para o processo de regionalização guiado pelo usuário em contraposição às abordagens automáticas é

apresentada na seção 5. Na última seção estão algumas conclusões preliminares em função dos resultados até agora obtidos.

## 2 Abordagens de regionalização

Existem três abordagens utilizadas para a condução da regionalização. Na primeira abordagem, o processo é realizado em dois estágios independentes. No primeiro estágio não é considerada a informação espacial e um procedimento de *clustering* convencional é executado, utilizando somente os atributos não-espaciais. No segundo estágio, os *clusters* são reavaliados observando as relações de vizinhança dos objetos. Assim, objetos similares agrupados em um *cluster* na fase inicial, mas sem contigüidade espacial, serão 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 et al. (1995), refere-se ao controle da similaridade entre os objetos de uma mesma região que é garantido pelo primeiro estágio do processo. O inconveniente desta abordagem está na falta de controle sobre o número de regiões resultantes. Os casos com pequena dependência espacial entre os objetos, por exemplo, tenderão a produzir um número elevado de regiões (Wise et al., 1997).

Na segunda abordagem, a similaridade entre os objetos é avaliada considerando simultaneamente a posição geográfica dos objetos e seus atributos não-espaciais. As coordenadas do centróide são consideradas como atributos adicionais ao problema. Então, são utilizados algoritmos comuns de *clustering*, onde a avaliação da similaridade contém duas componentes ponderadas, uma para o espaço de atributos e a outra para a distância geográfica. Se o peso dado para a componente correspondente à distância geográfica for forte o suficiente, os grupos resultantes do processo de classificação serão contíguos. Este tipo de abordagem possui a dificuldade de se estabelecer os pesos ideais para as duas componentes.

Esta abordagem é utilizada pelo sistema SAGE (*Spatial Analysis in a GIS Environment*) em seu método de regionalização (Ma et al., 1997). Ele utiliza um procedimento de *clustering* baseado na técnica de particionamento *k-médias*, cuja a função objetivo é formada a partir de três critérios: a) *homogeneidade*: regiões formadas por objetos similares, considerando atributos não-espaciais; b) *compacidade*: as coordenadas dos 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). O SAGE considera a componente *igualdade* para produzir regiões equilibradas em relação a um determinado atributo, de modo a possibilitar comparações apropriadas entre as regiões (exemplo: taxa de mortalidade por câncer). O atributo é definido pelo usuário do SAGE (Wise et al., 1997).

Openshaw et al. (1998) critica a abordagem utilizada no procedimento de regionalização do sistema SAGE e em outros trabalhos anteriores que utilizam estratégias semelhantes (Cliff et al., 1975), (Martin, 1998). Nestes métodos, cada componente da função objetivo, é uma função isolada e utilizam em seus cálculos variáveis que medem fenômenos distintos e são medidos em unidades diferentes. Openshaw et al. (1998) afirma também que uma estratégia melhor e mais simples é selecionar uma das funções componentes como 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). A forma indicada para tratar com restrições em problemas de otimização é adicionar à função objetivo, funções de penalização que refletem a violação das restrições. Esta solução foi utilizada pelos autores no desenvolvimento do seu sistema de zoneamento automático denominado de ZDES.

Na terceira abordagem utilizada em procedimentos de regionalização, o relacionamento de vizinhança entre os objetos é explicitado por meio de dispositivos auxiliares, como uma matriz ou um grafo. No caso do uso de uma matriz, ela é chamada de *matriz de contigüidade (C)*, onde cada elemento, $c_{ij}$, indica se os objetos $i$ e $j$ são contíguos ou não. Desta forma, $c_{ij} = 0$ para objetos não contíguos, e $c_{ij} = 1$ para objetos contíguos. De forma equivalente, quando é utilizado *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). A Figura 1.a mostra a representação da estrutura espacial por meio de um grafo.

## 3 Regionalização via Árvore Geradora Mínima

Nesta seção destacaremos um método de regionalização apresentado em Assunção et al. (2000). Neste método, é gerada uma árvore a partir do grafo correspondente ao conjunto de dados. Esta árvore é escolhida de forma a garantir que a soma dos custos associados às arestas seja a menor possível. Tal árvore é chamada de árvore geradora mínima (AGM).

