Análise de Conglomerados Espaciais

Renato M. Assunção
Departamento de Estatística - Universidade Federal de Minas Gerais

Resumo

Este capítulo trata do problema de particionar uma região em sub-regiões homogêneas.
Palavras-Chave: análise de conglomerados, medidas de dissimilaridade.

1 Introdução

Este capítulo apresenta uma metodologia para agrupamentos de áreas, tais como setores censitários ou municípios, levando em consideração restrições de contiguidade. Esses métodos são chamados de regionalização em economia regional, geografia e epidemiologia e objetivam agrupar áreas em regiões disjuntas tais que os grupos obtidos possuem uma grande homogeneidade interna com relação a atributos de interesse e, ao mesmo tempo, são heterogêneos entre si. As características em consideração constituem o perfil das áreas.

Métodos usuais de conglomerados abordam o problema de agrupar objetos em grupos homogêneos, nosso problema reside em realizar uma análise de conglomerados quando os objetos possuem uma localização espacial. Em particular, nosso foco recai sobre a situação em que temos um mapa particionado em áreas, cada uma delas com uma posição geográfica determinada. Em geral, a posição de cada área é o centróide geográfico (o centro de massa do polígono que determina a área).

Assim, nosso objetivo consiste em particionar um território em regiões que possuem uma grande homogeneidade interna com relação a atributos de interesse tais como características sociais e econômicas ou aspectos geofísicos. O mapa da Figura 1 apresenta o município de São João do Meriti dividido em 353 pequenas áreas (setores cenistários do IBGE), cada uma delas contendo entre 300 e 500 domicílios, aproximadamente. Para cada uma dessas pequenas áreas, existem informações sociais e econômicas dos seus habitantes, tais como a renda média de seus habitantes e a proporção de seus domicílios que são ligados à rede geral de esgoto. O objetivo da regionalização é produzir um novo mapa onde as pequenas áreas do mapa inicial são agrupadas de acordo com seu grau de similaridade em relação a estas variáveis sociais e econômicas. As regiões formadas contém pequenas áreas que são bastante homogêneas com relação a todas as variáveis utilizadas. Ao mesmo tempo, as pequenas áreas pertencentes a regiões distintas serão bastante diferentes, em geral.

A situação ideal que uma regionalização almeja é aquela em que as regiões são compostas de áreas bastante similares, quase réplicas umas das outras. Ao mesmo tempo, as áreas de regiões distintas devem ser muito diferentes. Dessa forma, é obtida uma partição da população em regiões muito distintas compostas de unidades quase idênticas. Por causa disso, uma boa regionalização possibilita uma atuação uniforme dentro de cada região homogênea e, possivelmente, atuações diferentes em regiões homogêneas distintas. A ação específica a ser adotada deverá considerar as características específicas de cada área.

O mapa da Figura 2 apresenta a regionalização de São João do Meriti utilizando 10 variáveis sociais e econômicas obtidas do Censo Demográfico de 1991. O método utilizado foi desenvolvido pelo autor e será explicado na próxima seção. O método foi implementado num software chamado SKATER, um acrônimo para Spatial 'K'luster Analysis by Tree Edge Removal. Este software é gratuito e pode ser obtido no site http://www.est.ufmg.br/leste.

Notas de aula do curso de Estatística Espacial, 2003. E-mail:assuncao@est.ufmg.br

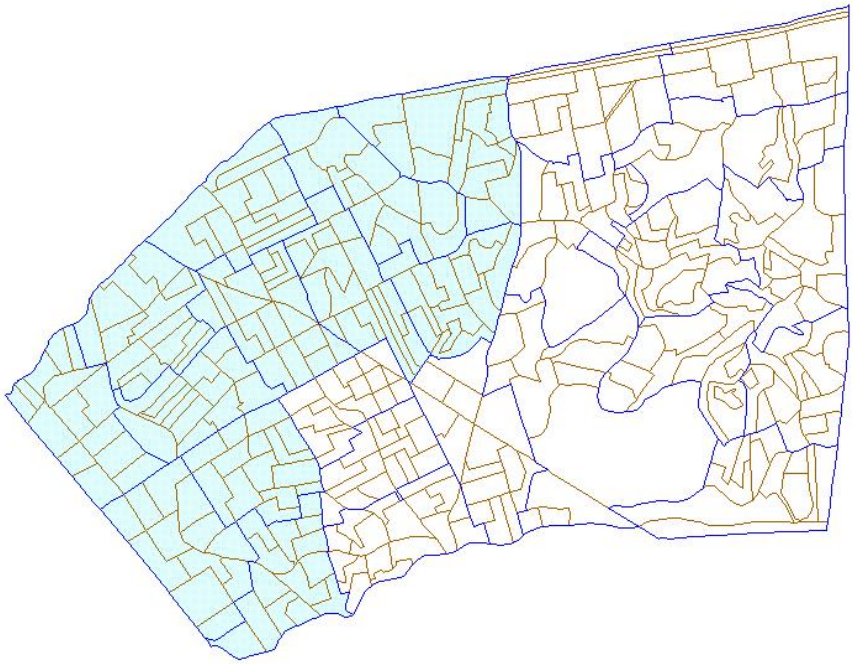

Figura 1: Mapa de São João do Meriti, RJ, dividido em 353 setores censitários de acordo com o Censo Demográfico de 1991.

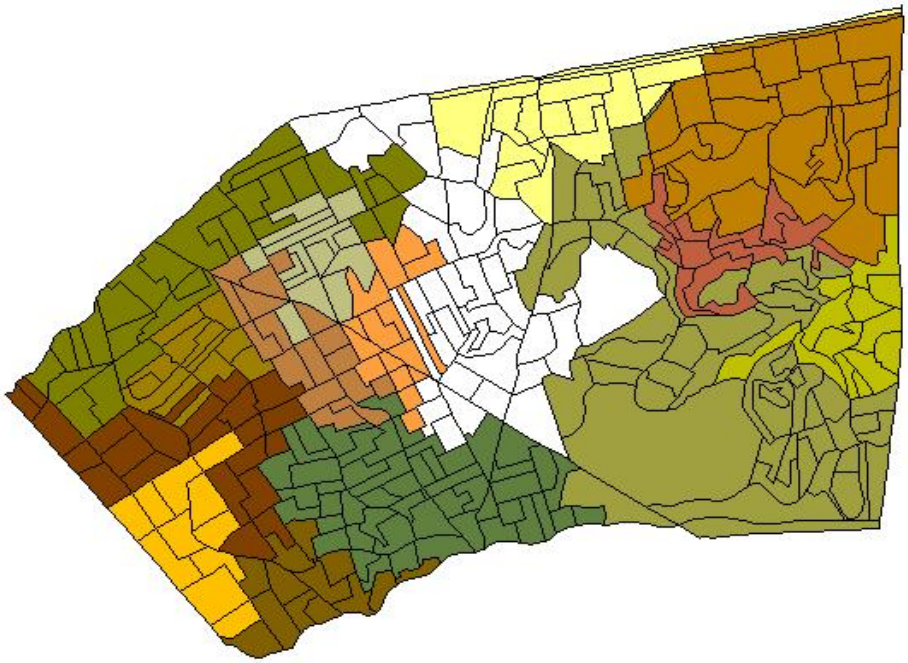

Figura 2: Mapa de São João do Meriti, RJ, dividido em regiões homogêneas de acordo com 10 variáveis sociais e econômicas de sua população coletadas pelo Censo Demográfico de 1991 do IBGE.

2 Metodologia

Vamos considerar áreas geográficas contíguas organizadas sob a forma de um mapa, tal como aquele mostrado na Figura 1. Associado a cada área, temos um vetor de atributos ou características $x = (x_1, \ldots, x_n)$ constituindo o perfil dessa área. Na análise de conglomerados espaciais, podemos ter duas medidas de distância entre quaisquer dois pares de áreas: uma baseada nas suas posições no espaço geográfico, e outra baseada numa medida de distância baseada no espaço das variáveis.

