AGRUPAMENTO DE REGIÕES:
UMA ABORDAGEM UTILIZANDO ACESSIBILIDADE

Carlos Eduardo Ribeiro de Mello

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

_____

Prof. Geraldo Zimbrão da Silva, D.Sc.

_____

Prof. Jano Moreira de Souza, Ph.D.

_____

Prof. Julia Celia Mercedes Strauch, D.Sc.

_____

Prof. Vania Bogorny, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MAIO DE 2008

MELLO, CARLOS EDUARDO RIBEIRO DE

Agrupamento de regiões: uma abordagem utilizando acessibilidade [Rio de Janeiro] 2008

XII, 84 p., 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2008)

Dissertação – Universidade Federal do Rio de Janeiro, COPPE

1. Mineração de dados espaciais
2. Análise de agrupamento

I. COPPE/UFRJ II. Título (série)

DEDICATÓRIA

Para meus pais, que sempre foram os alicerces da construção do meu caráter.

# Agradecimentos

Agradeço a Deus, por tudo, principalmente pela família que tenho, pois essa foi fundamental para a realização deste trabalho.

Aos meus pais, Isabel e Carlos, pelo amor, amizade, conselhos e educação. À minha mãe, um agradecimento especial pelas revisões desta dissertação.

À minha irmã, Mariana, pelo exemplo de perseverança e pelo carinho. À Giovana, minha sobrinha linda, que cada vez mais me convence que não sei nada – Entendeu Canjica?!.

À minha namorada Adria, pelo amor, paciência, amizade, conselhos e pela inspiração diária. Também pelas revisões desta dissertação e pela compreensão as minhas ausências.

À minha avó Irene, por todo seu carinho, amor e dedicação.

À minha avó Conceição (*in memoriam*) que sempre torceu por mim, me apoiando e me incentivando. Ao meu avô Ribeiro (*in memoriam*), pelo exemplo de humanidade.

A toda a minha família por compreender minhas ausências e desatenções.

Aos meus amigos Tuninho e Maura, pelas minhas ausências e furos nas viagens, mas principalmente, pelo apoio e amizade.

Aos meus amigos Zé, Diogo, Marzulo, Ivomar, Bruno, Thatiana e André, por contribuírem para esta pesquisa. Aos demais amigos da COPPE e da COPPETEC, pelo companheirismo.

Ao professor Geraldo Zimbrão, pela orientação, amizade e por acreditar no meu potencial e nas minhas idéias.

Ao professor Jano Souza, por aceitar participar da banca deste trabalho, por sua confiança no meu trabalho e pelo incentivo. À Vânia Bogorny, por aceitar participar da banca deste trabalho e pelos conselhos e sugestões acadêmicas. À professora Julia, por aceitar participar da banca e pela ajuda com os dados utilizados neste trabalho.

A todos o meu muito obrigado!

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

# AGRUPAMENTO DE REGIÕES: UMA ABORDAGEM UTILIZANDO ACESSIBILIDADE

Carlos Eduardo Ribeiro de Mello

Maio / 2008

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

O agrupamento de dados espaciais consiste em agrupar objetos espaciais do tipo ponto ou polígono, de maneira que seja considerada a componente espacial dos objetos. Embora muitos métodos de agrupamento para dados convencionais tenham sido desenvolvidos, existem poucos métodos que tratam de dados espaciais. Além disso, os métodos de agrupamento espaciais existentes utilizam a informação da vizinhança ou distância geográfica entre os objetos para agrupá-los. Nesta dissertação, propomos uma abordagem de agrupamento espacial através da informação de acessibilidade entre regiões. O objetivo é encontrar grupos de regiões semelhantes e acessíveis entre si. Além disso, adaptamos um método de agrupamento espacial para trabalhar com a abordagem proposta. Para avaliarmos nossa abordagem, esta foi comparada com a abordagem por vizinhança. Essa avaliação consistiu de um estudo de caso da distribuição do IDH (Índice de Desenvolvimento Humano) dos bairros da cidade do Rio de Janeiro. Neste, a abordagem proposta apresentou resultados de melhor qualidade.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

CLUSTERING REGIONS: AN ACCESSIBILITY APPROACH

Carlos Eduardo Ribeiro de Mello

May / 2008

Advisor: Geraldo Zimbrão da Silva

Department: Computer and Systems Engineering

Spatial clustering consists in grouping spatial objects, which can be points or polygons, considering the spatial component of the objects in the group formation. Although several clustering methods have been developed, there are few methods that work on spatial data. In addiction, those spatial clustering methods use neighborhood information or geographical distance among the objects to identify the groups. In this dissertation, we propose a clustering approach that uses accessibility information among the regions. It aims to identify clusters with similar and accessible regions. Besides, we modified a spatial clustering method to work on the proposed approach. In order to evaluate our approach, we compared both the accessibility and neighborhood approaches in an experiment. We conducted a case study of HDI (Human Development Index) distribution within Rio de Janeiro's neighborhoods. It showed that our approach achieves better results than the neighborhood approach.

# Índice

# Índice de Figuras

# Índice de Tabelas

# Capítulo 1 – Introdução

## 1.1 – Motivação

A informação geográfica é de suma importância na tomada de decisões e estabelecimento de estratégias em diversas áreas. Administradores de nações, indústrias e organizações de um modo geral fizeram e fazem uso de informações geográficas em muitas das suas decisões.

Devido ao crescimento da população mundial e sua pressão sobre os recursos naturais, diversos problemas que envolvem questões de localização geográfica estão cada vez mais em foco. Problemas relacionados ao meio ambiente, como o aumento da poluição, mudanças climáticas e análise de impacto ambiental chamam atenção de pesquisadores em todo o mundo. O crescimento desordenado de cidades também traz à tona questões como planejamento urbano, controle do tráfego de veículos e engenharia de transportes coletivos.

O crescimento do desenvolvimento tecnológico, principalmente na área da informática, possibilitou que organizações incorporassem informações geográficas a seus processos. Para lidar com esse tipo de informação, sugiram os sistemas de informação geográfica (SIG). Os SIG são sistemas utilizados para armazenar, analisar, manter e manipular dados geográficos de maneira automatizada (BOLSTAD, 2005).

Os dados geográficos utilizados pelos SIG podem ser imagens digitalizadas (*e.g.*, fotos de satélite) ou objetos que representam uma geometria no espaço, chamados objetos espaciais. Esses dados são armazenados e gerenciados por bancos de dados de imagem (imagens digitalizadas) e bancos de dados espaciais (objetos geométricos espaciais) (GUTING, 1994).

Nos últimos anos, os SIG vêm ganhando cada vez mais espaço no ambiente corporativo. Conseqüentemente, uma enorme quantidade de dados georreferenciados têm sido gerada, coletada e armazenada em bancos de dados espaciais. Além disso, institutos de pesquisa como o Instituto Brasileiro de Geografia e Estatística (IBGE) disponibilizam na Internet dados espaciais e não-espaciais (convencionais) sobre condições sociais e econômicas de cidades e regiões brasileiras.

Nesse contexto, extrair informação útil e conhecimento desses enormes volumes de dados disponíveis requer o uso de técnicas e ferramentas apropriadas. A área da

computação que trata dessas questões é chamada de Mineração de Dados (*Data Mining*) ou Descoberta de Conhecimento em Banco de Dados – DCBD (*Knowledge Discovery in Databases – KDD*) (CHEN *et al.*,1996, FAYYAD *et al.*, 1996).

A DCBD é definida como o processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em dados (FAYYAD *et al.*,1996). Em bancos de dados espaciais, a DCBD é utilizada para nos ajudar a entender os dados espaciais, descobrir relações entre os dados espaciais e não-espaciais, construir bases de conhecimento espacial, capturar características gerais de maneira simples e concisa *etc.* (KOPERSKI *et al.*, 1996).

Embora muitos trabalhos na área de mineração de dados em bancos de dados relacionais tenham sido de desenvolvidos, há poucos trabalhos que tratam de descoberta de conhecimento em bancos de dados espaciais (ESTER *et al.*, 2001). Em KOPERSKI *et al.*(1996), ESTER *et al.*(2000) e ESTER *et al.*(2001) são descritas as principais técnicas de mineração de dados espaciais. Uma dessas técnicas é o agrupamento (*clustering*).

O agrupamento consiste no processo de agrupar objetos físicos e abstratos em classes de objetos similares (HAN & KAMBER, 2001, CHEN *et al.*, 1996), isto é, alocar objetos semelhantes entre si em um mesmo grupo e objetos distintos em grupos diferentes (HAN *et al.*, 2001). O agrupamento também pode ser definido como um método de classificação não-supervisionada (CHEN *et al.*, 1996). Em outras palavras, o agrupamento serve para classificar um conjunto finito de objetos sem nenhum conhecimento *a priori*.

Os métodos de agrupamento para dados espaciais que encontramos na literatura são adaptações dos métodos para dados convencionais. A maior parte desses métodos utiliza a relação espacial de distância entre os objetos no espaço geográfico como critério de agrupamento. Essa simplificação acarreta a perda de informações espaciais potencialmente úteis, tais como a topologia de vizinhança e de acessibilidade entre regiões.

A utilização das relações topológicas entre objetos espaciais como polígonos e linhas pode trazer resultados mais eficazes para o agrupamento (ESTER *et al.*, 2001). Quando utilizamos regiões representadas por objetos espaciais do tipo polígono, podemos mapear sua topologia de vizinhança e utilizá-la para o agrupamento de regiões. Objetos do tipo linha representando rodovias podem ser utilizados para mapear a acessibilidade entre regiões.

As relações topológicas podem ser mapeadas através de consultas a bancos de dados espaciais. Os bancos de dados espaciais permitem realizar consultas com operadores espaciais para descoberta de relações entre os objetos como *intersecta*, *dentro de*, *cruza*, *corta etc.* (GUTING, 1994). Através da utilização de consultas com operadores espaciais é possível extrair topologias e utilizá-las nos métodos de agrupamento de regiões.

Uma das aplicações para os métodos de agrupamento que utilizam objetos espaciais do tipo polígono é a regionalização. Essa aplicação consiste em agrupar regiões ou unidades básicas de área semelhantes formando agregados contíguos. OPENSHAW (1977) aponta para o problema da necessidade de métodos automatizados de regionalização envolvendo dados de censo ou informações socioeconômicas.

Em ASSUNÇÃO *et al.* (2006) é apresentado o método de regionalização via árvore geradora mínima. Esse método agrupa regiões considerando a topologia de relações de vizinhança. O método utiliza uma das abordagens de classificação com restrição de contigüidade descrita em GORDON (1996). Com isso, o método identifica grupos de regiões contíguos, isto é, cada grupo é formado por regiões necessariamente vizinhas entre si.

Embora a abordagem de agrupamento através da topologia de vizinhança utilize relações espaciais mais proveitosas que a abordagem que utiliza as relações de distância, há casos em que essa abordagem também não é suficiente. Devido à existência de acidentes geográficos como rios, lagos, lagoas e montanhas, a topologia de vizinhança entre as regiões pode não ser um bom critério para o agrupamento de regiões. Há casos em que tais acidentes geográficos podem impedir a interligação de regiões semelhantes através das relações de vizinhança. Isso ocorre, pois os métodos que utilizam a topologia de vizinhança não contemplam a informação de acessibilidade entre as regiões.

A acessibilidade entre regiões pode ser de grande proveito para o agrupamento destas. Essa informação pode ser extraída através dos objetos espaciais do tipo linha que representam rodovias, aerovias, hidrovias *etc.* que interligam regiões. Dependendo do problema em questão, a topologia de acessibilidade consegue contemplar uma boa maneira de agregar informação geográfica sobre as conexões entre regiões.

Um exemplo de influência da acessibilidade é o caso de Rio de Janeiro e Niterói. Essas duas cidades não são vizinhas, devido à existência de uma baía entre elas. No entanto, o desenvolvimento humano dessas duas cidades é semelhante. Essa semelhança

entre as cidades ocorre porque existe uma ponte de apenas 13 quilômetros interligando-as. Outro fato importante é que o fluxo diário de veículos nessa ponte é grande, caracterizando uma alta acessibilidade entre Rio de Janeiro e Niterói. Portanto, as pessoas que vivem em Niterói ou no Rio de Janeiro podem usufruir dos serviços disponíveis de ambas as cidades facilmente.

As interligações entre regiões através de acessibilidade não garantem que as regiões formem grupos contíguos. Portanto, isto impede sua aplicação ao problema de regionalização. Por outro lado, acreditamos que uma abordagem de agrupamento de regiões através de suas relações de acessibilidade poderia identificar agrupamentos de melhor qualidade onde os dados não-espaciais em questão são influenciados pela acessibilidade.

Existem trabalhos como SILVA *et al.*(1998) e KREMPI *et al.*(2002) que exploram a distribuição da acessibilidade entre regiões urbanas. Em TUNG *et al.* (2001) temos um abordagem de agrupamento para objetos espaciais do tipo ponto na presença de obstáculos como, por exemplo, acidentes geográficos. No entanto, não encontramos na literatura nenhum trabalho que utilize a acessibilidade entre regiões para a identificação de grupos de regiões.

## 1.2 – Objetivo

O objetivo desta dissertação é desenvolver uma abordagem para o agrupamento de regiões utilizando a informação de acessibilidade. Para isso, propomos um método de agrupamento de dados espaciais que utiliza a acessibilidade entre as regiões como restrição para a formação de grupos. Esse método, chamado método de agrupamento com restrição de acessibilidade via Árvore Geradora Mínima (AGM), é uma modificação do método de agrupamento com restrição de contigüidade via AGM.

Um protótipo do método proposto foi implementado e utilizado em um experimento. O experimento consistiu em um estudo de caso cujos objetivos foram:

- Comparar os resultados obtidos utilizando o agrupamento por acessibilidade e por vizinhança. Para isso, foram utilizadas medidas de comparação para as duas abordagens; e
- Realizar uma análise de agrupamento das regiões do estudo de caso. Nessa tarefa foi realizada a análise exploratória dos dados e um estudo aprofundado dos grupos encontrados.