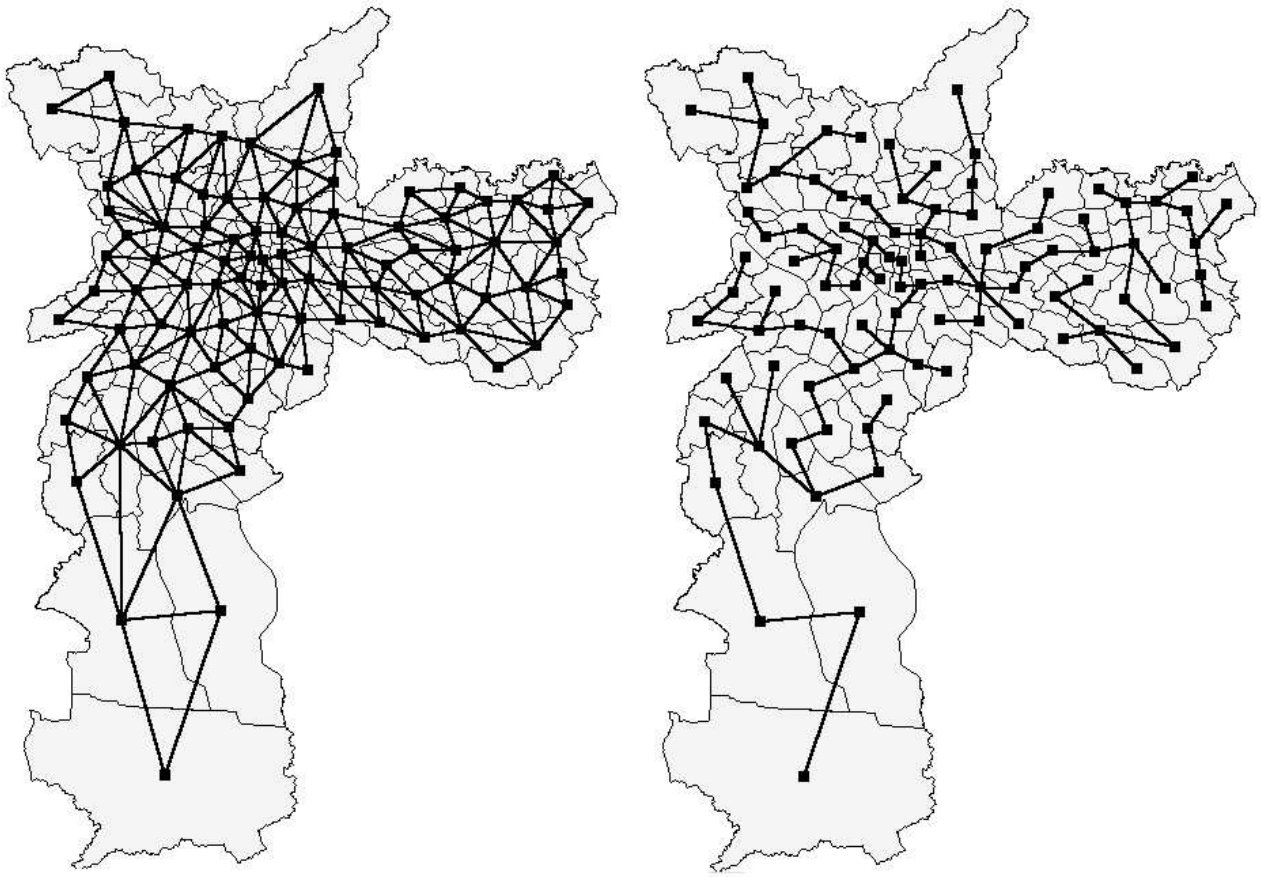

Figura 1: a) grafo representado as relações de vizinhança. b) uma árvore geradora mínima, a partir da escolha de um conjunto de atributos.