A medida de distância geográfica pode ser simplesmente a distância euclidiana entre os pontos que definem seus centróides no mapa num sistema de coordenadas qualquer ou então uma variável indicadora de que as áreas partilham uma fronteira comum. A partir desta medida de distância, uma relação de vizinhança pode ser criada. Por exemplo, duas áreas são vizinhas se a distância entre elas é menor que um certo valor ou então elas são vizinhas se elas compartilham uma fronteira comum.

A medida de distância no espaço das variáveis também é chamada de dissimilaridade entre as áreas. É comum tomar a distância euclidiana entre o vetor perfil de duas áreas como a medida de dissimilaridade entre elas ??fórmula ??. Outra possibilidade é tomar a distância de Mahalanobis ?? fórmula ??. É importante que as variáveis estejam padronizadas de alguma forma antes de calcular essa distância pois, caso contrário, as variáveis com maior variância vão tender a dominar o valor da dissimilaridade. Vários aspectos práticos sobre a escolha das variáveis,sobre as suas escalas ou padronizações, sobre as medidas de dissimilaridade e outros aspectos da análise de conglomerados espaciais serão discutidas na Seção ???.

2.1 Conglomerados espaciais

Para estudar métodos de análise espacial vamos introduzir algumas definições.

- **Definição:** Um conglomerado é qualquer subconjunto de áreas.

Conglomerados serão interessantes apenas se eles forem internamente homogêneos. Entretanto, esta característica não faz parte da definição de conglomerado, ela é apenas uma propriedade desejável a ser perseguida. Conglomerados podem ser constituídos por uma única área.

- **Definição:** Um conglomerado de áreas é conectado quando cada área componente é vizinha de, pelo menos, uma outra área pertencente ao conglomerado.

Assim, num conglomerado conectado, qualquer área possui algum vizinho no mesmo conglomerado. Vamos estender a definição acima permitindo que conglomerados formados por apenas um único objeto também sejam considerados conectados.

- **Definição:** Uma região subdividida em pequenas áreas é particionada em conglomerados espaciais quando as pequenas áreas forem agrupadas em conglomerados que sejam disjuntos e conectados.

O método usual de regionalização baseado na análise de conglomerados não garante que a alocação final será formada por conglomerados espaciais, a não ser através da ação subjetiva e manual do usuário. Portanto, a regionalização através do método usual de análise de conglomerado ainda sofre de uma boa dose de subjetividade.

2.2 Método da árvore geradora mínima

Nessa abordagem, reduzimos o mapa das áreas a um grafo onde cada nó v representa uma área e áreas vizinhas são ligadas por uma aresta quando as áreas são vizinhas. Vimos como fazer isso no Capítulo ?? desse livro. Vamos adotar a relação de adjacência, ou existência de fronteira comum, para definir quando duas áreas são vizinhas. Note que, dentre todas aquelas definições de vizinhança apresentadas no Capítulo ??, apenas esta de contigüidade ou compartilhamento de fronteiras pode ser adotada no procedimento de análise de aglomerados espaciais discutido neste capítulo.

Um caminho de v_1 a v_k é uma seqüência de nós $v_1, v_2, \ldots, v_k$ que são conectados pelas arestas $(v_1, v_2), (v_2, v_3), \ldots, (v_{k-1}, v_k)$. Um grafo é dito conexo se, para ir de um nó v do grafo a qualquer outro nó w, existe pelo menos um caminho de v a w. Usando as definições de vizinhança usuais, se o mapa não apresentar ilhas, vamos sempre encontrar o grafo correspondente ao mapa como um grafo conexo.

A cada aresta, vamos associar um custo, ou peso, relacionado ao grau de dissimilaridade entre as áreas. A escolha usual para esse custo é uma medida de distância entre os vetores de atributos, ou perfis, das duas áreas conectadas pela aresta. Geralmente, vamos adotar a distância euclidiana ou a distância de Mahalanobis (fórmulas ???) entre os vetores de atributos para representar essa distancia entre as áreas.

Nossa abordagem reduz o mapa de forma hierárquica. Primeiramente, defina quem é vizinho de quem e conecte esses vizinhos por uma aresta. Nesta etapa, geralmente estaremos olhando apenas a localização geográfica das áreas, independentemente de seus atributos. Isto é, apenas a distância geográfica entra em consideração. A seguir, atribua um custo a cada uma das arestas de acordo com a dissimilaridade de seus atributos, medida por uma distância entre os vetores de perfis das áreas. Nessa etapa, condicionada ao resultado da primeira, a localização geográfica é ignorada e levamos em conta apenas as variáveis sociais e econômicas componentes do perfil das áreas.

A idéia básica, nesse ponto, é simplificar o grafo apagando arestas de forma a ficar com um grafo reduzido, mas ainda conexo. Isto é, deverá ainda ser possível sair de uma área e chegar a qualquer outra área do mapa percorrendo sucessivamente arestas do grafo. Mas queremos apagar principalmente aquelas arestas de custo mais elevado de modo que, ao saltar de uma área a outra, teremos uma diferença pequena nos atributos das áreas envolvidas no salto. Além disso, queremos terminar com uma árvore tal que, se apagarmos qualquer aresta adicional, o grafo ficará dividido em dois subgrafos desconectados, os quais serão os candidatos a constituírem dois conglomerados espaciais. Isto é possível através da construção de uma árvore geradora mínima do grafo, como explicamos a seguir.

Vamos introduzir alguns dos conceitos básicos necessários nesse trabalho. Um *circuito* num grafo é um caminho onde os nós inicial e final são os mesmos. Uma *árvore* é um grafo conexo que não contém circuitos. Uma *árvore geradora* para um grafo G é um subgrafo que é uma árvore e que contém todos os nós de G. Assim, em uma árvore, quaisquer dois nós são unidos por um único caminho. Além disso, o número de arestas é igual a 1 mais o número de nós. Isso implica que, se qualquer aresta é apagada, a árvore estará desmembrada em duas subárvores desconectadas. O *custo de um grafo* é a soma dos custos das arestas do grafo. Uma *árvore geradora mínima* é uma árvore geradora que possui custo mínimo.

A figura 2 abaixo mostra à esquerda um grafo conexo com 10 nós localizados nos centróides das áreas de um mapa. A existência de uma aresta indica que as áreas são vizinhas e a espessura da linha é proporcional ao custo daquela aresta. Isto é, a espessura da aresta é proporcional à dissimilaridade das duas áreas medida como a distância entre os vetores de seus atributos. No gráfico do lado direito, encontra-se a árvore geradora mínima. Note que, se apagarmos qualquer aresta na árvore da direita, teremos dois subgrafos desconectados.

Note que um grafo pode ter mais que uma árvore geradora mínima, especialmente se os custos assumirem valores apenas num pequeno conjunto de valores possíveis. Dificilmente, esse será o caso nas situações onde as variáveis ou atributos das áreas são variáveis contínuas. O motivo é que o custo de uma aresta será uma medida de distância entre dois vetores de atributos, um para cada uma das duas áreas componentes da aresta. Como essa distância será, geralmente, a distância euclidiana ou a distância de Mahalanobis, dificilmente teremos duas arestas com custos idênticos.

2.3 Algoritmo de Prim para a árvore geradora mínima

O algoritmo que nós utilizamos para construir uma árvore geradora mínima é aquele devido a Prim (1957). A partir de um grafo conexo com custos associadas às arestas, construímos a árvore de forma recursiva, começando com a árvore T_1 e aumentado-a progressivamente $T_2, T_3, \ldots$ até T_n, que é a árvore geradora mínima.

- **Passo 1:** Tome qualquer nó v e faça $T_1 = v$.