## 1.3 – Estrutura

Este trabalho está organizado em 5 capítulos, sendo este o primeiro.

O segundo capítulo consiste na revisão da literatura de técnicas/métodos de agrupamento. Nesse capítulo temos uma descrição dos conceitos envolvidos, os principais métodos de agrupamento para dados convencionais e para dados espaciais. Além disso, o método de agrupamento com restrição de contigüidade via árvore geradora mínima é descrito detalhadamente.

No capítulo 3, apresentamos a problemática do uso da topologia de vizinhança nos cenários onde existem acidentes geográficos e o método proposto nesta dissertação para solucionar esse problema.

O capítulo 4 consiste do experimento realizado com o método de agrupamento proposto e da análise dos resultados obtidos.

Finalmente, o capítulo 5 consiste das conclusões, considerações finais e trabalhos futuros desta dissertação.

# Capítulo 2 – Conceitos básicos e trabalhos relacionados

## 2.1 – Introdução

Os seres humanos, subconscientemente, exercitam a atividade de agrupar coisas desde a infância, agrupando objetos, plantas e animais de acordo com suas características. Com o objetivo de aprender sobre um novo objeto ou um novo fenômeno, as pessoas procuram as características deste de modo que possam descrevê-lo e compará-lo a outros objetos ou fenômenos conhecidos (XU & WUNSCH II, 2005). Essa comparação é realizada através da similaridade ou da dissimilaridade entre os objetos conhecidos e os novos objetos em questão. Portanto, para aprender sobre objetos ou fenômenos novos os seres humanos tentam classificá-los.

Os sistemas de classificação podem ser tanto supervisionados como não-supervisionados (XU & WUNSCH II, 2005). A classificação supervisionada consiste em associar um conjunto finito de objetos a rótulos de classes pré-definidos. Na classificação não-supervisionada, também chamada de agrupamento, não há nenhum rótulo pré-definido (XU & WUNSCH II, 2005). O objetivo do agrupamento é separar um conjunto finito de dados sem rótulos em conjuntos "naturais" finitos e discretos (XU & WUNSCH II, 2005). HAN & KAMBER (2006) e CHEN *et al.* (1996) definem *agrupamento* como o processo de agrupar um conjunto de objetos físicos ou abstratos em classes de objetos similares. Esse processo é realizado através da alocação de objetos similares em um mesmo grupo e de objetos dissimilares (distintos) em grupos diferentes (HAN *et al.*, 2001).

HAN & KAMBER (2006) explicam que aplicações típicas de métodos de agrupamento são encontradas em diversas áreas. Nos negócios, o agrupamento pode ser utilizado para identificar grupos de clientes com mesmo perfil de consumo. Em Biologia, para categorizar genes de funções semelhantes, taxonomias de animais e plantas, detectar grupos de cadeias de DNA *etc.*. Na área de geoprocessamento, esses métodos podem ser utilizados para agrupar regiões cujas características climáticas ou socioeconômicas são semelhantes.

Diversas áreas do conhecimento como estatística, aprendizado de máquina, bancos de dados, biologia, *marketing etc.* contribuíram e/ou contribuem para o

desenvolvimento dos métodos e técnicas de agrupamento. Por exemplo, a área de análise de agrupamento da estatística contribuiu fortemente para o desenvolvimento de vários métodos de agrupamento como o *k-means* e o *k-medoids* (MACQUEEN, 1967, KAUFMAN & ROUSSEEUW, 1990).

## 2.2 – Processo de análise de agrupamento

JAIN *et al.* (1999) e XU & WUNSCH II (2005) apresentam a análise de agrupamento como um processo para extração de conhecimento. Este é semelhante ao processo de descoberta de conhecimento definido por FAYYAD *et al.* (1996). O processo de análise de agrupamento apresentado por XU & WUNSCH II (2005) está ilustrado na Figura 2.1. A seguir apresentamos suas etapas.

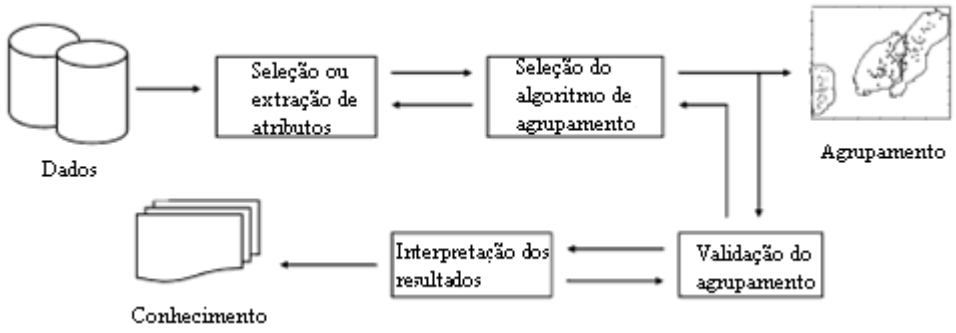


**Figura 2.1 – Processo de análise de agrupamento para extração de conhecimento**

1. **<u>Seleção ou extração de atributos</u>**

A análise de agrupamento é iniciada com a etapa de *seleção ou extração de atributos*. A *seleção de atributos* consiste em identificar o subconjunto dos atributos originais que seja mais apropriado para a realização do agrupamento. A *extração de atributos* consiste no uso de uma ou mais transformações nos atributos originais para produzir novos atributos. A normalização dos dados também é uma atividade realizada nessa etapa, seu objetivo é realizar transformações nos dados para que estes estejam em escala adequada para os métodos de agrupamento. XU & WUNSCH II (2005) afirmam que essa etapa é crucial para a qualidade dos resultados.

2. **<u>Definição ou seleção de métodos de agrupamento</u>**

A etapa de *definição ou seleção de algoritmos/métodos de agrupamento* está dividida em duas partes. A primeira é a *definição das medidas de similaridade*, que consiste na escolha da medida mais adequada para calcular a similaridade entre os objetos. Essa escolha depende dos tipos de dados dos atributos e do problema em questão. A segunda parte desta etapa é *definição ou escolha do método de agrupamento*. A escolha do método de agrupamento depende do tipo de problema, dos dados em questão e da análise que se deseja obter. Além disso, dependendo do método escolhido, alguns parâmetros são necessários para a sua execução. Por exemplo, há métodos onde precisamos utilizar como parâmetro o número de grupos que desejamos encontrar. Segundo XU & WUNSCH II (2005), não há nenhum método de agrupamento que seja universal para solução de todos os problemas. Portanto, uma investigação cuidadosa das características do problema em questão deve ser realizada para a escolha de um método de agrupamento apropriado.

Após a definição ou seleção do algoritmo/método de agrupamento, este é executado para o conjunto de dados selecionados. Como resultado dessa execução, temos os objetos divididos em grupos. Embora os métodos de agrupamento tentem encontrar uma estrutura de agrupamento para os objetos, isso não quer dizer que essa estrutura exista. Há casos em que os dados não estão distribuídos de maneira a formar grupos, isto é, os dados podem ser não agrupáveis (HAN & KAMBER, 2006). O problema de verificar se existe ou não uma estrutura de grupos presente nos dados é conhecido como tendência de agrupamento (XU & WUNSCH II, 2005). Mesmo assumindo que exista uma estrutura de grupos nos dados, os resultados do agrupamento devem ser avaliados e validados.

## 3. Validação do agrupamento

Nesta etapa do processo, iniciamos a *validação do agrupamento* (JAIN *et al.*, 1999). Segundo XU & WUNSCH II (2005), diferentes abordagens de agrupamento geralmente levam a diferentes agrupamentos. Mesmo para os casos onde são utilizados um mesmo algoritmo de agrupamento, os parâmetros ou a ordem de apresentação dos dados de entrada podem afetar o resultado final. Portanto, critérios e padrões efetivos de avaliação são importantes para prover aos usuários um resultado do agrupamento com certo grau de confiança (XU & WUNSCH II, 2005). Para isso, são utilizadas *medidas de validação*, que tentam avaliar de maneira objetiva os grupos encontrados. Essas medidas devem ser úteis para responder questões como quantos grupos estão presentes

nos dados, se os grupos obtidos possuem significado ou até mesmo para comparar, avaliar e escolher um método de agrupamento em detrimento a outro (XU & WUNSCH II, 2005). Uma leitura mais detalhada sobre métodos e medidas de validação de agrupamento pode ser encontrada em PAKHIRAA *et al* (2004), XIE & BENI (1991), DUBES (1987) e BEZDEK & PAL (1998).

**4. Interpretação dos resultados**

Após a validação dos grupos é realizada a etapa de *interpretação dos resultados*. Essa etapa consiste em generalizar os grupos de acordo com seus objetos, isto é, encontrar um rótulo para cada grupo. De um modo geral, essa tarefa é realizada através da análise estatística dos dados de cada grupo com o auxílio de um especialista no domínio do problema.

O processo de análise de agrupamento é um processo que inclui ciclos de *feedback* na sua execução, sendo um processo iterativo e interativo. Segundo XU & WUNSCH II (2005), em muitas circunstâncias, uma série de testes e repetições nesse processo são necessárias. Além disso, não há um critério universal e efetivo para guiar a seleção de atributos e métodos de agrupamento. Os critérios de validação fornecem alguns *insights* sobre a qualidade da solução, mas escolher um critério apropriado também é um problema que requer algum esforço.

## 2.3 – Medidas de similaridade

A medida de similaridade/dissimilaridade tem por objetivo mensurar o quanto dois objetos são semelhantes/distintos, conseqüentemente, esta depende dos tipos de dados dos objetos. Os dados podem ser quantitativos ou qualitativos, contínuos ou binários, nominais ou ordinais, para cada um desses tipos são adotadas medidas de similaridade apropriadas (XU & WUNSCH II, 2005).

Os objetos a serem agrupados geralmente são representados como um vetor de medidas ou por um ponto no espaço multidimensional (JAIN *et al.*, 1999). Cada componente do vetor corresponde a um atributo ou variável do objeto. Por exemplo, se quisermos agrupar um conjunto de pessoas de acordo com os dados da Tabela 2.1. Cada objeto é definido pelos atributos de uma pessoa. Portanto, para as pessoas Mariana, José

e Carlos temos os pontos no espaço: $(23;65;1.75)$, $(47;85;1.78)$ e $(24;95;1.82)$, respectivamente.

| *Objetos* | *Idade* | *Peso* | *Altura* |
|---|---|---|---|
| Mariana | 23 | 65 | 1.73 |
| José | 47 | 85 | 1.78 |
| Carlos | 24 | 95 | 1.82 |

**Tabela 2.1 – Dados de idade, peso e altura de pessoas**

A maioria dos métodos de agrupamento utiliza essa representação para identificar regiões no espaço onde há maior concentração de objetos (JAIN *et al.*, 1999). O processo de identificação de regiões é auxiliado por medidas de distância geométrica entre os pontos no espaço (HAN & KAMBER, 2001). Essas medidas indicam que quanto mais próximos os objetos, mais similares estes são. A Figura 2.2 ilustra um exemplo da distribuição dos objetos representados no espaço bidimensional e uma possível alocação desses objetos em grupos.

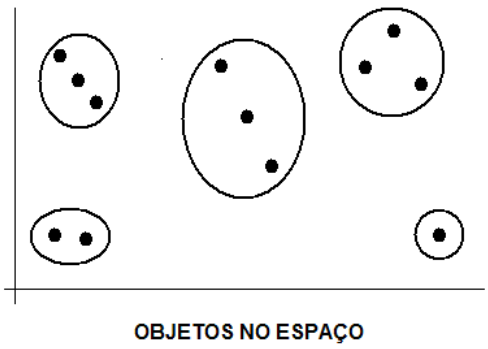


**Figura 2.2 – Distribuição dos objetos no espaço de variáveis**

Os tipos de dados mais utilizados nos métodos de agrupamento são os dados contínuos, conseqüentemente, existem inúmeras medidas de similaridade para estes. A distância entre os objetos no espaço pode ser adotada como medida de similaridade ou dissimilaridade (XU & WUNSCH II, 2005). As medidas de distância mais comuns são a distância Euclidiana, distância de Quarteirão (Manhattan) e a generalização destas, que é a medida de Minkowski (HAN & KAMBER, 2006).

A distância Euclidiana é dada pela fórmula:

$$D_{ij} = \left( \sum_{l=1}^{d} \left| x_{il} - x_{jl} \right|^{2} \right)^{1/2},$$

onde

- $D_{ij}$ é distância Euclidiana do objeto *i* para o objeto *j*;

- *l* é o índice do atributo do vetor de *d* atributos dos objetos; e

- $x_{il}$ é o *l*-ésimo atributo do objeto *i*.

A distância Euclidiana é amplamente utilizada, sua principal aplicação é no algoritmo de agrupamento *k-means*. Essa distância tende a encontrar grupos de objetos de formato esférico (XU & WUNSCH II, 2005).

A distância de Quarteirão é dada por:

$$D_{ij} = \sum_{l=1}^{d} \left| x_{il} - x_{jl} \right|,$$

onde as variáveis que a compõe possuem o mesmo significado que o da distância euclidiana. Nessa fórmula, notamos que, ao contrário de utilizarmos o quadrado da distância, utilizamos o módulo. Essa medida tende a formar grupos de formato retangular (XU & WUNSCH II, 2005).

A distância de Minkowski é a generalização das duas medidas anteriores. Esta é dada pela fórmula:

$$D_{ij} = \left( \sum_{l=1}^{d} \left| x_{il} - x_{jl} \right|^{n} \right)^{1/n}.$$

Nessa fórmula temos o parâmetro adicional *n*, que determina a forma dos grupos encontrados. Podemos notar que para *n=1* temos a distância de quarteirão e para *n=2* temos a distância euclidiana. Portanto, a distância euclidiana e de quarteirão são casos específicos da distância de Minkowski.