A AGM é obtida a partir do grafo por meio da utilização do algoritmo de PRIM. Este algoritmo está descrito em Assunção et al. (2000) e em livros de *Teoria de Grafo* como Jungnickel (1999). A Figura 1.b, ilustra uma AGM obtida a partir do grafo da Figura 1.a. O custo da aresta é inversamente proporcional à similaridade entre os objetos.

**Poda da AGM**

Após a geração da AGM, o método passa a uma segunda fase denominada de poda da árvore. Esta fase consiste em retirar as arestas mais caras. Cada aresta retirada provoca uma divisão na árvore, resultando em duas sub-árvores desconectadas. Serão escolhidas *k-1* arestas, para obter *k* regiões.

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

$$\textit{Custo da aresta } l = SSD_T - SSD_l \ ,$$

onde:

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

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

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 $j$, dado por:

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \ .$$

ii) $SSD_l$ é a soma das duas parcelas obtidas da *soma dos quadrados dos desvios* das duas sub-árvores , $T_a$ e $T_b$, geradas pela retirada da aresta $l$ da árvore $T$:

$$SSD_l = SSD_{Ta} + SSD_{Tb} \quad (1).$$

Para se obter a *soma dos quadrados dos desvios* para as duas sub-árvores, são calculados os valores médios dos $m$ atributos, tal como feito para o cálculo de $SSD_T$, porém, considerando-se apenas os atributos referentes aos objetos pertencentes a cada sub-árvore de $T$, $T_a$ e $T_b$.

Embora não exista uma forma objetiva para avaliação dos procedimentos de *clustering* em geral, 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 utilizado em métodos como *k-médias*, que reconhecidamente, produz bons resultados (Openshaw, 1995).

## 4 Heurística para a otimização

No método descrito na seção anterior, para cada partição da AGM, é avaliado os custos de todas as arestas existentes na árvore para se identificar a mais cara, ou seja, é realiza uma busca exaustiva pela melhor solução.

Em casos que envolvam um número elevado de objetos e atributos, o esforço computacional para avaliar todas as soluções cresce significativamente. Para melhorar a eficiência do método e sua aplicabilidade a maiores volumes de dados, propomos utilizar técnicas de otimização na fase de poda da AGM.

Visto como um problema de otimização, podemos caracterizar a procura pela a aresta de custo mais elevado, como a busca por uma solução ótima em um espaço de soluções $S = \{S_1, S_2, ..., S_{n-1}\}$. A exploração do espaço de soluções é feita de modo que uma solução aceitável seja identificada sem que haja necessidade de visitar todas as soluções possíveis. A função objetivo que orienta o processo de busca, é dada por:

$$f : S \rightarrow \Re,$$
$$f = SSD_l,$$

onde: $SSD_l$ é a *soma dos quadrados dos desvios* das duas sub-árvores, retirada a aresta $l$ da AGM, conforme definido anteriormente pela *equação (1)*.

Uma solução $S_i$ equivale a escolha de uma aresta $i$ dentre as $n-1$ arestas da árvore. Uma solução vizinha à $S_i$, é representada pela retirada de uma aresta que possui um dos vértices em comum com $S_i$. Assim, adotamos um mecanismo de busca pela vizinhança (*busca local*) procurando obter valores de $f(S_i)$ cada vez menores.

A estratégia de busca é apresentada a seguir. Ela combina elementos presentes em métodos de busca local como *Expansão da Vizinhança* e *Busca Tabu* (Laguna, 1994):

**Passo1:** escolha da solução inicial $S_i$ em $S$. Fazer $S^* = S_i$; $k^* = k = 0$; e incluir $S_i$ na lista de soluções visitadas ($T$);

**Passo 2:** Fazer $k = k + 1$. Escolher na lista $T$ a solução que terá a vizinhança expandida, gerando o conjunto de soluções promissoras ($V^*$);

**Passo 3:** Avaliar as soluções em $V^*$; Armazenar soluções avaliadas em $T$;escolher a melhor solução, $S_j$, em $V^*$.