Repita o passo 2 tanto quanto possível:

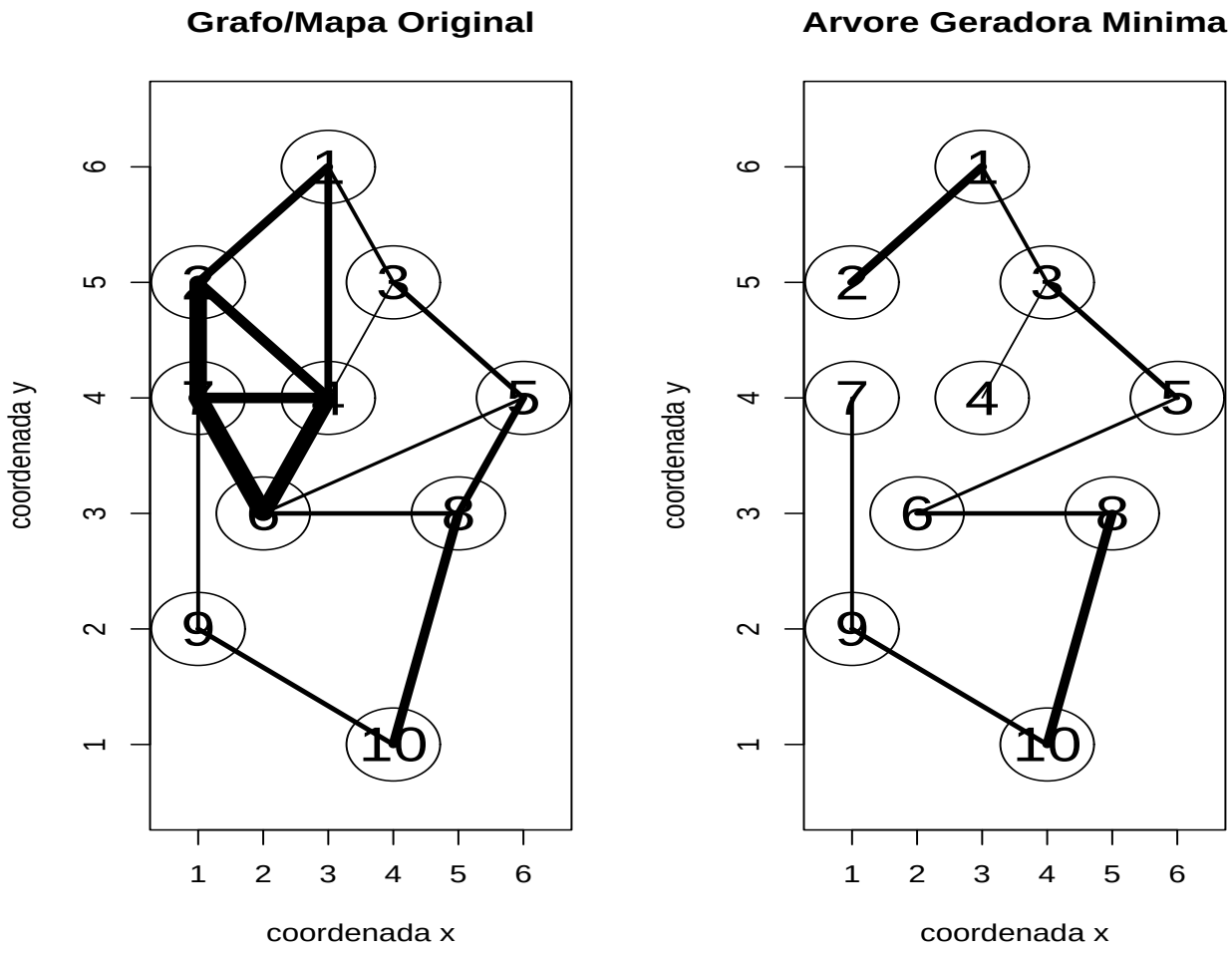

Figura 3: Grafo de mapa com 10 áreas com arestas proporcionais ao seu custo e sua árvore geradora mínima.

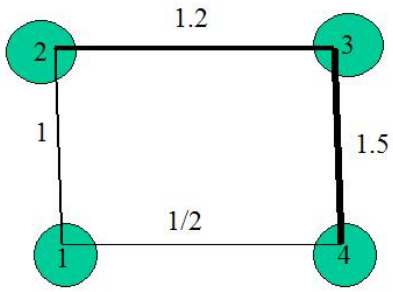

Figura 4: Grafo mostrando que a árvore geradora mínima possui propriedade ótima global e não local.

- **Passo 2:** Encontre uma aresta e_k de custo mínimo unindo um nó em T_k a um nó que não está em T_k. Se existir mais de um nó com essa propriedade, escolha um deles arbitrariamente. A árvore T_{k+1} é o grafo obtido adicionando esse nó e aresta a T_k.

Como o número de nós é finito, o algoritmo é interrompido em algum momento.

Numa árvore geradora mínima, nenhuma aresta da árvore geradora mínima pode ser substituída sob pena de gerar uma árvore de custo mais elevado. Apesar disso, não é verdade que o caminho entre quaisquer duas áreas na árvore geradora mínima seja aquele de custo mínimo entre essas duas área no grafo original. Para verificar isso, considere o exemplo abaixo exibido na Figura 4. Iniciando a árvore com o nó 1, sucessivamente adicionamos o nó 4 com a aresta 14, o nó 2 com a aresta 12 e o nó 3 com a aresta 23. No entanto, note que o caminho 341, que não pertence à árvore geradora mínima, tem custo 2 enquanto que o caminho 321, pertencente à árvore geradora mínima, tem custo 2.2, maior que o anterior. Isto mostra que o custo mínimo da árvore geradora mínima é uma propriedade global e não uma propriedade local.

2.4 Implementação do algoritmo de Prim

Nós adotamos uma implementação eficiente do algoritmo de Prim sugerida por Manber (1989) e descrita no quadro abaixo.

Algoritmo Árvore Geradora Mínima

Input: G, um grafo com custo nas arestas

Output: T, árvore geradora de Gcom custo mínimo

begin

Inicialmente T é o conjunto vazio $\emptyset$;

for todos nós v **do**

$v.mark \leftarrow$ False (é True se $v \in T$)

$v.cost \leftarrow \infty$

seja (x, y) uma aresta qualquer de G

$x.mark \leftarrow$ True

for todas arestas (x, z) **do**

$z.aresta \leftarrow (x, z)$ (uma aresta de custo mínimo de T a z)

$z.cost \leftarrow$ custo(x, z) (o custo de $z.aresta$)

while existir nó não marcado **do**

seja $w \in$ nós não marcado tal que $w.custo$ é mínimo

if $w.custo = \infty$ **then**

stop("G não é conexo")

else

$w.mark \leftarrow$True

adicione $w.aresta$ a T

(atualize agora os custos de nós não marcados conectados a w)

for todas arestas (w, z) **do**

if not $z.mark$ **then**

if $custo(w, z) < z.custo$ **then**

$z.aresta \leftarrow (w, z)$

$z.custo \leftarrow custo(w, z)$

end

2.5 Conglomerados espaciais a partir da árvore geradora mínima

Após a criação da árvore geradora mínima, passamos a particioná-la para obter os conglomerados espaciais. O problema combinatório de formação dos conglomerados espaciais está agora tremendamente reduzido pois basta verificar as n arestas da árvore e apagar uma delas para ter a árvore separada em dois subgrafos desconectados. Iterando essa função que apaga arestas em cada subgrafo resultante, vamos criando conglomerados de forma hierárquica.

Dada uma árvore geradora, uma escolha natural da aresta a ser apagada é aquela que possui o maior custo ou dissimilaridade. Ao apagar esta aresta numa árvore geradora, teremos como resultado dois subgrafos desconectados, cada um deles conexo, que podem ser vistos como dois conglomerados espaciais. O custo desse novo grafo seccionado em dois é a soma dos custos das arestas não apagadas nos dois subgrafos. É claro que, se qualquer outra aresta fosse apagada na árvore geradora inicial, o resultado seria um grafo seccionado em dois com um custo maior (ou igual). Nesse sentido, é natural escolhermos para apagar a aresta de custo máximo.

Continua-se seccionando a árvore apagando sucessivamente a aresta com o segundo menor custo (criando três subgrafos desconectados), apagando a aresta com o terceiro menor custo (criando quatro subgrafos), etc.

Considerando novamente o exemplo da Figura 2, mostramos na Figura 3 abaixo, os subgrafos sucessivos, e respectivos conglomerados espaciais, que resultam ao apagarmos as arestas de maior custo.