Outras medidas de similaridade para dados contínuos como distância de *sup*, distância de Mahalanobis, correlação de Pearson, distância de simetria de pontos e similaridade por cossenos podem ser encontradas em (XU & WUNSCH II, 2005).

Para que as medidas de similaridade baseadas em distância funcionem de maneira adequada é necessário que todas as variáveis estejam dentro do mesmo intervalo. Segundo HAN & KAMBER (2006), isso evita que as unidades de medida influenciem o agrupamento dos objetos, dando pesos diferentes às variáveis. Portanto, é necessário que a padronização ou normalização das variáveis seja realizada antes do

processo de agrupamento. Existem várias técnicas de padronização e normalização de dados na literatura. HAN & KAMBER (2001, 2006) e SHALABI *et al.* (2006) apresentam algumas dessas técnicas.

Um dos desafios para área de análise de agrupamento é definir medidas de similaridade para variáveis de outros tipos que não o contínuo. HAN & KAMBER (2006) e XU & WUNSCH II (2005) apresentam algumas dessas medidas. No entanto, nesta dissertação nos limitamos às medidas de similaridade para variáveis contínuas.

## 2.4 – Métodos de agrupamento

Existem diversos métodos de agrupamento na literatura, o que torna difícil suas categorizações, pois muitos destes sobrepõem várias categorias (HAN & KAMBER, 2006). Apesar disso, a categorização dos métodos de agrupamento é útil para apresentá-los de uma maneira organizada.

Em HAN & KAMBER (2001, 2006) e HAN *et al.* (2001), os métodos de agrupamento estão divididos em cinco categorias principais, conforme a seguir:

**Métodos de particionamento:** Para *n* objetos ou tuplas de dados, os métodos de particionamento constroem *k* partições de dados, onde cada uma destas representa um grupo de objetos. Nessas *k* partições, cada grupo deve conter pelo menos um objeto e cada objeto deve pertencer apenas a um grupo. O número de partições *k* é um parâmetro de entrada para os métodos de particionamento. Com esse valor, são construídas *k* partições iniciais e, através de uma técnica de realocação iterativa, o algoritmo tenta melhorar o esquema de partições movimentando os objetos de um grupo para o outro. Essa movimentação é realizada de acordo com algum critério que avalia a qualidade do particionamento. De maneira geral, o critério adotado para orientar o algoritmo é alocar objetos similares no mesmo grupo e objetos dissimilares em grupos diferentes. Para encontrar um valor ótimo global para a qualidade do agrupamento é necessário avaliar todas as possíveis combinações de partições. No entanto, esse procedimento é bastante custoso em termos de tempo. Para evitar isso, muitos métodos trabalham com heurísticas para a escolha das partições.

**Métodos hierárquicos:** Os métodos hierárquicos trabalham criando uma hierarquia de classes dos dados. Esses métodos podem ser classificados como aglomerativos ou divisivos, dependendo de como sua hierarquia de classes é construída.

Os métodos hierárquicos aglomerativos constroem sua hierarquia começando com *n* grupos, cada qual formado por um único objeto. A partir desses *n* grupos, os métodos aglomerativos juntam (ou unem) os grupos sucessivamente até que um único grupo seja formado. Os métodos divisivos seguem a filosofia oposta dos métodos aglomerativos. Estes iniciam com um único grupo contendo os *n* objetos. A partir deste, os grupos são separados sucessivamente até que *n* grupos de um único objeto sejam formados. Os métodos hierárquicos têm como característica não reavaliar as junções e separações dos grupos. Portanto, que uma vez que uma junção ou separação de grupos é realizada, esta nunca mais pode ser desfeita. Essa característica pode ser uma desvantagem, caso o critério para unir ou dividir os grupos não seja apropriado. Alternativas para resolver esse problema consistem em melhorar as interligações entre os grupos, isto é, a similaridade entre os grupos. Outra alternativa é utilizar métodos de particionamento em conjunto com os métodos hierárquicos.

**Métodos baseados em densidade:** Nos métodos de agrupamento baseados em densidade é utilizada a noção de densidade de objetos em um grupo. Sua idéia geral é definir um grupo baseado na densidade de sua vizinhança no espaço multidimensional. Os métodos baseados em densidade buscam novos objetos vizinhos no espaço de acordo com algum limite de densidade nas suas vizinhanças. A Figura 2.3, retirada de HAN & KAMBER (2006), ilustra o método DBSCAN, desenvolvido em ESTER *et al.* (1996). Esse método baseado em densidade procura por grupos através do uso de esferas no espaço multidimensional. De acordo com o número de pontos contidos na esfera, novos objetos são adicionados aos grupos. Os centros das esferas são definidos pelos objetos contidos em um grupo.

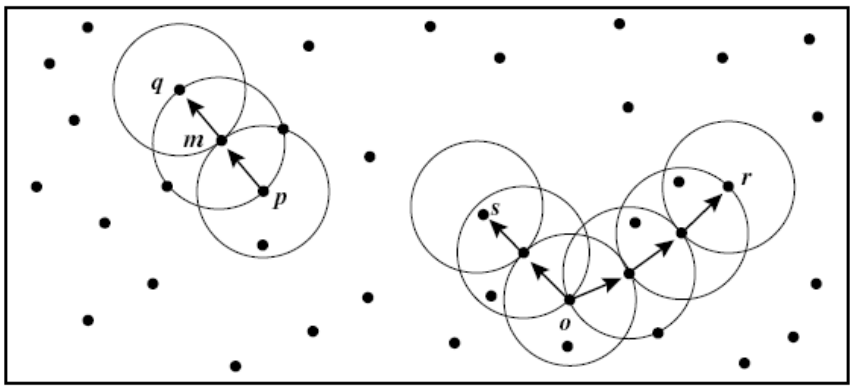


**Figura 2.3 – Execução do método de agrupamento DBSCAN**

O método DBSCAN e sua extensão OPTICS podem ser encontrados em ESTER *et al.* (1996) e ANKERST *et al.* (1999). Além destes, o método DENCLUE, que utiliza uma abordagem baseada em densidade através de funções, é descrito detalhadamente em HINNERBURG & KEIM (1998). A principal vantagem dos métodos baseados em densidade é que estes podem encontrar grupos de tamanhos variados e com formatos diferentes.

**Métodos baseados em GRID:** Os métodos baseados em GRID utilizam a abordagem de dividir o espaço de variáveis onde os objetos estão contidos em células que formam uma estrutura de malha. O agrupamento é realizado através de operações nessa estrutura de malha. O método STING, desenvolvido por WANG *et al.* (1997), é um exemplo de método de agrupamento baseado em GRID. Esse método divide o espaço em uma malha de células retangulares. Dependendo do tamanho da célula, esta pode conter um ou mais objetos, embora possam existir células sem nenhum objeto. Os objetos contidos em uma mesma célula formam um grupo. O STING aumenta o tamanho das células construindo uma hierarquia de grupos, conforme ilustrado na Figura 2.4.

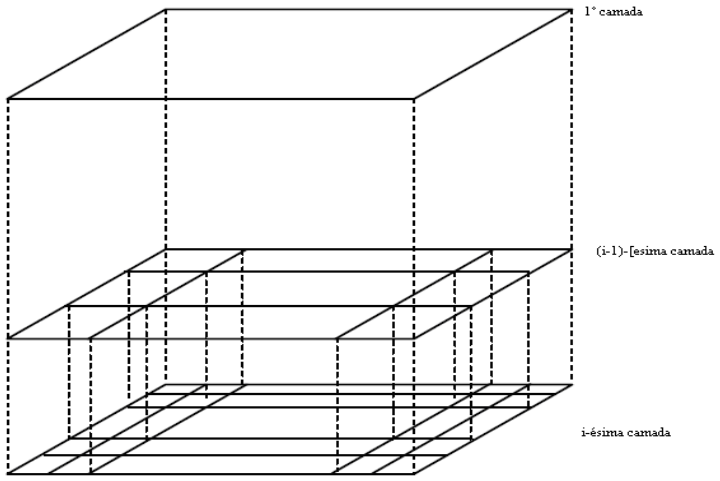


**Figura 2.4 – Divisão de células do método STING**

A principal vantagem dos métodos baseados em GRID é seu rápido tempo de processamento. Estes não dependem do número de objetos, mas sim do número de células que formam a estrutura em malha no espaço de variáveis. O método STING e o

método WaveCluster, que também utiliza a abordagem de agrupamento em GRID, encontram-se descritos em WANG *et al.* (1997 e SHEIKHOLESLAMI *et al.* (1998).

**Métodos baseados em modelo:** Os métodos de agrupamento baseados em modelo têm por objetivo utilizar um modelo estatístico para cada grupo e agrupar os objetos de acordo com esses modelos. Esses métodos podem definir grupos através de funções de densidade que refletem a distribuição espacial dos objetos. O EM (*Expectation Maximization)* é um método baseado em modelo, onde cada grupo é representado matematicamente por uma distribuição de probabilidade paramétrica (DEMPSTER *et al.*, 1977). Esse método consiste em estimar os parâmetros das distribuições de probabilidade, de forma que estas se adéqüem à distribuição dos dados. O COBWEB é outro método baseado em modelo, desenvolvido por FISHER (1987). Este utiliza um algoritmo de aprendizado conceitual que realiza uma análise de probabilidade e, a partir da definição de conceitos, os utiliza como modelos para os grupos. Outro método desta categoria é o SOM, um método baseado na utilização de redes neurais. A descrição detalhada desses métodos encontra-se em HAN & KAMBER (2006).

Embora HAN & KAMBER (2006) separem os métodos de agrupamento nessas cinco categorias, JAIN *et al.*(1999) defende que os métodos de agrupamento estão organizados em duas categorias principais: particionamento e hierárquicos. As outras categorias são subcategorias ou derivações dos métodos de particionamento e dos métodos hierárquicos.

Os principais métodos de agrupamento para dados espaciais são adaptações dos métodos de particionamento, portanto descrevemos detalhadamente seus principais métodos na seção 2.5. Por outro lado, o método de agrupamento com restrições que é objeto de estudo desta dissertação segue a abordagem dos métodos hierárquicos. Portanto, apresentamos os dois principais métodos hierárquicos na seção 2.6.

## 2.5 – Métodos de particionamento

Os métodos de particionamento têm sido, dos métodos de agrupamento, os mais populares antes do surgimento da mineração de dados (HAN *et al.*, 2001). O objetivo desse métodos é construir $k$ partições de $n$ objetos com $k \leq n$, onde cada partição

representa um grupo (HAN & KAMBER, 2006). Além disso, as seguintes condições devem ser satisfeitas:

(1) Cada grupo deve conter pelo menos um objeto; e

(2) Cada objeto deve pertencer a exatamente um único grupo.

Os métodos de particionamento organizam os objetos em *k* grupos de forma que uma função ou medida de erro para a distribuição dos grupos seja minimizada (HAN *et al.*, 2001). Essa medida é baseada nos desvios dos objetos em relação aos pontos que definem os centros dos seus grupos. O desvio de cada objeto em relação ao centro de seu grupo pode ser computado através de uma medida de similaridade (HAN *et al.*, 2001).

Os métodos de particionamento são baseados na técnica de realocação iterativa (HAN & KAMBER, 2006). Dado o número de partições *k* e um conjunto de *n* objetos, a realocação iterativa inicialmente seleciona arbitrariamente *k* pontos que definem os *k* grupos. A partir destes, os *n-k* objetos restantes são associados aos *k* grupos. A partir dos grupos formados, são calculados novos pontos para representar os grupos. Um critério ou medida de erro é calculado para o agrupamento formado. O processo de descoberta de novos centros para representar os grupos e realocar os objetos de acordo com estes é repetido até que a medida de erro não possa mais ser reduzida.

O objetivo do critério ou medida de erro é tentar formar grupos de objetos o mais homogêneos possível, diminuindo a soma dos desvios dos objetos aos centros dos seus respectivos grupos. Nesse contexto, os métodos de particionamento mais conhecidos e utilizados são o *k-means* e o *k-medoids* (HAN & KAMBER, 2006).

### 2.5.1 – k-Means

O algoritmo de agrupamento *k-means* divide um conjunto de dados em *k* partições de modo a que a similaridade intragrupos (dentro) seja alta e a similaridade intergrupos (entre) seja baixa (MACQUEEN, 1967). O *k-means* utiliza como parâmetro de entrada o número de partições *k*. A similaridade dos grupos é medida através das distâncias no espaço multidimensional definido pelos atributos dos objetos. O *k-means* define os grupos de acordo com o ponto médio dos grupos. Este é definido pelos valores médios dos objetos que compõem o grupo ao qual pertencem, isto é, o centróide ou centro de gravidade do grupo (HAN & KAMBER, 2006).

A execução do *k-means* começa com a escolha aleatória de *k* objetos para representar os centros dos grupos. Definidos os *k* centros, os objetos restantes são associados aos centróides mais próximos. Essa associação é realizada através da distância entre um objeto e o centróide. Construídos os grupos, o algoritmo calcula a suas respectivas médias, definindo novos centróides. Novamente, os objetos são associados aos grupos de acordo com sua distância aos centróides. O processo de recalcular as médias dos grupos e realocar os objetos é repetido até que o critério de agrupamento convirja.

Tipicamente, o critério utilizado para convergência dos grupos é a soma dos erros quadráticos, dado por:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 ,$$

onde *E* é a soma dos erros quadráticos, $m_i$ é a média ou centróide do grupo $C_i$ e *p* é o vetor que representa um objeto no espaço. Dessa forma, o método pára de realocar os objetos quando encontra um mínimo local para o valor de *E*. Ou seja, o algoritmo pára quando os objetos não mudam mais de grupo.