**Passo 4:** Se $f(S_j) < f(S^*)$, então: $S^* = S_j$; $k^* = k$;

**Passo 5:** Verificar condição de parada ($k - k^* > 8$). Senão satisfeita, voltar ao *passo 2*.

O conjunto de soluções promissoras, $V^*$, indica as soluções que serão avaliadas na iteração corrente e correspondem à expansão na vizinhança de uma solução já avaliada, escolhida em $T$ no passo 2. Para esta escolha é utilizada uma função de avaliação:

$$f' = \max(SSD_{Ta}, SSD_{Tb}).$$

Será escolhida a solução que apresentar um menor valor para $f'$ e que ainda tenha ao menos uma solução vizinha ainda sem avaliação.

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. Esta função evita que estes ramos sejam verificados dando prioridade para soluções que particionam a árvore em dois grupos mais homogêneos e equilibrados.

A Figura 2 mostra as soluções investigadas em um experimento, onde foi escolhido como ponto de partida uma aresta incidente em um nó-folha distante da solução ótima. Na Tabela de Soluções Visitadas, apresentada a seguir, estão os valores de $f$ e $f'$ utilizados para direcionar o processo de busca. 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 3 mostra os valores obtidos para a função objetivo, $f$, para cada iteração. Podemos observar um mínimo local neste gráfico ($k = 6$). Para ultrapassá-lo,

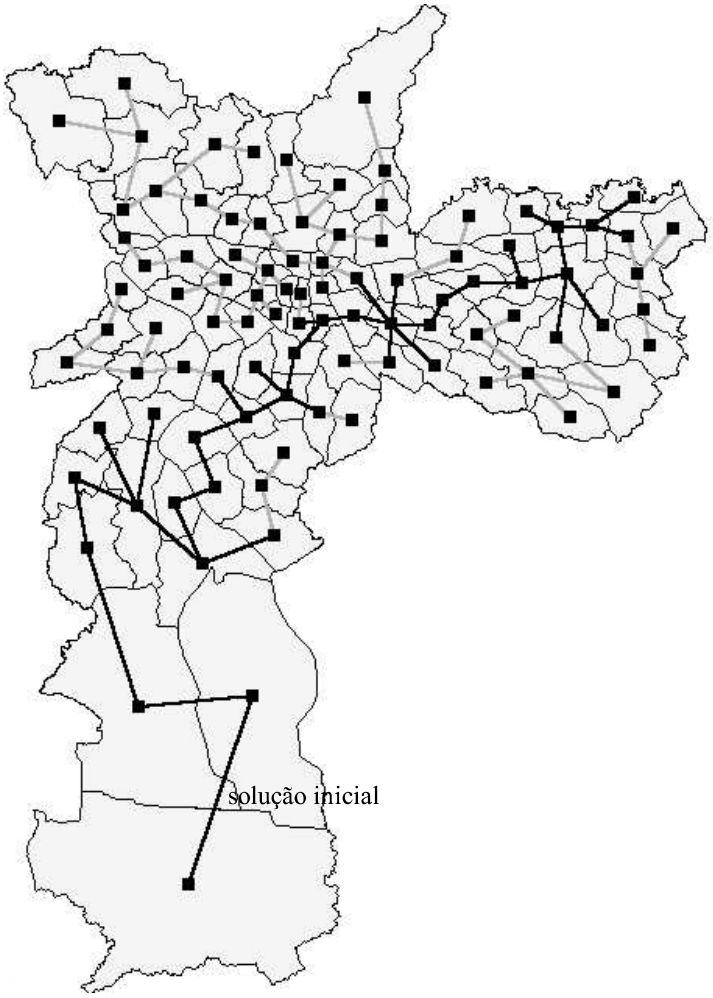


Figura 2: Soluções visitadas no processo de busca.