???????

2.5.1 Outro critério para particionar a árvore geradora mínima

A escolha da aresta a ser apagada usando a mesma medida de dissimilaridade usada para construir a árvore geradora mínima tem uma grande desvantagem, no entanto. Ocorre que as últimas arestas a serem adicionadas na árvore geradora mínima tendem a ter os maiores custos. Afinal, não é sem motivos que essas áreas foram as últimas a serem conectadas. Ao apagar as arestas de maior custo na árvore geradora

mínima, estaremos tendendo a quebrar o grafo nas últimas arestas que foram adicionadas à árvore as quais, por sua vez, tendem a ligar áreas isoladas no grafo. Isto é, apagar as últimas arestas tende a gerar dois conglomerados, um formado por apenas umas poucas áreas e o outro, com o restante das áreas. Esse resultado geral não ocorre no caso do exemplo simples com apenas 10 áreas mas ele poderá ser visualizado de forma muito clara na aplicação com os dados de São João do Meriti na seção 6.

Assim, buscamos uma alternativa para a escolha das arestas a serem sucessivamente apagadas do grafo da árvore geradora mínima. Decidimos associar um custo diferente à cada aresta para esta segunda etapa de apagar arestas. Numa árvore, denotamos por SSTO a soma de quadrados dos desvios no espaço das variáveis em relação à média de todas as áreas da árvore. Para cada um dos dois conglomerados resultantes de apagar-se uma aresta numa árvore, calculamos a soma de quadrados dos desvios no espaço das variáveis em relação à média do conglomerado resultante. A seguir, somamos as duas somas de quadrados, uma de cada conglomerado resultante, obtendo SSA. Quanto menor SSA, mais homogêneos serão os conglomerados resultantes. Como SSA está entre 0 e SSTO, definimos o custo de apagar a aresta como sendo SSTO-SSA. Assim, um custo alto associado com a aresta indica que seu desaparecimento vai gerar conglomerados homogêneos.

Este novo critério de apagar as arestas que maximizam a redução da soma de quadrados gera resultados muito bons, como veremos no caso de São João do Meriti. No futuro, vamos buscar alguma propriedade de otimalidade que porventura venha a satisfazer a árvore geradora mínima com arestas apagadas pela redução de mínimos quadrados.

A Figura 4 abaixo mostra o resultado de adotarmos esse novo critério para particionar a árvore geradora mínima do exemplo simples com 10 áreas. Note que, embora a primeira partição seja idêntica à anterior, as demais diferem na escolha da aresta a ser apagada.

???????

3 Aspectos práticos de análise de conglomerados espaciais

3.1 Padronização das variáveis

Alguma forma de padronização das variáveis é necessária devido ao impacto que diferentes escalas podem ter na função de dissimilaridade e na soma de quadrados dentro dos conglomerados. As escolhas usuais para a função distância levam ao fato de que a dissimilaridade entre duas áreas quaisquer é influenciada pelas escalas arbitrárias em que as variáveis são medidas. Por exemplo, uma variável carrega a mesma informação, seja medida como proporção, seja medida como porcentagem. No entanto, a diferença dos valores desta variável hipotética entre duas áreas é 100 vezes maior no caso de porcentagem do que no caso de proporção. Assim, sugerimos que todas as variáveis possuam uma escala padronizada com média zero e o desvio padrão igual a 1. Observe que, implicitamente, este procedimento de padronização termina dando pesos idênticos às variáveis. No entanto, é possível dar pesos diferentes às variáveis, caso o usuário queira. Basta definir uma função distância apropriada que incorpore esta ponderação diferenciada.

3.2 Escolha das variáveis

É comum que características sociais e econômicas medidas em pequenas áreas (no caso, os municípios de Minas Gerais) apresentem uma alta correlação estatística já que refletem aspectos associados da mesma estrutura social. Isto implica numa grande redundância de informação, com muitas variáveis carregando basicamente a mesma informação que outras.

Por exemplo, a porcentagem do Fundo de Participação Municipal sobre a Receita Corrente Municipal (FPM) é negativa e fortemente correlacionada (r = -0.79) com a porcentagem de arrecadação do ICMS sobre a Receita Corrente Municipal (ICMS). O gráfico da Figura 1 mostra um diagrama de dispersão destas variáveis onde cada ponto representa um município. Um valor muito alto de FPM está, em geral, associado com um valor muito baixo de ICMS. De fato, a redundância da informação é de tal ordem que, a partir da posição do FPM de um dado município, seria possível predizer o seu valor de ICMS com uma razoável precisão. No entanto, observe no gráfico a presença de dois municípios com valores de ambas as variáveis, tanto FPM quanto ICMS, muito baixos, contrariando a tendência geral. Além disto, note qe a

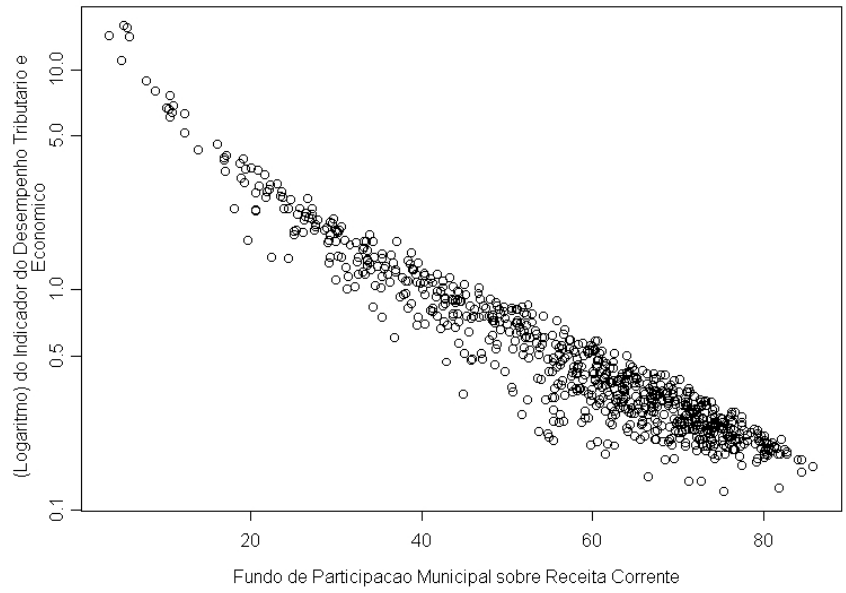

Figura 5: iagrama de dispersão da variável (logaritmo do) Indicador de Desempenho Tributário e Econômico versus o Fundo de participação dos municípios sobre a receita corrente. Cada ponto representa um município de Minas Gerais.

relação entre FPM e ICMS não é linear.

Figura ???

Outro par de variáveis medidas nos municípios mineiros exibindo um alto grau de correlação é o IDH-Renda e o (logaritmo do) PIB municipal per capita. Neste caso, a correlação é um pouco menor (r=0.59) mas ainda assim apresenta uma grande quantidade de informação redundante. A razão para considerarmos a transformação logarítmica será dada mais tarde. A capacidade de predição do IDH-Renda a partir do valor do PIB municipal é menor que antes mas ainda é bastante alta. Assim, existe um certo grau de redundância entre estas duas variáveis, menor que o anterior, mas ainda substancial. Esta redundância ou correlação entre as variáveis mostra que parte da informação contida numa das variáveis para discriminar os municípios também está presente na outra variável.

No entanto, nem todo par de variáveis exibe um grau de correlação tão alto quanto esses. Por exemplo, considere a Figura 3 que mostra o diagrama de dispersão das variáveis Crescimento percentual do PIB entre 85 e 96 e o valor do (logaritmo) do PIB em 1997. Neste caso, a correlação é bastante pequena (r=0.29). Isto é, o valor de uma variável informa muito pouco sobre o valor da outra: o fato de uma variável ter um valor muito alto diz muito pouco sobre o valor provável que a outra variável pode ter. Isto implica que não há muita redundância de informação nestas variáveis, elas representam partes independentes de informação para discriminar os municípios de Minas Gerais.