A Figura 2.5 ilustra um exemplo de execução do *k-means* no plano para *k=3*. Os *k* centróides são representados através do sinal de "+". Na Figura 2.5(a), os *k* objetos são escolhidos aleatoriamente como centróides. Em seguida, o algoritmo associa os objetos restantes aos centróides. A linha pontilhada indica os grupos formados. Na Figura 2.5(b), são calculadas as médias dos grupos formados e definidos novos centróides. Novamente, de acordo com esses centróides, os objetos são realocados nos grupos. O algoritmo segue até que, conforme apresentado na Figura 2.5(c), os objetos não mudam mais de grupo, isto é, o valor do erro quadrático não mais pode ser reduzido.

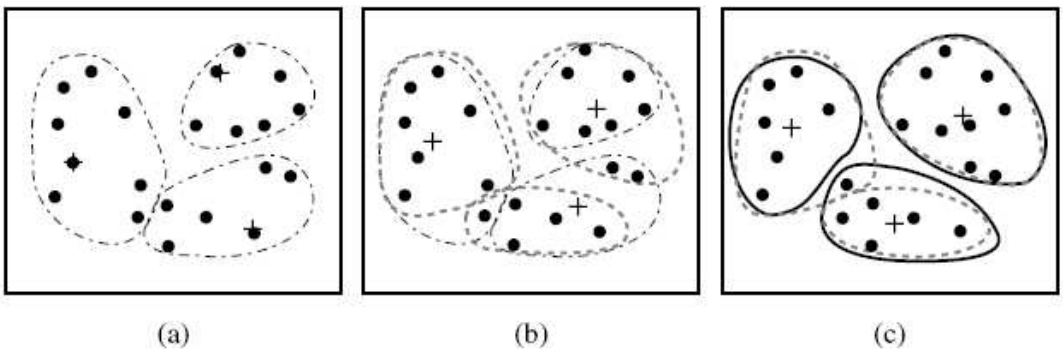


**Figura 2.5 – Execução do *k-means* (HAN & KAMBER, 2006)**

Segundo (HAN & KAMBER, 2006), o *k-means* funciona bem quando os grupos são compactos e separados uns dos outros. O método é relativamente escalável e eficiente para o processamento de grandes conjuntos de dados.

No entanto, o *k-means* só pode ser aplicado quando é possível definir a média dos grupos, o que não ocorre para variáveis categóricas. Além disso, o método não é adequado para a descoberta de grupos com formatos côncavos ou com grupos de tamanhos diferentes (HAN & KAMBER, 2006). Outra desvantagem do método é a necessidade de estimar o parâmetro *k*.

O *k-means* é sensível a presença de ruído ou valores extremos nos dados, pois estes influenciam fortemente a média. Conseqüentemente, isto afeta a posição dos centróides e a avaliação da homogeneidade dos grupos.

Existem algumas variações do *k-means*, estas podem diferir deste na seleção das *k* partições iniciais, na medida de similaridade ou na estratégia do cálculo da média (CHATURVEDI *et al.*, 1994, 2001). Uma iniciativa que pode trazer bons resultados é aplicar um método hierárquico para descobrir o valor de *k* e executar o *k-means* com esse valor de *k*. Para trabalhar com variáveis categóricas uma iniciativa é utilizar, ao invés da média, a moda dos dados e trabalhar com medidas de similaridade para variáveis categóricas (GANTI *et al.*, 1999).

### 2.5.2 – k-Medoids

Conforme apresentamos, o *k-means* é sensível a valores extremos. Esses valores distorcem a distribuição dos dados, afetando a média dos grupos e, conseqüentemente, o valor do erro quadrático (HAN & KAMBER, 2006).

O *k-medoids* tem por objetivo ser menos sensível à presença de valores extremos que o *k-means* (KAUFMAN & ROUSSEEUW, 1990). Para isso, esse método escolhe um objeto presente nos dados, chamado *medoid* ou objeto representativo, para representar o grupo. Portanto, ao invés de utilizar a média dos objetos, o *k-medoids* utiliza os *medoids* (KAUFMAN & ROUSSEEUW, 1990). Os objetos restantes nos dados são associados aos grupos de acordo com o valor de sua similaridade com os *medoids*.

O particionamento é realizado de maneira semelhante ao *k-means*. Entretanto, o critério de agrupamento consiste em minimizar a soma dos **erros absolutos**, dado por:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - o_i| \ ,$$

onde $E$ é a soma dos erros absolutos, $o_i$ é o objeto representativo ou *medoid* do grupo $C_i$ e $p$ é o vetor que representa um objeto no espaço.

### 2.5.3 – PAM (Partitioning Around Medoids)

O algoritmo PAM (Partinioning Around Medoids), desenvolvido por Kaufman e Rousseeum, utiliza a abordagem do *k-medoids* para encontrar *k* grupos através de objetos representativos (*meidoids*) como centróides dos grupos (KAUFMAN & ROUSSEEUW, 1990, NG & HAN, 1994, 2002).

O PAM começa com *k* objetos representativos (*medoids*) selecionados de maneira arbitrária. Os objetos restantes são associados aos grupos de acordo com sua similaridade com os *medoids* selecionados. Para todos os pares de objetos formados por um *medoid* e por um objeto não representativo é avaliada a troca do *medoid* pelo objeto não-representativo. Para isso, cada objeto não-representativo $o_{random}$, selecionado de um grupo *i* é confrontado com o *medoid* $o_i$ de seu grupo. O algoritmo calcula os valores de $E_{o_{random}}$ e $E_{o_i}$, isto é, os erros absolutos com $o_{random}$ e $o_i$ como *medoids* do grupo *i*, respectivamente. Caso o valor de $E_{o_{random}}$ seja menor que o valor de $E_{o_i}$, então $o_i$ deixa de ser o *medoid* do grupo *i* e $o_{random}$ assume esse papel. Isto faz com que um novo conjunto de *k* objetos representativos seja formado. Novamente, os objetos não-representativos são realocados de acordo com sua similaridade entre novos *medoids*. Esse processo é repetido até que nenhum objeto representativo seja trocado.

Segundo HAN & KAMBER (2006) e NG & HAN (1994), o PAM funciona bem com pequenos conjuntos de dados. Entretanto, sua eficiência é baixa para volumes de dados médios e grandes.

### 2.5.4 – CLARA (Clustering LARger Applications)

Uma abordagem baseada na técnica do *k-medoids* para grandes volumes de dados é o algoritmo CLARA (Clustering LARge Applications) (KAUFMAN & ROUSSEEUW, 1990, NG & HAN, 1994, 2002, HAN & KAMBER, 2006). Ao invés de

procurar objetos representativos em todo conjunto de dados, esse método seleciona uma amostra do conjunto de dados e aplica o algoritmo PAM (NG & HAN, 1994). Desse modo, o CLARA encontra os *medoids* que estão contidos em uma amostra sem ter que avaliar todo o conjunto de dados.

No entanto, os objetos da amostra devem ser selecionados de modo suficientemente aleatório para garantir que os *medoids* contidos na amostra se aproximem dos *medoids* para todo o conjunto de dados. Para isso, o CLARA seleciona múltiplas amostras do conjunto de dados e executa o PAM em cada uma delas. Com os *medoids* de cada uma das amostras, é calculada a soma dos erros absolutos desses *medoids* para todo o conjunto de dados. Os *medoids* que resultam no menor erro para todo o conjunto de dados são utilizados para definir o agrupamento.

Segundo NG & HAN (1994, 2002), o CLARA tem um desempenho satisfatório para grandes bases de dados. Entretanto, a qualidade dos resultados depende do número de amostras, de objetos e de grupos envolvidos na análise de agrupamento.

### 2.5.5 – CLARANS (Clustering LARge Applications based on RANdomized Search)

No método PAM, os melhores *k medoids* são procurados em todo o conjunto de dados, tornando-o proibitivo para grandes bases de dados. O método CLARA possui a mesma abordagem que o PAM, entretanto, este utiliza amostras do conjunto de dados para procurar os melhores *k medoids* dentre essas amostras. O CLARA pode não encontrar o melhor agrupamento para todo o conjunto de dados. Isso ocorre, caso nenhum dos melhores *k medoids* para todo o conjunto de dados estejam em uma das amostras selecionadas. Portanto, um bom agrupamento baseado em amostragem não representa necessariamente um bom agrupamento para todo o conjunto de dados (HAN & KAMBER, 2006).

O método CLARANS (Clustering LArge Applications based on RANdomized Search) utiliza a abordagem baseada na aplicação do PAM em amostras do conjunto de dados (NG & HAN, 1994, 2002). Entretanto, diferentemente do método CLARA, esse método escolhe as amostras do conjunto de dados dinamicamente durante sua execução.

No CLARA, a busca por *k medoids* em todo o conjunto de dados é modelada como um problema de grafos (NG & HAN, 2002). Os nós do grafo representam um conjunto de *k* objetos. Dois nós são vizinhos (isto é, estão conectados), se a interseção

entre eles possuir exatamente *k-1* objetos. Dessa forma, o CLARANS organiza em um grafo todos os possíveis conjuntos de *k medoids* contidos nos dados.

De acordo com abordagem do CLARANS, o PAM pode ser visto como uma busca por um nó mínimo em um grafo $G_{n,k}$ de *n* nós e *k* arestas (NG & HAN, 2002). A cada passo do algoritmo, todos os nós vizinhos de um determinado nó corrente são examinados. O nó corrente é trocado pelo nó vizinho que retorna um agrupamento de menor erro absoluto. A busca por nós vizinhos termina quando não houver mais nenhum nó vizinho que tenha erro absoluto menor que o nó corrente. Portanto, o PAM precisa avaliar $k*(n-k)$ nós vizinhos para cada nó corrente, o que torna esse método ineficiente para valores de *n* grande.

Por outro lado, o CLARA examina poucos vizinhos e restringe sua busca a pequenos subgrafos do grafo original $G_{n,k}$. Cada um desses subgrafos consiste de todos os nós que são subconjuntos (de cardinalidade *k*) de uma determinada amostra. Portanto, o CLARA possui um espaço de busca restrito aos nós contidos nesses subgrafos, tornando seu desempenho melhor que o do PAM. Entretanto, as possíveis soluções para o agrupamento dos dados ficam confinadas dentro dos subgrafos de $G_{n,k}$. Com isso, se um nó de $G_{n,k}$ que produz o agrupamento ótimo do conjunto de dados não pertencer a nenhum desses subgrafos gerados pelas amostras, o método nunca encontrará a solução ótima de agrupamento.

Assim como o CLARA, o método CLARANS não examina todos os vizinhos de cada nó do grafo. No entanto, o CLARANS não restringe sua busca a nenhum subgrafo particular de $G_{n,k}$. A diferença é que enquanto o CLARA define suas amostras no início do processamento, o CLARANS, a medida que avalia os nós, define dinamicamente as próximas amostras a serem avaliadas.

O CLARANS inicia seu algoritmo com um nó de $G_{n,k}$ arbitrário, chamado nó corrente ou atual. O custo do nó atual é comparado com o custo de seus vizinhos. Caso o nó vizinho tenha custo menor que o nó atual, este passa a ser o nó atual. O número máximo de vizinhos do nó atual a serem avaliados é definido como parâmetro para o algoritmo. Caso não exista nenhum nó vizinho com custo menor que o nó atual, o método marca o nó atual como sendo um mínimo local para o custo. Novamente, um novo nó corrente é escolhido arbitrariamente e o mesmo processo é executado. Isso é feito para que o algoritmo não fique preso em mínimos locais. O número de mínimos

locais a serem avaliados é passado como parâmetro para o algoritmo. O nó que obtiver o menor de todos os mínimos locais é considerado a solução do agrupamento.

Segundo NG & HAN (2002), o CLARANS apresenta experimentalmente maior eficiência que os algoritmos CLARA e PAM. Além disso, o CLARANS apresenta melhor qualidade em seus resultados que os outros métodos. A desvantagem do CLARANS é que além dos mesmos parâmetros dos outros métodos, este possui mais dois parâmetros de entrada (número de vizinhos a serem avaliados e número de mínimos locais). Dependendo do ajuste desses parâmetros, o CLARANS pode encontrar resultados de melhor ou pior qualidade em menos ou mais tempo de execução.

## 2.6 – Métodos hierárquicos

Nesta seção, apresentamos os dois principais métodos de agrupamento hierárquicos. O objetivo é apresentar a filosofia básica dos métodos hierárquicos, pois o método de agrupamento objeto de estudo desta dissertação utiliza essa abordagem.

Os métodos de agrupamento hierárquicos trabalham agrupando os objetos em uma estrutura de árvore de grupos (HAN & KAMBER, 2006). Essa estrutura organiza os grupos formando uma hierarquia. Dependendo da maneira como esta é construída, podemos classificar os métodos hierárquicos em aglomerativos ou divisivos (HAN & KAMBER, 2006, HAN *et al.*, 2001).

Os métodos aglomerativos constroem a hierarquia de grupos iniciando cada grupo com um único objeto. Ou seja, inicialmente temos o mesmo número de grupos que o número de objetos. Iterativamente, os grupos são unidos de acordo com um critério de similaridade entre eles. Isto é, os grupos mais similares são aglomerados, formando novos grupos. O método segue essa iteração até que algum limite de similaridade intergrupos (entre grupos) seja alcançado ou até que seja formado um único grupo com todos os objetos.

Os métodos divisivos utilizam a idéia oposta dos aglomerativos. Esses métodos iniciam seu algoritmo com um único grupo contendo todos os objetos. Em seguida, os grupos são subdivididos de acordo com a similaridade entre eles. Essas subdivisões são realizadas iterativamente até que algum limiar de similaridade seja alcançado ou até que todos os objetos pertençam cada um a um grupo diferente. Os métodos AGNES (AGglomerative NESting) e DIANA (DIvisia ANAlysis) utilizam a abordagem

aglomerativa e divisiva, respectivamente (KAUFMAN & ROUSSEEUW, 1990, HAN *et al.*, 2001).