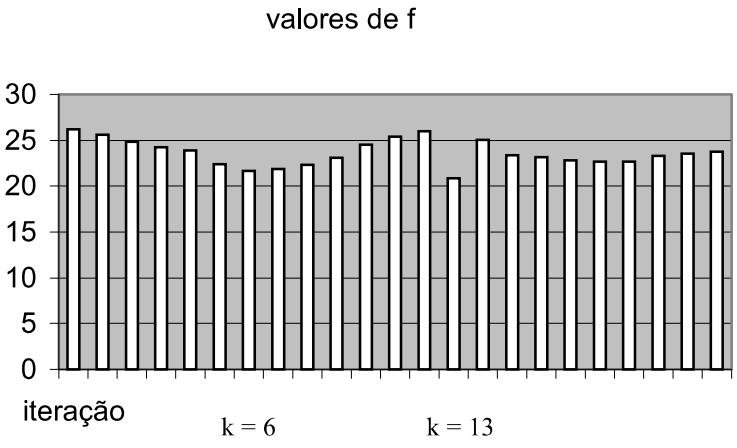


Figura 3: Valores de *f* ao longo da busca.

Tabela de soluções visitadas: Valores obtidos para as soluções avaliadas durante o processo de busca.

Obs.: as colunas *elems.A* e *elems.B* indicam o número de objetos nas sub-árvores Ta e Tb.

| k | k* | f | f' | $SSE_{Ta}$ | $SSE_{Tb}$ | elems.A | elems.B |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 26.190 | 26.190 | 26.190 | 0.000 | 95 | 1 |
| 1 | 1 | 25.574 | 25.406 | 25.406 | 0.168 | 94 | 2 |
| 2 | 2 | 24.848 | 24.608 | 24.608 | 0.239 | 93 | 3 |
| 3 | 3 | 24.238 | 23.942 | 23.942 | 0.296 | 92 | 4 |
| 4 | 4 | 23.860 | 23.502 | 23.502 | 0.357 | 91 | 5 |
| 5 | 5 | 22.422 | 21.870 | 21.870 | 0.552 | 88 | 8 |
| 5 | 5 | 26.367 | 26.368 | 0.000 | 26.368 | 1 | 95 |
| 5 | 5 | 25.762 | 25.762 | 0.000 | 25.762 | 1 | 95 |
| 6 | 6 | 21.632 | 20.794 | 20.794 | 0.838 | 84 | 12 |
| 6 | 6 | 26.090 | 26.041 | 0.049 | 26.041 | 3 | 93 |
| 7 | 6 | 21.839 | 20.726 | 20.726 | 1.112 | 83 | 13 |
| 8 | 6 | 22.300 | 20.629 | 20.629 | 1.670 | 82 | 14 |
| 9 | 6 | 23.086 | 20.308 | 20.308 | 2.779 | 81 | 15 |
| 10 | 6 | 24.523 | 19.468 | 19.468 | 5.055 | 73 | 23 |
| 10 | 6 | 26.530 | 26.028 | 0.501 | 26.028 | 7 | 89 |
| 11 | 6 | 25.801 | 17.547 | 17.547 | 8.254 | 69 | 27 |
| 11 | 6 | 26.151 | 26.125 | 0.025 | 26.125 | 2 | 94 |
| 11 | 6 | 25.531 | 25.358 | 0.000 | 25.358 | 1 | 95 |
| 12 | 6 | 26.019 | 17.192 | 17.192 | 8.826 | 68 | 28 |
| 13 | 6 | 25.500 | 15.453 | 9.747 | 15.453 | 53 | 43 |
| 13 | 13 | 20.870 | 18.954 | 1.916 | 18.954 | 14 | 82 |
| 14 | 13 | 25.014 | 15.533 | 9.481 | 15.533 | 52 | 44 |
| 15 | 13 | 26.515 | 26.513 | 0.002 | 26.513 | 2 | 94 |
| 15 | 13 | 23.354 | 21.006 | 2.348 | 21.006 | 23 | 73 |
| 15 | 13 | 26.160 | 26.106 | 0.054 | 26.106 | 3 | 93 |
| 15 | 13 | 26.411 | 26.411 | 0.000 | 26.411 | 1 | 95 |
| 15 | 13 | 26.553 | 22.038 | 4.515 | 22.038 | 22 | 74 |
| 16 | 13 | 23.150 | 21.008 | 2.142 | 21.008 | 22 | 74 |
| 17 | 13 | 22.785 | 21.032 | 1.752 | 21.032 | 21 | 75 |
| 18 | 13 | 22.643 | 21.045 | 1.598 | 21.045 | 20 | 76 |
| 19 | 13 | 26.579 | 26.579 | 0.000 | 26.579 | 1 | 95 |
| 19 | 13 | 22.698 | 21.209 | 1.489 | 21.209 | 18 | 78 |
| 20 | 13 | 23.265 | 22.863 | 0.402 | 22.863 | 9 | 87 |
| 20 | 13 | 25.697 | 25.453 | 0.243 | 25.453 | 7 | 89 |
| 20 | 13 | 26.571 | 26.571 | 0.000 | 26.571 | 1 | 95 |
| 21 | 13 | 26.478 | 26.478 | 0.000 | 26.478 | 1 | 95 |
| 21 | 13 | 23.539 | 23.280 | 0.259 | 23.280 | 7 | 89 |
| 22 | 13 | 26.497 | 26.497 | 0.000 | 26.497 | 1 | 85 |
| 22 | 13 | 23.765 | 23.719 | 0.046 | 23.719 | 5 | 91 |