Podemos ter um par de variáveis possuindo uma redundância tão grande que a informação contida numa delas é praticamente uma transformação matemática exata da informação contida na outra. Isto é, não existe de fato quase nenhuma informação nova na outra variável, que seria apenas a variável antiga numa nova escala, algo semelhante a medir a temperatura em graus Celsius ou em graus Farenheit. Este é o caso de dois pares de variáveis, aquele formado pelas variáveis Fundo de Participação Municipal sobre Receita Corrente (FPM) e o (logaritmo do) Indicador do Desempenho Tributário e Econômico e aquele formado pelas variáveis Renda Familiar per Capita em 1991 e o IDH Renda. No caso do primeiro par, mostrado na Figura 4, a correlação é de -0.96 e, no caso do segundo par, mostrado na Figura 5, a correlação é de 0.98. Observe que no caso do segundo par, a relação ????

SO WHAT ?? ENOUGH ABOUT (OF??) REDUNDANCY...

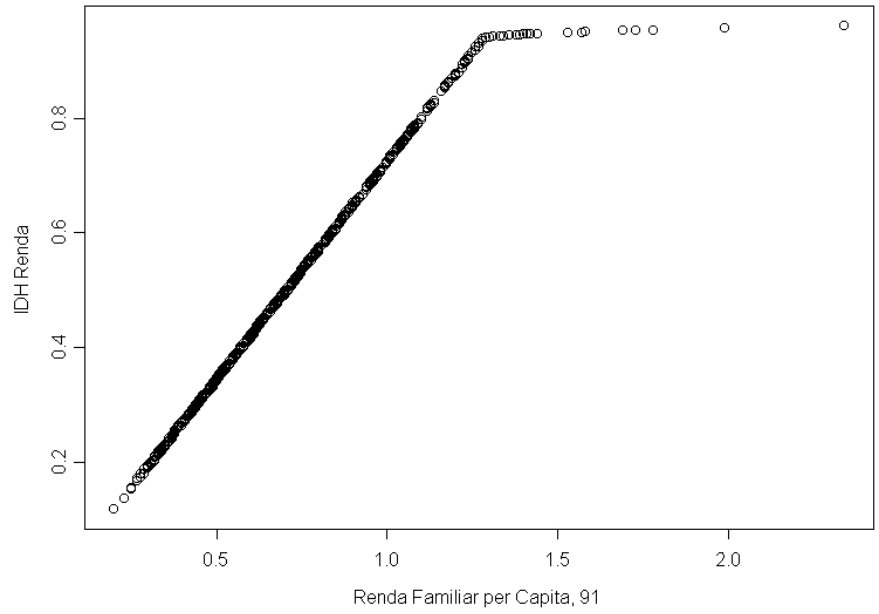

Figura 6: Diagrama de dispersão da variável IDH-Renda versus a Renda Familiar per Capita. Cada ponto representa um município de Minas Gerais.

3.3 Peso das variáveis

Na função de distância, devemos ponderar diferentemente as variáveis apenas se houver razões teóricas ou de outra ordem. Geralmente, não teremos estas razões teóricas e qualquer ponderação seria feita com base em resultados empíricos que podem variar de região para região. Além disto, a redundância da informação contida em várias variáveis servirá como uma ponderação implícita. Fazendo uma analogia, podemos imaginar um espaço de todas as possíveis características/variáveis de interesse e uma distribuição de probabilidade associada a este espaço dando igual chance a cada uma dessas características. Um subconjunto de características com alto grau de correlação mútuo deve acabar representado mais freqüentemente numa "amostra" de variáveis.

Associado ao peso, existe a questão das variáveis de perfil (idade, renda, etc.) que devem ser resumidas numa única variável a menos que tenhamos interesse em dois ou mais aspectos diferentes da distribuição de forma independente (proporção de idosos e de crianças, por exemplo). A razão é que usar várias variáveis para traçar um perfil de uma única característica vai terminar dando mais peso a esta característica que a outras que entrarem com uma única variável.

Como são muitas as características a serem analisadas, deseja-se que a homogeneidade seja obtida em várias variáveis simultaneamente. No entanto, as variáveis que devem ser utilizadas para realizar a agregação espacial para formação das áreas foram...BLA BLA BLA ?????

Correlação Espacial não é o mesmo que a correlação usual....Mostrar mapas.????

(ver GAT???COMENT para escolher quais). Idéia: correlação espacial é diferente e info espacial é diferente. ????

4 Dados Geográficos Sociais e Econômicos

O IBGE divide os municípios brasileiros em pequenas áreas com o objetivo de organizar o trabalho de coleta de dados dos Censos Demográficos decenais. Estas áreas utilizam referenciais geográficos claramente definidos e cobrem todo o território nacional, área urbana e rural. Nas áreas urbanas, tipicamente,

os setores possuem em média 300 domicílios e variando entre 200 e 500 domicílios com uma população variando entre 700 e 2000 habitantes. Estas são as menores unidades geográficas brasileiras para as quais existem informações populacionais confiáveis e cobrindo todo o território nacional. Como as informações são coletadas na época dos Censos, a informação mais recente refere-se ao ano de 1991. Houve uma Contagem Populacional efetuada em 1996 mas ela basicamente apenas contou a população por faixas de idade e sexo, não coletando outras informações sobre a estrutura dos domicílios ou da população.

Em Belo Horizonte, não existe outra base de dados, além das informações censitárias de 1991 por setor censitário, tão detalhada geograficamente, tão confiável e cobrindo todo o município. Por este motivo, utilizamos esta base para realizar a análise deste relatório.

Em 1991, Belo Horizonte foi dividida em 1999 setores censitários dos quais 11 não possuíam domicílios (eram áreas militares, asilos, etc.) e foram eliminados da regionalização. O mapa de Belo Horizonte particionado nestes setores censitários encontra-se na Figura 3. O número médio de domicílios por setor é 255 e 90% dos setores possuem entre 128 e 383 domicílios. Por causa de seu pequeno número de domicílios, de sua pequena extensão geográfica e dos critérios para demarcação dos seus limites, cada setor é composto de domicílios e população com um perfil social e econômico muito similar. Desta forma, nós ignoramos as diferenças internas a um setor censitário e trabalhamos com os valores médios de características sociais calculadas em cada setor. Devido a sua similaridade interna extremamente alta, estes valores médios são bem representativos das características da população residente em cada setor.

As variáveis sociais e econômicas utilizadas neste capítulo referem-se a características de infra-estrutura dos domicílios e da população residente em cada setor censitário. Elas são as seguintes:

- Proporção de domicílios no setor que são casas (em contraste com apartamentos).
- Proporção de domicílios no setor que são próprios (em contraste com alugados, cedidos, etc.)
- Número médio de cômodos por domicílios do setor.
- Número médio de cômodos que são utilizados como dormitórios nos domicílios do setor.
- Número médio de banheiros por domicílios do setor.
- Número médio de pessoas por domicílios do setor.
- Proporção de domicílios do setor que estão ligados à rede geral de água.
- Proporção de domicílios do setor que estão ligados à rede geral de esgoto.
- Proporção de domicílios do setor que têm seu lixo coletado pela SLU.
- Proporção dos residentes no setor que são do sexo masculino.
- Proporção dos chefes de domicílios do setor que são do sexo masculino.
- Proporção dos residentes do setor que são crianças (abaixo de 10 anos de idade).
- Proporção dos residentes do setor que são idosos (com idade igual ou superior a 60 anos).
- Renda média dos chefes de domicílios do setor.

Estas 14 variáveis fornecem um quadro bastante detalhado e rico das características sociais e econômicas da população de cada setor. Todas possuem um alto grau de correlação espacial fazendo com que mapas dessas variáveis mostrem áreas vizinhas com valores geralmente similares como será mostrado mais a frente.

Várias das 14 variáveis listadas anteriormente possuem alta correlação já que refletem aspectos associados da estrutura social. Por exemplo, a renda média dos chefes dos setores está associada com o número médio de cômodos. O gráfico da Figura 4 mostra um diagrama de dispersão destas variáveis onde cada ponto representa um setor censitário. É claro que uma alta renda média está fortemente associada com uma grande número de cômodos por domicílio.