O critério para juntar ou dividir dois grupos é baseado em uma medida de similaridade intergrupos. As quatro principais medidas para similaridade intergrupos são: a distância mínima, a distância máxima, a distância da média e a distância média. A seguir temos suas respectivas fórmulas, onde $|p - p'|$ é a distância entre dois objetos ou pontos *p* e *p'*, $m_i$ é a média do grupo $C_i$ e $n_i$ é o número de objeto em $C_i$.

**Distância mínima:** $d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$

**Distância máxima:** $d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$

**Distância da média:** $d_{mean}(C_i, C_j) = |m_i - m_j|$

**Distância média:** $d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$

O método AGNES utiliza como critério para unir dois grupos a distância mínima. O método DIANA utiliza a distância máxima para avaliar as possíveis separações dos grupos. Entretanto, dependendo do problema, outras medidas de similaridade entre grupos podem ser utilizadas.

A Figura 2.6 apresenta o funcionamento dos algoritmos AGNES e DIANA, conforme apresentado em HAN & KAMBER (2006). Nessa figura podemos verificar que a diferença básica entre esses dois métodos é a direção da construção da estrutura de hierarquia dos grupos.

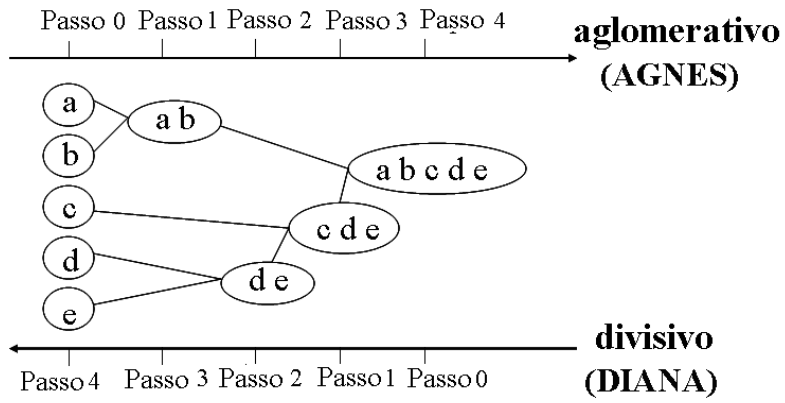


**Figura 2.6 – Agrupamento hierárquico utilizando o AGNES e DIANA (HAN & KAMBER, 2006)**

Essa estrutura em árvore pode ser representada através de um dendrograma. Este apresenta os objetos analisados e como estes são agrupados a cada passo do algoritmo. Além disso, o dendrograma permite que sejam visualizados os valores das similaridades intergrupos. Por exemplo, na Figura 2.7, temos um dendrograma para o agrupamento da Figura 2.6. Neste, podemos verificar que a similaridade entre os grupos (ou intergrupos) {a,b} e {c,d,e} é próxima de 0.2, ou que a similaridade entre os grupos (ou intergrupos) {c} e {d,e} é próxima de 0.4.

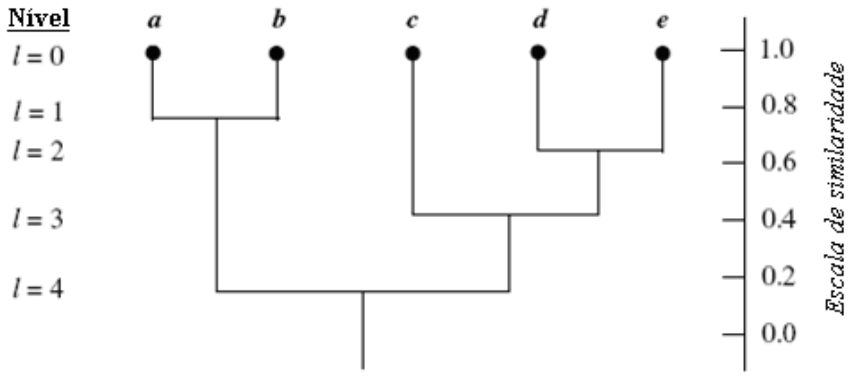


**Figura 2.7 – Estrutura de grupos através do Dendrograma (HAN & KAMBER, 2006)**

A principal vantagem dos métodos de agrupamento hierárquicos é a possibilidade de visualizar a hierarquia dos grupos. Essa hierárquica é útil quando precisamos conhecer como os grupos estão organizados entre si. Além disso, a visualização da hierarquia auxiliada por um dendrograma facilita a escolha do número de grupos que melhor divide os dados.

Os métodos hierárquicos têm como desvantagem seu desempenho. Para avaliar como os grupos serão unidos ou divididos é necessário que os *n* objetos sejam avaliados a cada *n* iterações do algoritmo. Portanto, isso faz com que o tempo de execução aumente exponencialmente em função da quantidade de dados. Além disso, os métodos hierárquicos não reavaliam suas junções ou divisões de grupos. Isto é, uma vez que uma operação de junção ou divisão de dois grupos é realizada, estes não serão divididos ou unidos novamente nos próximos passos do algoritmo. Ou seja, se uma junção ou divisão de grupos for realizada erroneamente durante a execução do algoritmo, esta não poderá ser desfeita. Uma iniciativa para resolver essa questão é a aplicação de algoritmos de particionamento em conjunto com algoritmos hierárquicos (HAN *et al.*, 2001).

Outros métodos de agrupamento hierárquicos como o BIRCH , ROCK, CURE e CHAMELEON podem ser encontrados em ZHANG *et al.* (1996), GUHA *et al.* (1999), GUHA *et al.* (1998), KARYPIS *et al.* (1999), respectivamente.

## 2.7 – Bancos de dados espaciais

Nesta seção apresentamos uma visão geral dos bancos de dados espaciais, suas características e funcionalidades.

Em várias áreas do conhecimento existe a necessidade de gerenciar e armazenar dados geométricos, geográficos ou espaciais, isto é, dados relacionados ao espaço. Por exemplo, um espaço de interesse na área de Geografia é a superfície da terra. Neste, temos interesse em armazenar a localização e a extensão de regiões, rios, ruas, *etc.*. Segundo GUTING (1994), para representar esses dados são utilizados tipos espaciais de dados, como LINHA, PONTO e REGIÃO (ou POLÍGONO).

Na Figura 2.8, estão ilustrados os três tipos espaciais básicos dos bancos de dados espaciais apontados por GUTING (1994). O tipo PONTO é utilizado para representar um objeto no espaço, indicando apenas sua localização. Esse tipo de dado espacial pode ser aplicado em situações onde a extensão do objeto no espaço não é relevante. O tipo LINHA, que representa uma curva no espaço, geralmente é utilizado para representar rotas no espaço, como rodovias, rios, cabos telefônicos *etc.*. A REGIÃO ou POLÍGONO é o tipo de dado utilizado em objetos nos quais queremos representar abstrações cuja forma e extensão no espaço são definidas. Por exemplo, cidades, países, bairros *etc.*.

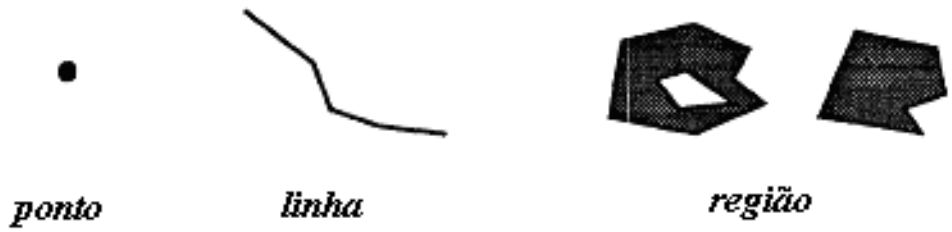


**Figura 2.8 – Três tipos espaciais básicos (GUTING, 1994)**

Os tipos de dados espaciais são utilizados em conjunto para modelar a estrutura geométrica das entidades no espaço, suas relações com outras entidades (*r intersecta l*),

propriedades (*área(r) > 1000*) e operações (*interseção*(*l*, *r*) – parte de *l* que está dentro de *r*) (GUTING, 1994).

Na Figura 2.9, temos os dois tipos mais importantes de coleções de objetos espacialmente relacionados apresentados por GUTING (1994). O primeiro tipo é chamado *partição*, onde este é um conjunto de regiões disjunto no espaço. Nessa coleção de objetos existem pares de regiões que possuem fronteira comum, conseqüentemente, uma topologia de vizinhança ou adjacência entre as regiões da *partição* pode ser estabelecida. Além disso, essa coleção de objetos pode ser utilizada para representar mapas temáticos. O segundo tipo de coleção de objetos espaciais é chamado de *rede*. Este pode ser visto como um grafo contido no plano. Neste, os pontos contidos nas linhas e as interseções entre linhas podem ser interpretados como os nós de um grafo e as linhas que conectam os pontos como arestas. O tipo *rede*, geralmente, é utilizado para representar malhas rodoviárias, mapas hidrográficos, redes de transporte *etc*.

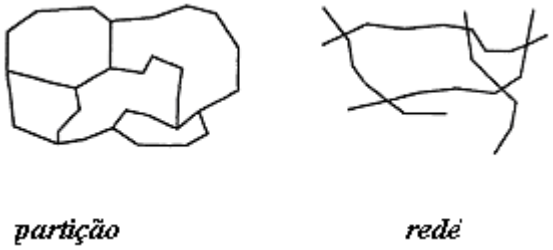


**Figura 2.9 – Coleções de objetos espaciais (GUTING, 1994)**

Os sistemas de bancos de dados espaciais (*Spatial Database Systems* - *SDBS*) são sistemas de bancos de dados que gerenciam e armazenam tipos de dados espaciais (ESTER *at al.*, 2000, 2001). Além disso, os bancos de dados também trabalham com os dados não-espaciais. A principal aplicação para os bancos de dados espaciais são os Sistemas de Informação Geográfica (SIG) (GUTING, 1994). Esses sistemas necessitam que o banco de dados dê suporte à consultas que utilizem operações e propriedades espaciais, além do suporte para os dados relacionais ou não-espaciais.

Um exemplo de tabela de um banco de dados espaciais é ilustrado na Figura 2.10. Nesta, temos os dados dos bairros da cidade do Rio de Janeiro com seus atributos não-espaciais e espaciais. O bairro da Barra da Tijuca está ilustrado com os valores de seus atributos não-espaciais. Além destes, há o atributo **espacial**, em que este é do tipo

região ou polígono, representando a localização e extensão desse bairro na cidade do Rio de Janeiro.

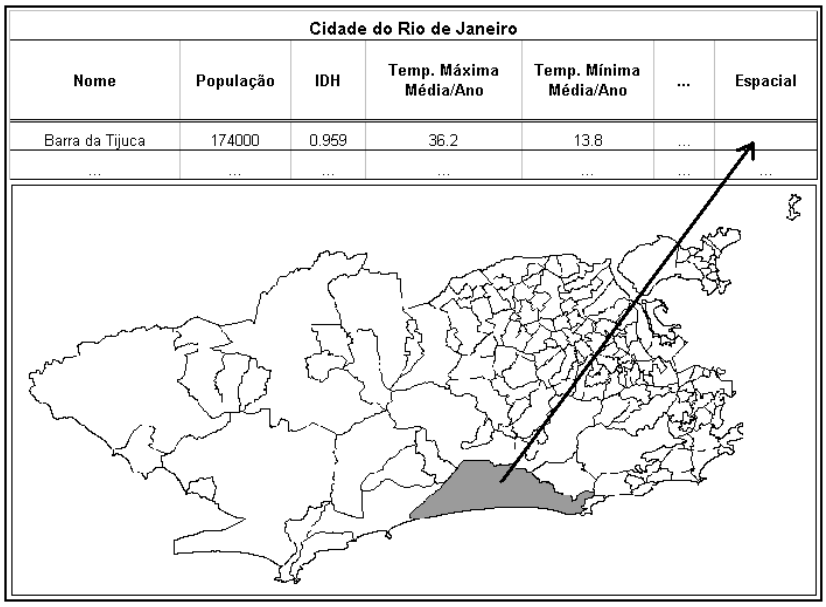


**Figura 2.10 – Tabela espacial com dados dos bairros da cidade do Rio de Janeiro**

Os bancos de dados espaciais fornecem suporte a consultas utilizando operações sobre relações espaciais entre os objetos (GUTING, 1994). Essas relações podem ser de três tipos:

- **Relações topológicas** como *toca*, *adjacente*, *dentro*, *disjunto* e *sobrepõe*. Estas são independentes de transformações nos objetos como translação, escala ou rotação (Figura 2.11(a));
- **Relações de direção** como *acima*, *abaixo*, *norte_de*, *sul_de*, *etc.* (Figura 2.11(b)); e
- **Relações de medida** como "*distância > 100*" ou "área > 100" (Figura 2.11(c)).

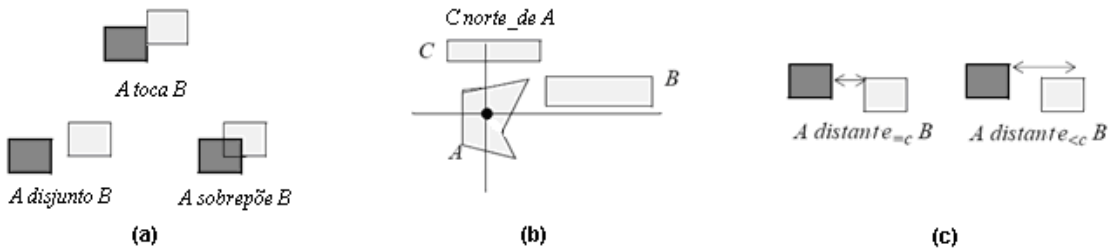


**Figura 2.11 – Exemplos de relações espaciais**

As operações através das relações entre os objetos espaciais permitem que consultas complexas sejam realizadas. Por exemplo, a consulta que retorna todos os bairros que são vizinhos a Barra da Tijuca poderia ser expressa através da seguinte consulta SQL no PostGis (POSTGRESQL, 2008):

```
"SELECT b.nome FROM BAIRROS a, BAIRROS b WHERE a.nome = 'BARRA DA TIJUCA' AND TOUCHES(a.espacial, b. espacial)"
```

Nessa consulta, o atributo **espacial** é o atributo que armazena a geometria dos objetos espaciais. A primitiva **TOUCHES** representa a relação espacial *toca*. Caso as geometrias dos objetos que são parâmetros dessa primitiva possuam fronteira comum, a mesma retorna o valor *verdadeiro*, caso contrário, retorna *falso*.