a estratégia de busca teve que insistir durante 6 iterações até que o valor de *f* voltasse a diminuir.

Uma melhoria para esta estratégia de busca é escolher um ponto de partida mais conveniente. Considerando que os nós-folhas correspondem a objetos com valores de atributos mais extremos e os nós centrais, a objetos com valores intermediários, é mais indicado que o algoritmo parta de uma aresta posicionada mais ao centro do grafo. Este fato é facilmente verificado na tabela, onde soluções correspondentes a arestas que dividem de forma equilibrada o número de objetos em cada sub-árvores, tendem a apresentar menores valores para *f*. Evidentemente existe um custo adicional para se identificar um ponto de partida conveniente.

A estratégia descrita atente apenas à retirada de uma única aresta. Para aplicarmos a um problema geral, envolvendo a retirada de *n* arestas sucessivas, há necessidade de se incluir um laço de controle externo que será executado *n-1* vezes. Após cada subdivisão da árvore, a nova busca pela aresta mais cara pode, por exemplo, ser orientada para a sub-árvore que apresentar uma maior *soma dos quadrados dos devios*.

Esta estratégia de busca ainda não foi testada em profundidade, mas os primeiros resultados são promissores. A definição do critério de parada é um ponto crítico. Se o critério escolhido for muito sensível poderá conduzir o processo para um mínimo local. Se for muito elevado poderá ocasionar um número exagerado de verificações antes do seu encerramento.

## 5 Abordagem semi-automática

Nesta seção discutiremos a utilidade de conduzir a regionalização como um processo interativo. Deste modo, o analista tem a possibilidade de interferir, não somente no início do procedimento, mas direcionando a sua execução, analisando, comparando e reagrupando os objetos. Defendemos esta idéia baseados nas características dos procedimentos de classificação, mais especificamente nas técnicas de *Análise de Cluster*, e na possibilidade de utilizar a AGM como elemento estruturante dos objetos.

Uma característica negativa dos procedimentos de classificação automáticos é a falta de controle durante o processo. O analista define parâmetros de entrada (como número de regiões, atributos, restrições,...) e obtém uma classificação sem identificar exatamente o que ocorreu durante o processo e como cada parâmetro interferiu no resultado. Na regionalização dirigida pelo analista, a classificação pode ocorrer passo a passo, permitindo obter medidas quantitativas sobre os objetos e verificar a influência dos parâmetros e restrições, extraindo assim um maior conhecimento dos dados.

Na abordagem dirigida que propomos, a AGM desempenha um papel central. Como já dissemos, a *Análise de Cluster* é utilizada para se descobrir estruturas existentes em um conjunto de dados. A AGM explicita os relacionamentos existentes entre os objetos, pois cada objeto aparece na árvore ligado à objetos vizinhos, com os quais possui maior similaridade. Assim, a árvore mostra claramente as 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 obtidos por subdivisões sucessivas na árvore.