Nem todo par de variáveis exibe o mesmo alto grau de correlação daquele entre renda e número de cômodos. Considere, por exemplo, a renda média dos chefes do setor e a proporção de idosos no setor

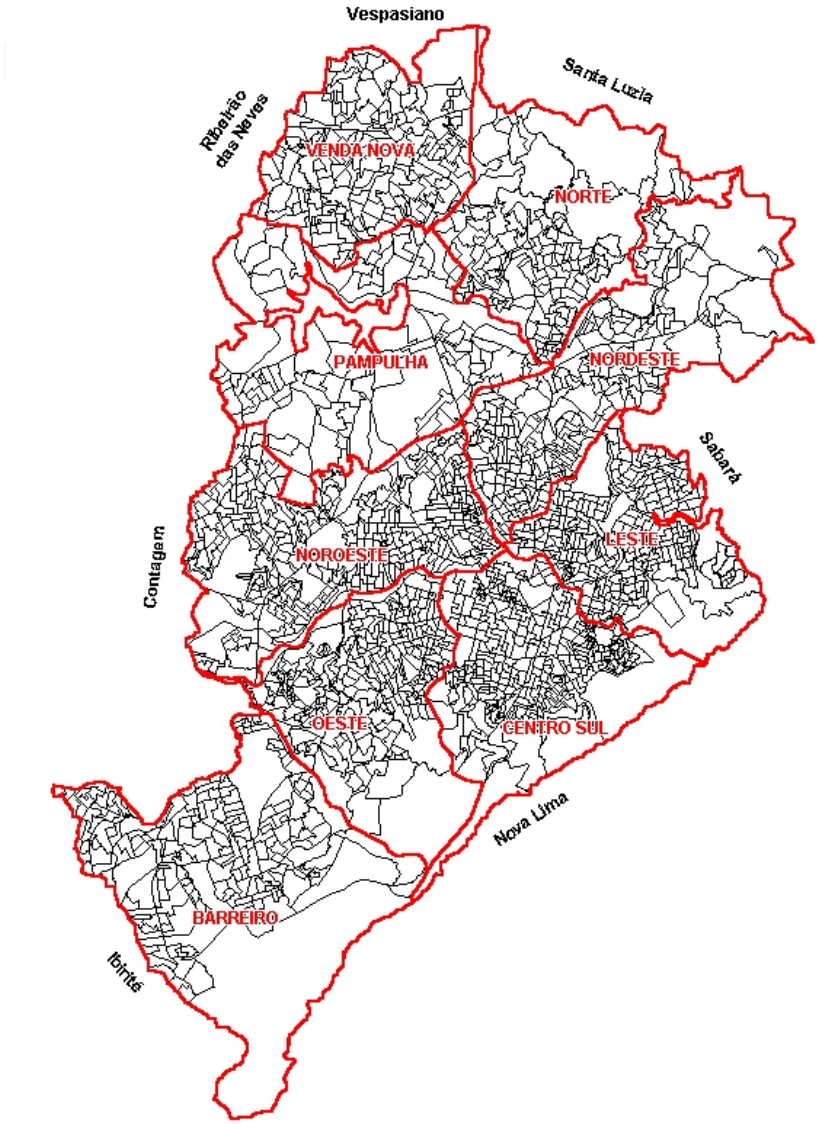

Figura 7: Mapa de Belo Horizonte dividido em seus 1999 setores censitários de acordo com o Censos Demográfico de 1991.

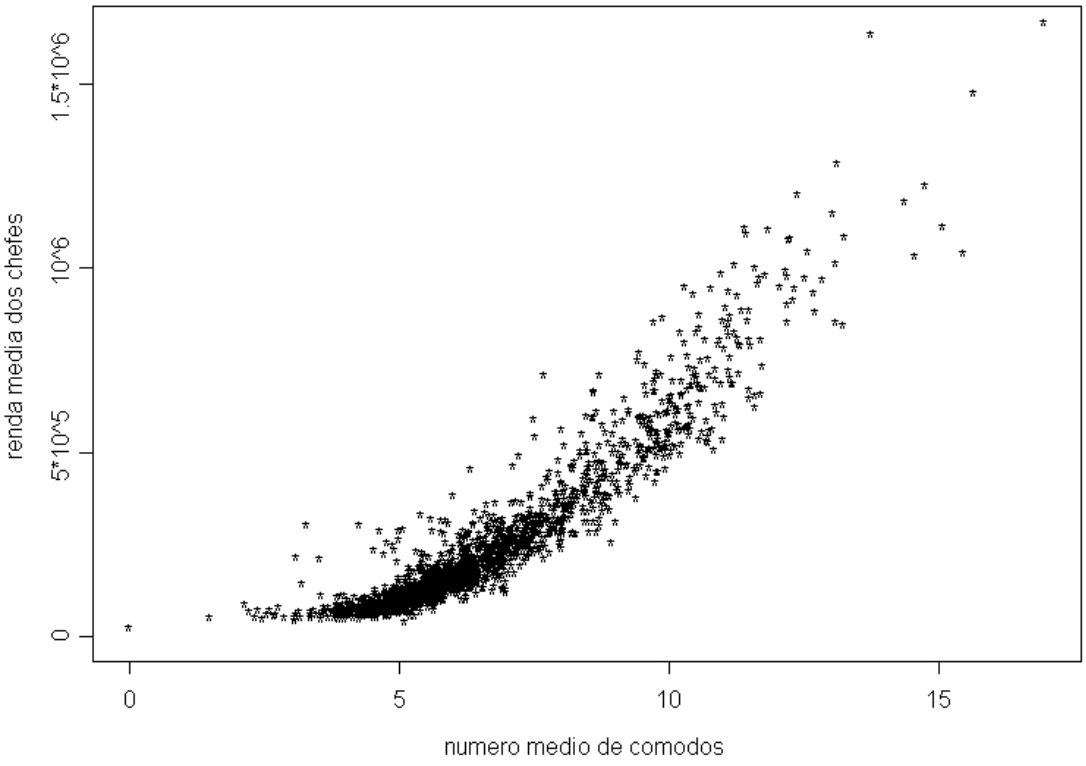

Figura 8: Diagrama de dispersão do número médio de cômodos (eixo horizontal) e da renda média do chefes (eixo vertical). A unidade é o setor censitário de Belo Horizonte, cada um deles representado por um ponto no gráfico.