Portanto, a consulta retorna todos os bairros cuja geometria faz fronteira com o bairro Barra da Tijuca. Esse tipo de consulta é bastante comum em aplicações de SIG. Além disso, esta torna transparente para o SIG a complexidade das operações com objetos espaciais.

## 2.8 – Métodos de agrupamento espaciais

Nesta seção apresentamos os principais métodos de agrupamento espaciais. Embora esses métodos possam ser aplicados em diversas áreas, limitamos nosso escopo ao agrupamento de dados geográficos em duas dimensões.

Os métodos de agrupamento espaciais consistem em alocar objetos espaciais em grupos (NG & HAN, 1994). A diferença desses métodos para os métodos convencionais, descritos anteriormente, está no tratamento dos atributos espaciais (ESTER *et al.*, 2001). Nestes, as componentes espaciais são consideradas na execução do agrupamento, de maneira que os objetos dentro de um mesmo grupo possuam relações espaciais entre si.

Aplicações dos métodos de agrupamento espaciais podem ser encontradas em diversas áreas como Geologia, Geografia, Epidemiologia *etc.*. Exemplos dessas aplicações como a detecção de falhas sísmicas, a identificação de regiões de influência, o controle de epidemias *etc.* podem ser encontrados na literatura (ESTER *et al.*, 2001, HAN & KAMBER, 2006).

O problema de agrupar objetos espaciais pode ter duas abordagens em relação ao espaço de atributos. A primeira abordagem é quando desejamos agrupar os objetos considerando somente os atributos espaciais. Nesta, os objetos de um mesmo grupo estão relacionados espacialmente, mas não necessariamente seus atributos não-espaciais são similares. Na segunda abordagem, os atributos não-espaciais dos objetos também são considerados para a formação dos grupos. Portanto, nessa abordagem, os objetos de um mesmo grupo, além de serem espacialmente relacionados, também possuem atributos não-espaciais similares.

A escolha da relação espacial que estará presente nos objetos de um mesmo grupo também pode variar. A relação espacial de distância entre os objetos no espaço geográfico é amplamente utilizada. Essa relação espacial fornece uma noção natural de vizinhança entre os objetos espaciais. No entanto, esta pode não funcionar eficientemente para objetos do tipo polígono. Para esse tipo de objetos, as relações topológicas como *tocam*, *intersectam etc.* podem ser mais apropriadas, principalmente se esses objetos possuírem extensão e formas muito diferentes (ESTER *et al*., 2002).

Os principais métodos de agrupamento espaciais encontrados na literatura são o PAM, o CLARA e o CLARANS (NG & HAN, 1994, 2002, ESTER *et al.*, 2001, KOPERSKI *et al.*, 1996). Esses métodos utilizam a relação de distância entre os objetos como critério de agrupamento. Além destes, temos suas variações como o (SD) CLARANS e o (NSD) CLARANS (NG & HAN, 1994). Nestes, o agrupamento espacial considera tanto os atributos espaciais como os atributos não-espaciais dos objetos. Na seção 2.8.1, descrevemos como o PAM, o CLARA e o CLARANS são utilizados como métodos de agrupamento espaciais e descrevemos o funcionamento dos métodos SD(CLARANS) e NSD(CLARANS).

GORDON (1996) apresenta várias técnicas para o problema de classificação com restrições. Nesse trabalho, as técnicas apresentadas são aplicadas ao agrupamento de regiões espaciais através do uso de relações topológicas. Uma dessas técnicas consiste em utilizar a topologia de vizinhança entre as regiões. Esta é representada por dispositivos auxiliares como grafos ou matrizes. A partir desses grafos ou matrizes, métodos de agrupamento hierárquicos podem ser aplicados para detectar os grupos.

O método AZP (*Automatic Zoning Procedure*) utiliza um grafo de vizinhança para representar a topologia entre os objetos espaciais. Este utiliza a técnica de particionamento do *k-Means* para alocação dos objetos em grupos, sempre mantendo as restrições de contigüidade representadas no grafo de vizinhança (NEVES, 2003). Com

isso, o AZP encontra grupos de regiões contíguas no espaço, isto é, todas as regiões de um mesmo grupo são vizinhas entre si.

Em ASSUNÇÃO *et al.* (2006) é apresentado o método de agrupamento com restrições de contigüidade via árvore geradora mínima. Este utiliza a técnica de mapeamento da topologia espacial apresentada por GORDON (1996) em conjunto com a técnica de agrupamento em grafos utilizando árvore geradora mínima descrita por JAIN *et al.*(1999). Esse método apresenta melhores resultados que o método AZP (NEVES, 2003).

### 2.8.1 – Agrupamento espacial com PAM, CLARA e CLARANS

Os métodos de agrupamento convencionais podem ser aplicados aos dados espaciais, de forma que estes considerem como medida de similaridade a distância entre os objetos representados no espaço. Isso faz com que os métodos convencionais encontrem grupos de objetos espaciais, onde objetos pertencentes a um mesmo grupo sejam próximos ou "vizinhos". Em outras palavras, esses métodos definem áreas no espaço geográfico com alta densidade de objetos espaciais (ESTER *et al.*, 2001).

Existem diversas abordagens para medir a distância entre objetos espaciais (NG & HAN, 2002). A abordagem mais simples para o agrupamento espacial consiste em representar os objetos espaciais como pontos no espaço. Dessa forma, podemos aplicar métodos de agrupamento convencionais, onde medidas de distância como a distância de Quarteirão ou a distância Euclidiana podem ser utilizadas para calcular a distância entre os objetos (pontos no espaço geográfico).

NG & HAN (1994, 2002) apresentam experimentos comparando os métodos de *particionamento* PAM, CLARA e CLARANS utilizando dados espaciais do tipo ponto. Nestes, os objetos são definidos por pares de coordenadas geográficas *(x,y)*. A diferença do resultado do agrupamento espacial para o convencional está na interpretação do resultado. Ao invés de estarmos trabalhando com os objetos no espaço de atributos não-espaciais, estamos trabalhando com os objetos no espaço físico ou geográfico.

A Figura 2.12 ilustra um exemplo de agrupamento de casas em uma região. Estas são representadas por pontos no espaço, definidos pelas suas coordenadas geográficas (latitude e longitude). Através desses pontos, um método de *particionamento* identifica 11 grupos. Estes indicam regiões onde há maior densidade de casas, possivelmente condomínios, vilas ou ruas residenciais.

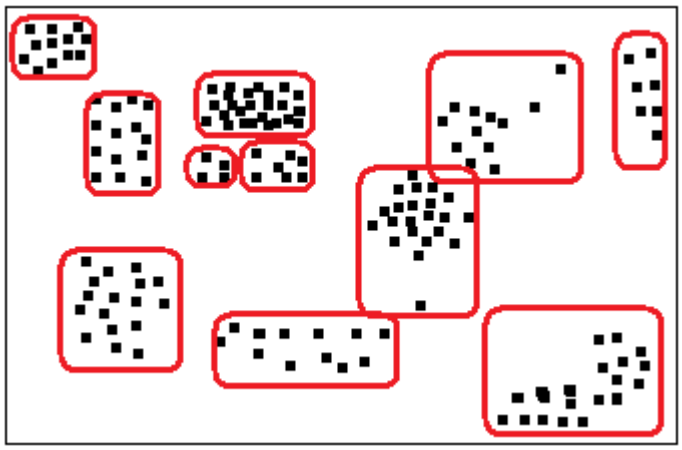

**Figura 2.12 – Exemplo de agrupamento de casas representadas por pontos no espaço.**

A representação de objetos espaciais através de pontos no espaço geográfico pode alcançar bons resultados, principalmente quando a extensão e forma dos objetos espaciais não são relevantes. No entanto, há casos em que queremos agrupar objetos cujas extensões e formas são relevantes para o agrupamento como, por exemplo, shoppings, parques, estacionamentos, bairros, cidades *etc.*. Nesse caso, a redução dos objetos espaciais em pontos acarreta uma grande perda de informação espacial. Os objetos do tipo região ou polígono carregam, além da localização espacial, informações sobre suas extensões e formas. No entanto, a dificuldade de trabalhar com esses tipos de objetos reside em como medir a distância entre dois polígonos.

NG & HAN (2002) apresentam e comparam três maneiras diferentes de calcular a distância entre dois polígonos. A primeira maneira consiste em calcular a distância exata de separação entre dois polígonos convexos (Figura 2.13(a)). A segunda consiste em calcular a distância mínima entre os vértices dos polígonos (Figura 2.13(b)). A terceira consiste em encontrar o menor retângulo que contém todo o polígono e calcular a distância exata entre os dois retângulos (Figura 2.13(c)).

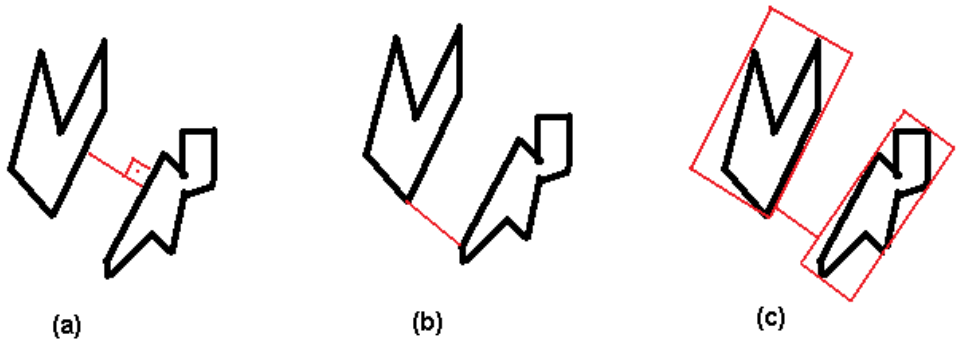


**Figura 2.13 – Medidas de distância entre polígonos**

Além disso, os autores descrevem os resultados de experimentos com as três medidas entre polígonos aplicadas aos métodos PAM, CLARA e CLARANS. Os grupos encontrados pelos três métodos formam regiões contíguas no espaço.

Para considerar os atributos não-espaciais dos objetos no agrupamento, podemos simplesmente adicioná-los nos vetores de atributos. No entanto, pesos diferentes nas componentes espaciais devem ser considerados para que a contigüidade dos grupos seja mantida. O método SAGE utiliza essa abordagem, sua principal desvantagem é necessidade da escolha dos pesos apropriados para as componentes espaciais (NEVES, 2003).

NG & HAN (1994) apresentam outras duas abordagens para considerar os atributos não-espaciais no agrupamento. A primeira abordagem é chamada de dominante espacial e a segunda é chamada de dominante não-espacial. Essas duas abordagem utilizam um método de particionamento, no caso o CLARANS, em conjunto com um método que constrói regras a partir da generalização de atributos, chamado DBLearn (HAN *et al.*, 1992).

Na abordagem de dominante espacial, SD(CLARANS), os objetos são agrupados considerando apenas as componentes espaciais utilizando o CLARANS. Em seguida, o DBLearn é aplicado em cada um dos grupos encontrados. Os resultados são regras que generalizam os atributos não-espaciais encontrados em cada um dos grupos formados. A abordagem de dominante não-espacial, NSD(CLARANS), funciona de forma oposta. Inicialmente, são encontradas regras a partir da generalização dos atributos não-espaciais em todo conjunto de dados através da execução do DBLearn. Em seguida, as componentes espaciais das regras mineradas são agrupadas utilizando o CLARANS. Em geral, o SD(CLARANS) é mais eficiente que o NSD(CLARANS), no

entanto, dependo do tipo de conhecimento que se deseja encontrar, a escolha dos métodos pode variar (NG & HAN, 1994, KOPERSKI *et al.*, 1996).

Os métodos que utilizam a relação de distância entre os objetos espaciais podem não alcançar bons resultados, pois estes não utilizam as relações topológicas entre os objetos para o agrupamento.

### 2.8.2 – Método de agrupamento com restrição de contigüidade via árvore geradora mínima

O método de agrupamento de dados espaciais com restrição de contigüidade via Árvore Geradora Mínima (AGM) foi proposto por NEVES *et al.* (2002) e ASSUNÇÃO *et al.* (2006). O objetivo desse método é a regionalização, isto é, identificar grupos de regiões contíguas e com características (atributos não-espaciais) semelhantes. Para isso, esse método utiliza um grafo ou matriz de vizinhança (ou contigüidade) para mapear a topologia de vizinhança entre as regiões. A partir desse mapeamento, o problema de encontrar grupos de regiões é abordado como um problema de agrupamento em grafos descrito em JAIN e*t al.*(1999).

Esse método é aplicado apenas em objetos do tipo região onde estes estejam distribuídos de modo a formar uma coleção de objetos espaciais do tipo *partição* (veja seção 2.7). Essa limitação do método ocorre porque o mapeamento das relações de vizinhança é feito a partir da operação espacial *tocam*, descrita na seção 2.7 deste capítulo.

A seguir, descrevemos os passos realizados pelo método de agrupamento com restrição de contigüidade via AGM.

**1. Construção do grafo ou matriz de contigüidade**

Nesse passo, as relações de vizinhança entre as regiões são mapeadas através um grafo de vizinhança. Nesse grafo, cada nó é representado por uma região e cada aresta representa uma relação de vizinhança entre duas regiões. A Figura 2.14, retirada de NEVES (2003), ilustra um exemplo.