Além das relações de similaridades entre os objetos, outras informações podem ser extraídas da árvore. Por exemplo: Nós com apenas uma aresta incidente (nós-folha) tendem a corresponder a objetos com valores de atributos extremados; Nós com mais de uma aresta, são objetos que tendem a ter valores de atributos centrais, em relação aos nós e aos ramos vizinhos. Portanto, a AGM fornece informações importantes que não estão visíveis nos mapas em cores, que é a forma tradicional de visualização dos resultados da regionalização. Estas informações adicionais corroboram para que a AGM seja utilizada como uma ferramenta auxiliar para análise dos dados e não somente como um passo intermediário e escondido dentro de um procedimento automático.

Nesta nova abordagem a estrutura da AGM é utilizada também na seleção de um subconjunto de objetos similares, para os quais podem ser extraídos gráficos, informações quantitativas ou sumários estatísticos, ajudando a identificar as características de diferentes agrupamentos de objetos e a desenvolver um conhecimento sobre os dados e a sua relação com o espaço.

Uma terceira função desempenhada pela AGM é que sua estrutura é ainda aproveitada como representação gráfica do processo de regionalização. A medida que regiões são definidas, por exemplo, arestas da árvore são apagadas. O analista pode atuar no processo por meio da "edição" da árvore, apagando ou criando arestas. Esta funcionalidade é importante para atender casos onde haja necessidade de: corrigir um problema provocado pela restrição espacial; redistribuir as áreas segundo um critério não considerado anteriormente; dividir regiões homogêneas demasiadamente extensas; ou ainda, aglutinação de regiões pequenas.

Um sistema experimental está em desenvolvimento onde as duas propostas aqui apresentadas poderão ser melhor avaliadas. Este sistema utiliza a TerraLib que é uma biblioteca de classes e funções para a construção de aplicativos geográficos (www.dpi.inpe.br/terralib.html).

## 6 Conclusão

A utilização de técnicas de otimização pode melhorar a eficiência do método via árvore geradora mínima permitindo sua aplicação a problemas de regionalização que envolvam um número maior de objetos e atributos.

A AGM é um mecanismo útil para os procedimentos de regionalização pois explicita os relacionamentos entres os objetos similares e reduzem o número de soluções possíveis.

A AGM contém 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.

Acreditamos que a possibilidade de conduzir o procedimento de regionalização como um processo dirigido pelo analista pode trazer algumas vantagens. A AGM pode ser utilizada como um elemento estruturante dos dados, permitindo ao usuário selecionar, comparar e editar os objetos e as regiões, atendendo a casos específicos e a critérios subjetivos difíceis de serem considerados em procedimentos totalmente automáticos.

## Referências


ANDEBERG, M.R. *Cluster analysis for applications*. New York: Academic Press, 1973.

ASSUNÇÃO, R.M.; LAGE, J.P.; A.REIS, E.; SILVA, P.L.N. *Análise de conglomerados espaciais via árvore geradora mínima*. 2000.(comunicação pessoal).

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

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

GORDON, A.D. *Classification - methods for the exploratory analysis of multivariate data*. 1 ed. London: Chapmam and Hall, 1981.

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

JUNGNICKEL, D. Graphs, Networks and Algorithms. In: BECKER, E.; BRONSTEIN, M.; COHEN, H.; EISENBUD, D.; GILMAN, R. eds. *Algorithms and Computation in Mathematics*. Berlin: Springer, v.5, 1999.

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

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

MA, J.; HAINING, R.P.; WISE, S.M. *SAGE user's guide*. 1997.

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

MARTIN, D. Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science*, v. 12, 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*, Santiago, 1994.

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

OPENSHAW, S.; ALVANIDES, S.; WHALLEY, S. *Some further experiments with designing output areas for the 2001 UK census*. University of Leeds, 1998.

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

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, Lecture Notes in Artificial Intelligence*. Berlin: Springer, 2001, p. 31-45.