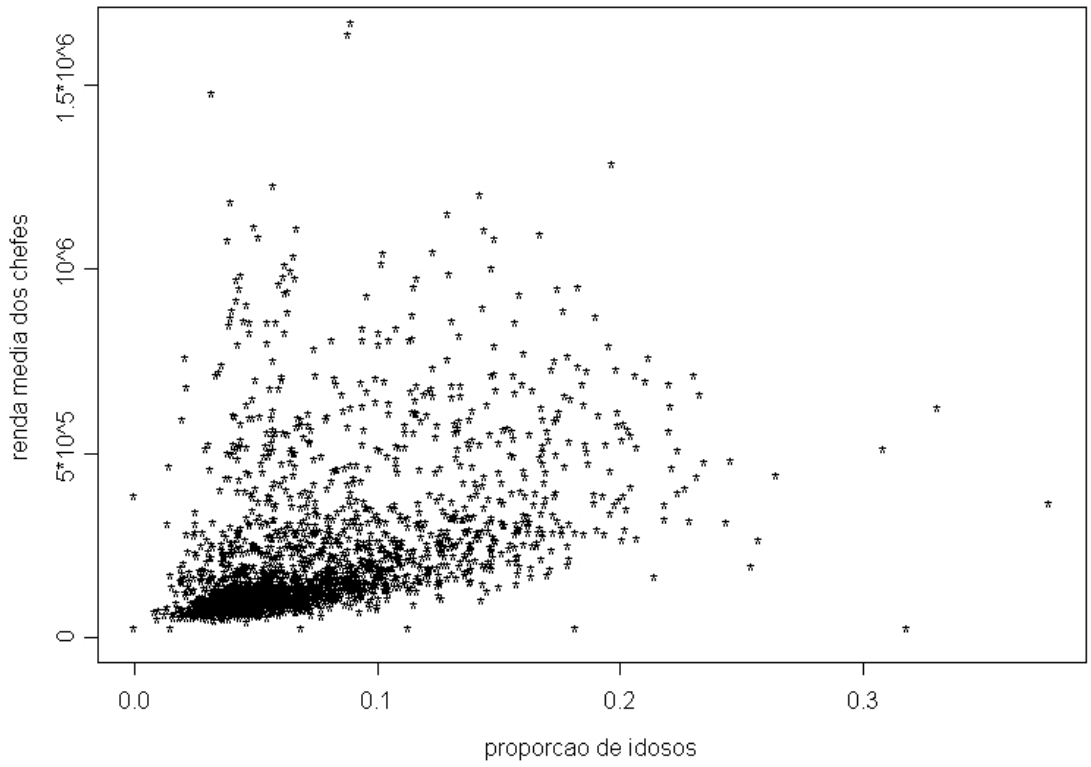

Figura 9: Diagrama de dispersão da proporção de idosos (eixo horizontal) e da renda média do chefes (eixo vertical). A unidade é o setor censitário de Belo Horizonte, cada um deles representado por um ponto no gráfico.

cujo diagrama de dispersão encontra-se na Figura 5. A conseqüência é que, embora exista um certo grau de redundância na informação contida nesse conjunto de variáveis, elas fornecem informações diferentes que, devidamente combinadas, compõem um perfil multifacetado dos setores censitários. Assim, a agregação de setores com base na similaridade de seu perfil deve levar em conta todas as variáveis simultaneamente.

5 Mapas das variáveis

Apresentamos a seguir mapas de Belo Horizonte apresentando como tema algumas das variáveis mais significativas da lista anterior. A maior parte dos mapas das 14 variáveis utilizadas neste trabalho mostra um padrão espacial semelhante. Este padrão reflete os diferentes estratos de status social distribuídos no espaço.

O mapa da renda média do chefe de família mostra os valores mais altos na região centro sul da cidade e em volta da Lagoa da Pampulha. Pequenos enclaves de renda mais alta são encontrados no Coração Eucarístico (Regional Noroeste) e na Cidade Nova. Em tornos das regiões de renda mais alta, encontram-se áreas de renda média e, finalmente, no anel mais externo a estas últimas, encontram-se as regiões de renda mais baixa. As maiores regiões de renda mais baixa são as regiões Norte, Nordeste, Noroeste, Oeste e do Barreiro.

MAPA DE RENDA ???

O mapa da proporção de idosos na Figura 10 reflete uma distribuição espacial diferente. Os valores apresentam um gradiente de aumento do proporção de idosos à medida em que aproxima-se do centro da

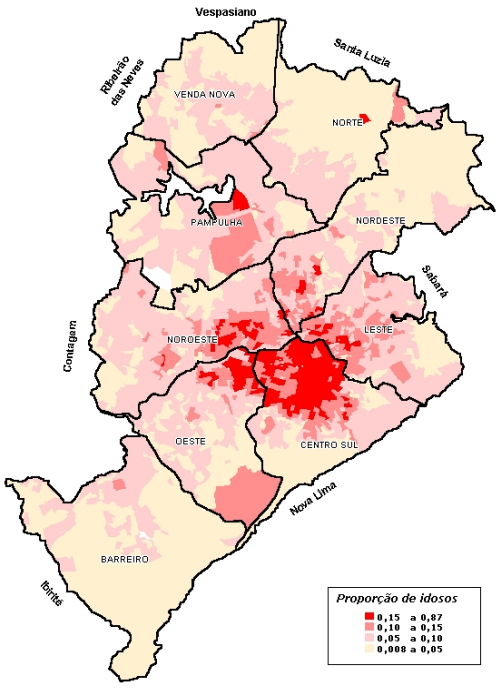

Figura 10: Mapa temático da proporção de idosos (acima de 60 anos) por setor censitário em Belo Horizonte em 1991.

cidade, onde encontram-se o valores mais elevados, acima de 15% de idosos. Na periferia da cidade os valores são inferiores a 10%, sendo muitas vezes inferior a 5%, um valor pelo menos três vezes menor que no centro de Belo Horizonte.

A maior presença de idosos nas áreas centrais da cidade é resultado de dois processos distintos. O primeiro processo é aquele de ocupação do espaço geográfico. O segundo processo é aquele de transição demográfica. Quanto ao primeiro processo, as áreas centrais são aquelas de assentamento mais antigo enquanto que as da periferia são de ocupação mais recente. Uma grande parcela desta ocupação mais recente é feita por migrantes que chegaram há relativamente pouco tempo em Belo Horizonte. Como a maior parte desses migrantes é constituída por pessoas jovens, isto explica a baixa presença relativa de idosos nestas áreas.

O segundo processo, o da transição demográfica, é constituído pelo declínio rápido das taxas de mortalidade seguido no tempo pelo declínio das taxas de fecundidade. Os demógrafos estudam este processo com atenção e têm encontrado vários fatores explicativos para ele. Populações de renda maior e com mais escolaridade iniciam o processo da transição demográfica mais cedo e com taxa mais rápida. Assim, a queda da fecundidade das mulheres de renda mais alta e com maior escolaridade, leva a uma mudança da pirâmide etária aumentando a proporção de idosos e diminuindo a presença relativa de crianças.

De modo inteiramente análogo, as duas explicações acima, em termos dos processos de ocupação do espaço e da transição demográfica, servem para explicar a distribuição espacial da proporção de crianças na cidade. Elas podem também ser usados para explicar o padrão espacial da proporção de residentes do sexo masculino. Por um lado, como é mais comum que migrantes recentes sejam homens, a periferia apresenta maior proporção de homens em sua população. Por outro lado, como a mortalidade em qualquer faixa etária atinge mais aos homens que as mulheres, populações mais idosas tendem a ser mais femininas.

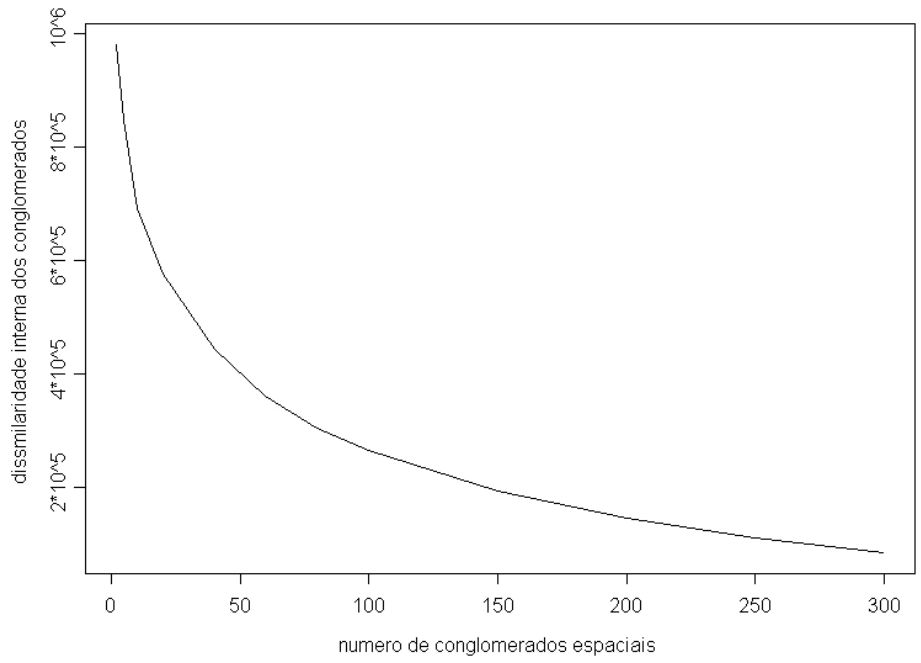

Figura 11: Medida de heterogeneidade das regiões formadas a medida em que o número de regiões aumenta. Dados de 14 variáveis sociais e econômicas do Censo Demográfico de 1991 nos setores censitários de Belo Horizonte.