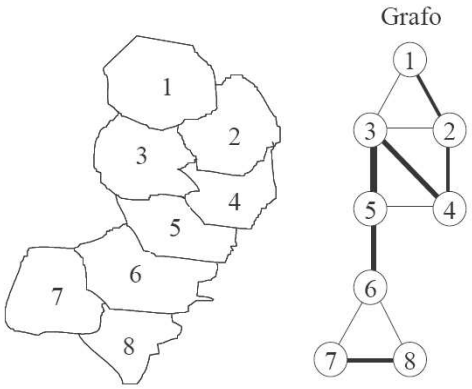


**Figura 2.14 – Construção do grafo de vizinhança (NEVES, 2003)**

A topologia de vizinhança também pode ser representada através de uma matriz de vizinhança $A_{n\times n}$, onde *n* é o número total de regiões a serem agrupadas. Cada relação de vizinhança entre duas regiões *i* e *j* é representada na matriz pelas componentes $a_{i,j} = a_{j,i} = 1$. O restante das componentes da matriz possui valor $a_{x,y} = a_{y,x} = 0$, indicando que não há relação de vizinhança entre as regiões *x* e *y*.

**2. Geração da árvore geradora mínima**

Nesse passo, os pesos para as arestas do grafo de vizinhança são calculados e sua árvore geradora mínima é descoberta.

Os pesos das arestas são definidos pelas medidas de similaridade entre os objetos conectados por estas. A medida de similaridade adotada é a distância Euclidiana entre os vetores de atributos não-espaciais das regiões representadas pelos nós. Com isso, quanto mais semelhantes são os atributos não-espaciais, menor é medida de similaridade.

A partir do grafo de vizinhança com os pesos de suas arestas calculados, executamos o algoritmo PRIM (CORMEN *et al*., 2001). Esse algoritmo descobre a árvore geradora mínima de um grafo a partir da eliminação das arestas de maior peso que formam ciclos. A Figura 2.15 ilustra um exemplo de execução do PRIM.

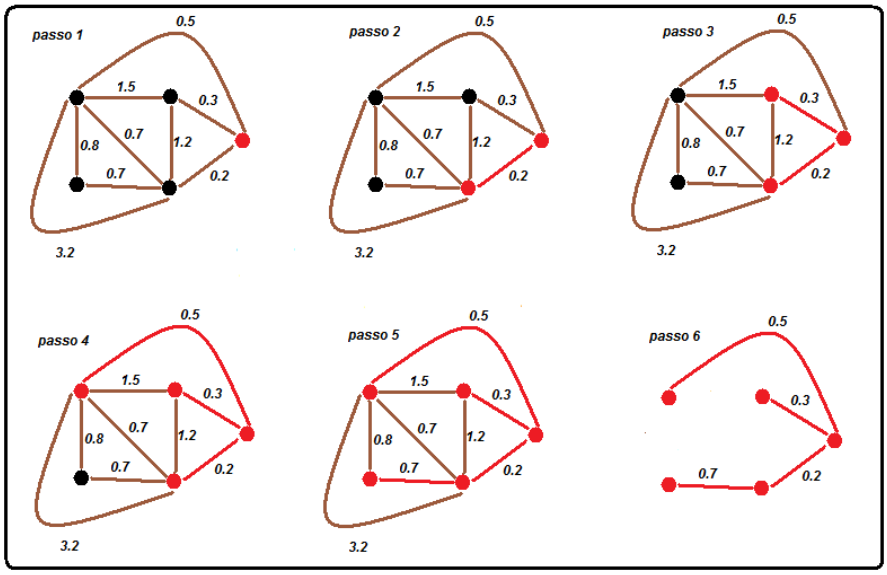


**Figura 2.15 - Exemplo de execução do algoritmo PRIM**

Portanto, a AGM representa as relações de vizinhança entre os objetos de maior similaridade.

## 3. Poda da AGM

O procedimento de retirada de arestas da árvore geradora mínima (AGM) é chamado de poda. Esse procedimento é semelhante aos métodos de agrupamento hierárquicos divisivos, apresentados na seção 2.6.

Na poda da AGM, cada aresta eliminada gera duas novas sub-árvores, isto é, dois grupos. Dessa maneira, se quisermos encontrar 3 grupos, basta eliminarmos 2 arestas. Ou seja, a partir de um grupo são gerados subgrupos deste, formando uma hierarquia entre os grupos. Portanto, é necessário que um critério seja utilizado para a divisão da árvore, isto é, um critério de escolha das arestas a serem podadas.

Para isso, é definido um custo para todas as arestas do grafo. Este é calculado através seguinte fórmula:

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

onde $SQD_T$ é a *soma dos quadrados dos desvios* dos atributos não-espaciais dos objetos contidos na árvore $T$ na qual a aresta $l$ se encontra. Esse valor é obtido através da fórmula:

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

onde:

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

$x_{ij}$ é o atributo não espacial *j* do objeto *i*;

*m* é o número de atributos não-espaciais considerados na análise;

$\bar{x}_j$ é o valor médio do atributo *j*.

A parcela $SQD_l$ do *custo da aresta l* consiste na *soma dos quadrados dos desvios* das duas sub-árvores geradas a partir da poda da aresta *l* em *T*. Essa parcela é dada por:

$$SQD_l = SQD_{Ta} + SQD_{Tb},$$

onde $SQD_{Ta}$ é a *soma dos quadrados dos desvios* da sub-árvore *Ta* de *T* e $SQD_{Tb}$ da sub-árvore *Tb* de *T*, conforme ilustrado na Figura 2.16.

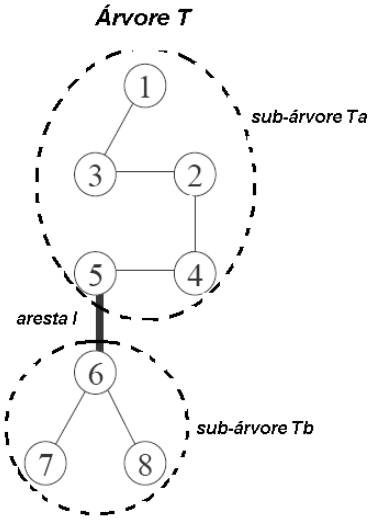


**Figura 2.16 – Sub-árvores geradas de *Ta* e *Tb*.**

O custo da aresta representa uma medida da diferença de homogeneidade entre a árvore *T* e as sub-árvores *Ta* e *Tb*. Portanto, quanto maior o valor do custo, menor é a dispersão entre os grupos formados pelas sub-árvores *Ta* e *Tb* em relação à árvore *T*. À medida que arestas são retiradas, novas árvores são geradas e os custos de suas arestas são re-calculados.

O critério de minimizar a soma dos quadrados dos desvios utilizado na poda da AGM é o mesmo critério de agrupamento adotado pelo *k-means* (vide seção 2.5.1).

Portanto, a escolha das arestas de maior custo resultará em um agrupamento de regiões contíguas de menor dispersão intragrupos.

O método de agrupamento com restrição de contigüidade via AGM está implementado na ferramenta SKATER (SKATER, 2008). Esta permite a visualização dos grafos de vizinhança, da AGM e dos grupos detectados. Embora essa ferramenta não seja de código-aberto, seu uso é gratuito.

## 2.9 – Considerações finais

Neste capítulo realizamos a revisão da literatura sobre métodos de agrupamento e métodos de agrupamento espaciais. Nesta, constatamos que a maioria desses métodos agrupa os objetos através das relações de distância no espaço geográfico. No entanto, ESTER *et al.* (2001) sugerem que o uso de relações espaciais topológicas podem apresentar melhores resultados para o agrupamento de objetos do tipo região.

Uma iniciativa de método de agrupamento que utiliza a topologia espacial de vizinhança é o método de agrupamento com restrição de contigüidade via AGM. Este apresenta bons resultados comparado a outros métodos espaciais (NEVES, 2003).

Nenhum dos métodos de agrupamento espaciais pesquisados na literatura utiliza a informação de acessibilidade como critério de agrupamento. Portanto, concluímos a necessidade de desenvolver um método que utilize a informação de acessibilidade.

# Capítulo 3 – Agrupamento de regiões utilizando acessibilidade

## 3.1 – Introdução

Na seção 2.8.2 do capítulo 2, descrevemos o método de agrupamento com restrição de contigüidade via AGM. Esse método realiza o agrupamento de regiões, de modo que cada grupo formado seja contíguo. Em outras palavras, esse método agrupa os objetos espaciais semelhantes respeitando as restrições de contigüidade dentro dos grupos.

No entanto, nem sempre a restrição da contigüidade é uma boa maneira de agregar informação espacial à formação de agrupamentos. Existem casos em que essa restrição pode acabar dividindo regiões com características semelhantes. Isso ocorre devido à existência de acidentes geográficos, como montanhas, matas, lagos e rios.

Um exemplo onde o método de agrupamento com restrição de contigüidade via AGM não gera resultados satisfatórios é ilustrado na Figura 3.1. Nessa figura, temos um mapa de três regiões: A, B e C. As regiões A e C são muito semelhantes, ambas possuem aproximadamente a mesma população e atividade econômica. Por outro lado, a região B, formada por um grande morro, possui população e atividade econômica bem menor que suas vizinhas.

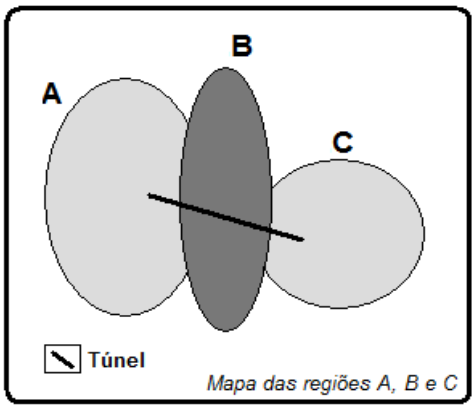


**Figura 3.1 – Mapa das regiões A, B e C**

O método de agrupamento com restrição de contigüidade via AGM pode agrupar as três regiões do mapa em dois ou três grupos. Para dois grupos, temos necessariamente um grupo formado pela região B junto com alguma das outras duas, gerando um grupo bastante heterogêneo. No agrupamento para três, cada grupo é formado de uma única região. Nos dois casos, a região B funciona como um divisor entre as duas regiões mais semelhantes, conforme apresentado na Figura 3.2 para o agrupamento em 3 grupos.

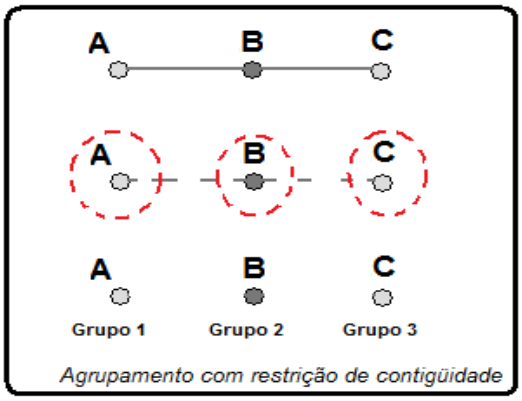


**Figura 3.2 – Esquema do agrupamento em 3 grupos**

No entanto, as regiões A e C não são semelhantes por acaso. Na região B há um túnel interligando as regiões A e C. Através desse túnel, as pessoas transitam entre essas duas regiões, de maneira que as economias das duas funcionam como se fossem uma única região. A informação da existência do túnel estava presente no mapa, mas não é utilizada no processamento do agrupamento.

Outro exemplo onde a restrição de contigüidade pode gerar resultados insatisfatórios está ilustrado no mapa da Figura 3.3. Nesse mapa, retirado do *googlemaps*[1], temos a região do estado do Rio de Janeiro com os municípios próximos a Baía de Guanabara.

[1] Para maiores informações acesse http://googlemaps.com.

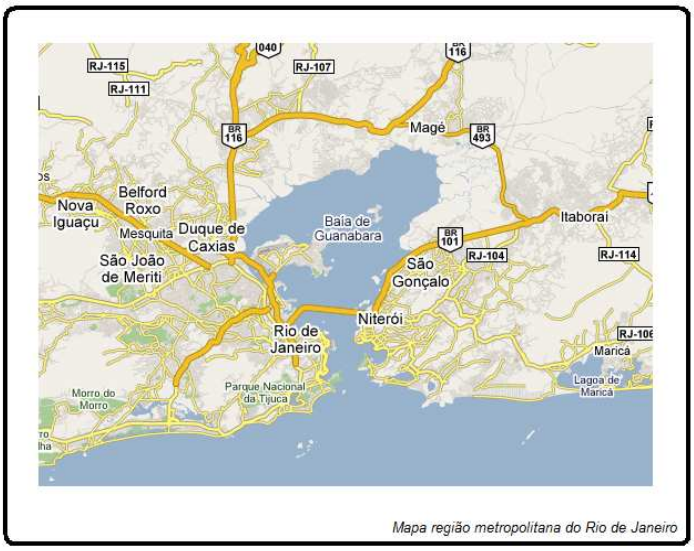


**Figura 3.3 – Mapa da região metropolitana do Rio de Janeiro**

Os municípios Rio de Janeiro, Duque de Caxias e Niterói possuem índices de desenvolvimento humano (IDH) altos (PNUD, 2008). Em contrapartida, os municípios Magé e São Gonçalo possuem IDH médios. No entanto, o método de agrupamento com restrição de contigüidade via AGM aloca os municípios Niterói e Rio de Janeiro em grupos diferentes, conforme ilustrado na Figura 3.4. Isso ocorre porque, embora exista uma ponte interligando Rio de Janeiro à Niterói, esses dois municípios não são vizinhos, ou seja, não geram um grupo contíguo.