Como a periferia é uma população relativamente mais jovem que o centro, ela também vai apresentar mais homens na sua composição.

6 Regionalização de Belo Horizonte

A regionalização de Belo Horizonte foi feita utilizando-se o método descrito neste capítulo e as 14 variáveis do Censo Demográfico de 1991 avaliadas em cada um dos 1999 setores censitários de Belo Horizonte. Uma das principais dificuldades em qualquer procedimento de regionalização é a decisão sobre quantas regiões devem ser criadas. No nosso método, esta decisão repousa sobre uma análise do padrão de diminuição da heterogeneidade interna dos grupos formados à medida em que o número de grupos aumenta.

A Figura 11 mostra a dimunição da heterogeneidade dentro dos conglomerados formados a medida que aumenta o número de conglomerados. O gráfico exibe os resultados de várias regionalizações, começando com uma partição do município em apenas duas regiões, passando para uma nova regionalização baseada em cinco regiões e, assim por diante, até terminar com uma regionalização baseada em 300 regiões. O eixo horizontal mostra o número de regiões das sucessivas regionalizações. O eixo vertical mostra o grau de heterogeneidade social e econômica dos residentes das regiões componentes de uma dada regionalização.

A medida em que a partição de Belo Horizonte vai sendo sucessivamente refinada pela criação de mais e menores regiões, o resultado é a formação de grupamentos espaciais menos heterogêneos. A diminuição da heterogeneidade interna das regiões é rápida e abrupta nas primeiras partições. Isto é, passar de uma regionalização baseada em cindo regiões para outra baseada em 20 regiões faz a medida de heterogeneidade diminuir substancialmente. Entretanto, a taxa de diminuição da heterogeneidade vai decrescendo até o ponto em que refinar uma regionalização quase não contribui mais para criar regiões mais homogêneas. Analisando o gráfico, pode-se verificar que regionalizações que utilizam entre 70 a 100 regiões parecem atender a um ponto ótimo: diminuem substancialmente a heterogeneidade inicial presente no município e, ao mesmo tempo, aumentar a regionalização não vai contribuir significativamente para diminuir a

heterogeneidade. Experimentando com alguns valores no intervalo 70 a 100, analisamos as regionalizações resultantes concluindo por uma partição do município em 80 regiões. O mapa resultante encontra-se na Figura ?? e na folha cartográfica fornecida junto com este relatório.

MAPA DA DIVISAO ????

7 Tamanho dos conglomerados

A visualização da regionalização obtida mostra que os 80 conglomerados possuem extensões territoriais muito distintas. De fato, pode-se verificar que aproximadamente metade deles ocupam territórios relativamente extensos, enquanto a outra metade é constituída por conglomerados bastante localizados e ocupando uma porção relativamente pequena do espaço geográfico da cidade. Vários desses pequenos conglomerados representam enclaves de populações circundadas por regiões muito discrepantes.

Por exemplo, a região Centro-Sul da cidade possui um pequeno conglomerado abrigando a favela do Morro do Papagaio. Este bolsão de pobreza, geograficamente pequeno, está cercado das áreas mais afluentes do município que juntas, ocupam grande extensão territorial. Estas áreas afluentes não são homogêneas e, por isto, são divididas em grandes conglomerados que separam, por exemplo, a região do Belvedere e Santa Lúcia daquela formada pelo Santo Antônio, São Pedro e outras. Entretanto, a transição social entre estes conglomerados não é tão abrupta quanto aquela que separa o Morro do Papagaio das demais. Assim, o pequeno enclave representado pelo conglomerado do Morro do Papagaio tem limites que demarcam estratos sociais muito distintos e distantes na escala se status sócio-econômico.

Outros enclaves com características semelhantes podem ser observados no mapa. Por exemplo, note a separação da favela Vila Cafezal de suas áreas vizinhas. Ou então, considere o pequeno conglomerado formado pelo Conjunto Califórnia, que é bem diferente que as áreas vizinhas dos bairros João Pinheiro, Alto dos Pinheiros, etc. O Conjunto Califórnia é formado basicamente por conjuntos habitacionais que abrigam uma população muito distinta daquela residentes nesses bairros.

O Coração Eucarístico é outro enclave no município. Comparado com as áreas vizinhas, este bairro abriga uma população com renda superior, com maior presença de profissionais liberais e grande proporção de apartamentos, entre outras características.

8 Separação de regiões

A regionalização de Belo Horizonte separa regiões que não possuem transições abruptas como aquelas que caracterizam vários dos pequenos enclaves e suas áreas vizinhas. Por exemplo, os bairros Belvedere e Santa Lúcia formam uma região distinta (de número 76) daquela formada pelos bairros Santo Antônio, São Pedro, Cruzeiro, Serra e Santo Agostinho (de número 19). Estas regiões não se diferenciam muito em todas as 14 variáveis sociais e econômicas utilizadas no procedimento de regionalização. De fato, as principais responsáveis pela separação entre estas duas regiões são: proporção de chefes do sexo masculino, renda média do chefe e número médio de banheiros, cômodos e dormitórios. Em relação a estas variáveis, as duas regiões são relativamente distintas mas, em relação às demais, elas não se diferenciam muito. Isto pode ser constatado no gráfico da Figura 12.

Este gráfico mostra 14 pontos, cada um representando uma das variáveis utilizadas para regionalizar Belo Horizonte. Estas variáveis estão padronizadas para terem média zero e desvio padrão unitário, possuindo assim uma escala única e permitindo comparações entre variáveis. O eixo vertical mostra os valores médios referentes a região do Belvedere/Santa Lúcia e o eixo horizontal, àqueles da região do Santo Antônio e demais bairros. A linha diagonal é a reta y=x. Pontos que estão próximos da reta significam que as variáveis correspondentes são similares nas duas regionalizações. Já os pontos que se afastam da reta indicam que as duas regiões são muito distintas nas variáveis correspondentes a esses pontos. Desta forma, observa-se que o que distingue estas duas regiões são as variáveis listadas anteriormente.

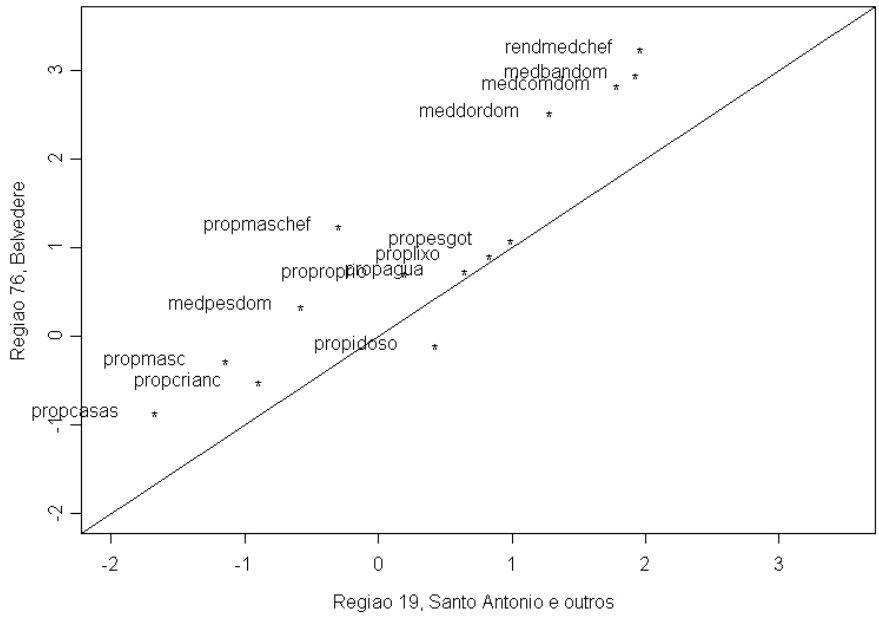

Figura 12: Perfis dos conglomerados 19 e 76.