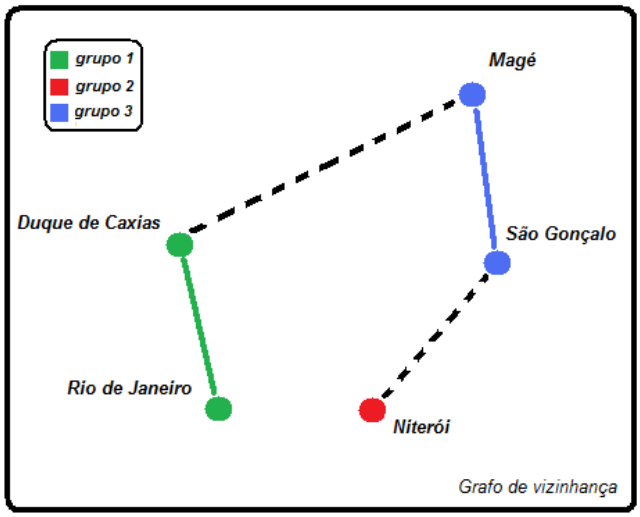


**Figura 3.4 – Esquema de agrupamento para os municípios próximos à Baía de Guanabara**

Nos dois exemplos apresentados, os resultados dos agrupamentos poderiam ter sido melhores. A informação sobre meios de acesso entre as regiões foram ignoradas e desperdiçadas pelo método de agrupamento com restrição de contigüidade via AGM. Em muitos casos, a informação dos meios de acesso entre regiões está disponível e pode exercer grande influência na formação de agrupamentos.

Nesse contexto, para melhorarmos a qualidade dos agrupamentos, é preciso relaxar a restrição de contigüidade sem que a relação espacial entre os objetos seja ignorada. Uma maneira de fazer isso é considerar o acesso entre regiões, independentemente destas serem ou não vizinhanças.

Neste capítulo, apresentamos uma proposta baseada no método de agrupamento com restrição de contigüidade via AGM. Esta consiste no método de agrupamento com restrição de acessibilidade via AMG cujo objetivo é considerar as informações dos acessos entre as regiões para agrupá-las.

## 3.2 – Acessibilidade e coeficiente de acessibilidade

A informação dos acessos entre regiões pode ser uma maneira de explorar relações entre regiões contidas em um mapa. Dependendo dos dados não-espaciais considerados para o agrupamento, uma topologia das relações de acesso entre as regiões pode agregar grande valor ao agrupamento.

Nesse sentido, precisamos de uma relação para construir uma topologia de relações de acesso entre regiões. Para isso, definimos acessibilidade ou relação de acessibilidade no contexto de banco de dados espaciais como:

***Acessibilidade (relação de acessibilidade) é a existência de um objeto espacial do tipo linha que cruza dois objetos espaciais do tipo polígono (ou região).***

A acessibilidade entre duas regiões pode ser representada através de um grafo ou matriz de acessibilidade. Nesse grafo, os nós representam as regiões e as arestas representam a existência de acessibilidade entre os nós. A matriz de acessibilidade possui *n* linhas e *n* colunas, onde *n* é o número de regiões. Cada linha e cada coluna da matriz representa uma região. As componentes da matriz com valor 1 indicam a existência de acessibilidade entre as regiões indicadas pela coluna e linha da componente. As componentes com valor 0 indicam a ausência de acessibilidade entre as regiões. A Figura 3.5 apresenta o exemplo das regiões A, B e C (seção 3.1) com seu grafo e matriz de acessibilidade.

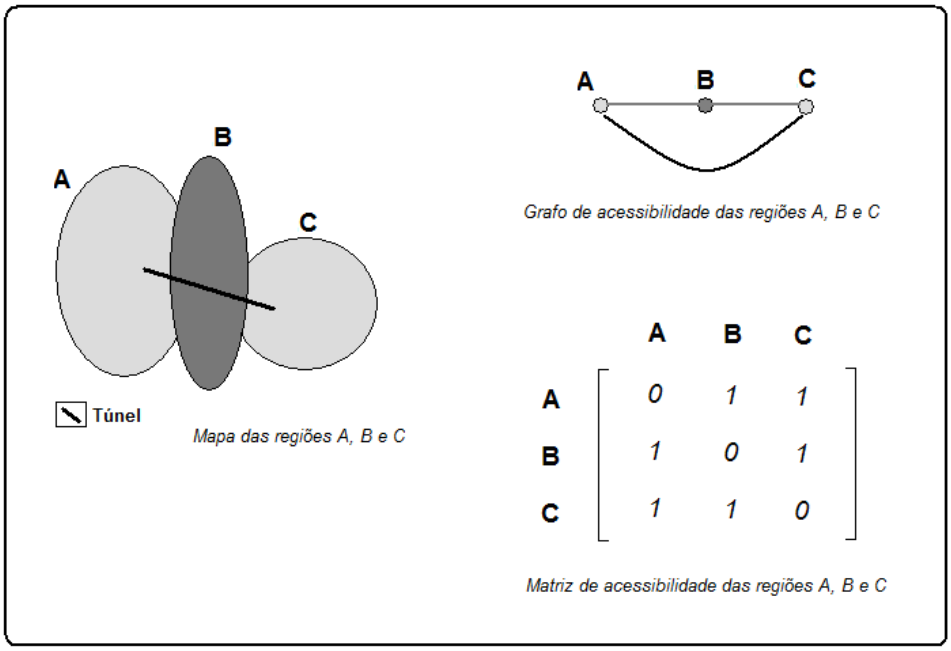


**Figura 3.5 – Exemplo das regiões A , B e C com seu grafo e matriz de acessibilidade**

No entanto, a acessibilidade não considera diferenças entre os objetos do tipo linha. Por exemplo, se temos uma rodovia em péssimo estado de conservação (pouco utilizada) e outra em excelente estado (muito utilizada), as duas devem ser consideradas com importâncias diferentes.

Portanto, para agregar mais informação sobre a acessibilidade entre as regiões, precisamos que pesos sejam atribuídos a essas relações. Esses pesos fazem com que as diferenças entre os meios de acesso entre regiões agreguem mais informação ao processo de agrupamento. Chamamos esses pesos de coeficientes de acessibilidade.

O ***coeficiente de acessibilidade*** é dado por um valor real no intervalo entre 0 e 1. Este é atribuído a cada relação de acessibilidade. Quanto mais próximo de 1 o *coeficiente de acessibilidade*, mais acessíveis entre si são as regiões. Por outro lado, quanto mais próximo de 0, menos acessíveis entre si são as regiões.

Apesar do coeficiente de acessibilidade agregar mais informação para as relações de acessibilidade, definir seus valores não é uma tarefa fácil. O coeficiente de acessibilidade pode ser definido através de dados não-espaciais para os objetos do tipo linha, através de funções que utilizam dados das relações espaciais entre as regiões ou, simplesmente, através de algum conhecimento *a priori* do problema.

O uso dos dados sobre fluxo de veículos em rodovias é um exemplo de definição de valores para os coeficientes de acessibilidade utilizando dados não-espaciais. Com os dados do fluxo de veículos podemos calcular os coeficientes de acessibilidade normalizando os valores dos fluxos. Dessa forma, as rodovias com maior fluxo de veículos terão coeficientes de acessibilidade mais próximos de 1, representando forte acessibilidade entre as regiões interligadas por elas. As rodovias com fluxos de veículos pequenos terão coeficientes de acessibilidade próximos de 0, portanto, pouca acessibilidade entre as regiões interligadas por elas.

Outro exemplo de coeficiente de acessibilidade é uma função que utiliza a distância entre duas regiões interligadas por uma rodovia. Definimos uma função que retorne coeficientes de acessibilidade mais próximos de 1 à medida que a distância entre as regiões é menor. Assim, quanto mais próximas duas regiões interligadas por uma rodovia, mais acessíveis entre si estas são. Por outro lado, à medida que a distância entre as regiões aumenta, a função retorna coeficientes de acessibilidade mais próximos de 0.

A escolha dos objetos do tipo linha que são utilizados para construir a matriz de acessibilidade e a definição dos valores dos coeficientes de acessibilidade devem ser feitas com cautela. Os objetos do tipo linha devem estar diretamente relacionados à natureza dos dados não-espaciais dos objetos de tipo região que queremos agrupar. Portanto, a escolha dos objetos do tipo linha e a definição do coeficiente de acessibilidade devem corresponder a elementos da realidade que agreguem informação coerente aos dados não-espaciais utilizados no agrupamento.

Por exemplo, se queremos agrupar regiões com atividades econômicas industriais semelhantes, utilizar rios não-navegáveis como meio de acesso entre as regiões não faz muito sentido. No entanto, se queremos agrupar regiões com o mesmo perfil de qualidade de água potável ou de atividade pesqueira nos rios, então os rios não-navegáveis fazem bastante sentido na conexão entre regiões semelhantes.

Portanto, a aplicação correta da acessibilidade e da definição de valores para o coeficiente de acessibilidade depende de algum conhecimento *a priori* do domínio do problema. O uso incorreto da acessibilidade ou de valores de coeficiente de acessibilidade pode conduzir os resultados a um agrupamento onde as relações espaciais entre os objetos não signifiquem nada. Assim, recomendamos cautela no seu uso, de forma que possamos melhorar os agrupamentos e mantermos as relações espaciais coerentes entre os objetos.

### 3.2.1 – Método de agrupamento de restrição de acessibilidade via AGM

Nesta seção descrevemos os detalhes do método de agrupamento com restrição de acessibilidade via AGM, proposto nesta dissertação. O método utiliza como entrada de dados as relações de acessibilidade entre regiões, os coeficientes de acessibilidade para cada ligação de acessibilidade e os dados não-espaciais das regiões a serem agrupados. A seguir descrevemos todos os passos executados pelo método.

#### 3.2.1.1 – Carga do grafo ou matriz de acessibilidade

O primeiro passo do algoritmo é a carga da matriz de acessibilidade. Nesse passo, é feita uma leitura das relações de acessibilidade dadas como entrada de dados. Estas são carregadas em uma matriz de acessibilidade $A_{n\times n}$, onde *n* é o número total de regiões a serem agrupadas. Cada ligação de acessibilidade entre duas regiões *i* e *j* é representada na matriz pelas componentes $a_{i,j} = a_{j,i} = 1$. O restante das componentes da matriz possui valor $a_{x,y} = a_{y,x} = 0$, indicando que não há acessibilidade entre as regiões *x* e *y*.

#### 3.2.1.2 – Geração da árvore geradora mínima

A matriz de acessibilidade carregada no passo anterior representa o grafo de acessibilidade. Nesse grafo, as arestas representam as acessibilidades entre as regiões. Para eliminarmos os ciclos de arestas entre os nós do grafo, precisamos de um critério para eliminar arestas. Para isso, estabelecemos pesos para cada aresta, de maneira que possamos eliminar as arestas de maior peso.

O peso é composto por uma medida de similaridade e outra de acessibilidade entre os nós conectados pelas arestas. A medida de similaridade adotada é a distância Euclidiana entre os vetores de dados dos nós. Os vetores de dados correspondem aos dados não-espaciais das regiões, nos quais os nós representam. Assim, quanto mais semelhantes são os dados não-espaciais, menor é medida de similaridade. A fórmula para o cálculo da medida de similaridade é dada por:

$$\sigma_{x,y} = \sqrt{(x_1 - y_1) + (x_2 - y_2) + \ldots + (x_i - y_i) + \ldots + (x_{n-1} - y_{n-1}) + (x_n - y_n)},$$

onde:

- $x_k$ é a componente *k* do vetor de dados do nó (região) *X*;

- $y_k$ é a componente *k* do vetor de dados do nó (região) Y; e

- $\sigma_{x,y}$ é a medida de similaridade entre os nós X e Y.

A medida de acessibilidade entre os nós é o valor do coeficiente de acessibilidade. Esses coeficientes são dados de entrada para o método, onde cada aresta possui um coeficiente de acessibilidade associado, pertencente ao intervalo entre 0 e 1.

Dessa maneira, o valor do peso para cada aresta é calculado através da seguinte fórmula:

$$p_{x,y} = \sigma_{x,y} \times (2 - \alpha_{x,y}),$$

onde:

- $p_{x,y}$ é o peso da aresta que conecta o nó *x* ao nó *y*;

- $\sigma_{x,y}$ é a medida de similaridade entre o nó *x* e o nó *y*; e

- $\alpha_{x,y}$ é o coeficiente de acessibilidade (ou medida de acessibilidade) entre o nó *x* e o nó *y*.

Na fórmula de $p_{x,y}$, utilizamos a função $2 - \alpha_{x,y}$, pois queremos que quando $\alpha_{x,y}$ tender a um ou for um, $p_{x,y}$ se aproxime de $\sigma_{x,y}$. Por outro lado, à medida que $\alpha_{x,y}$ se aproxima de zero, $p_{x,y}$ aumenta linearmente.

A partir do grafo de acessibilidade com os pesos de suas arestas calculados, executamos o algoritmo PRIM (CORMEN *et al.*, 2001).

### 3.2.1.3 – Poda da AGM

O procedimento de poda da AGM no método de agrupamento com restrição de acessibilidade é o mesmo do método de agrupamento com restrição de contigüidade (maiores detalhes podem ser encontrados na seção 2.8.2 do capítulo 2 desta dissertação). O resultado da poda da AGM é uma floresta. Cada componente conexa, isto é, cada árvore da floresta é um grupo de regiões semelhantes e acessíveis entre si.

## 3.3 – Complexidade do método

O método de agrupamento com restrição de acessibilidade via AGM possui a mesma complexidade do método de agrupamento com restrição de contigüidade via AGM. Conforme apresentado, o algoritmo de construção e poda da AGM é o mesmo para os dois métodos. O que muda são os pesos das arestas e a topologia das regiões representadas no grafo. Geralmente, no grafo de acessibilidade encontramos uma grande quantidade de arestas. Devido a isso, é possível que ocorra um pequeno aumento no tempo de execução.

## 3.4 – Considerações finais

Neste capítulo descrevemos a necessidade de uma abordagem de agrupamento de regiões utilizando informações de acessibilidade. Além isso, apresentamos a proposta de um método que considera a acessibilidade entre regiões através de objetos do tipo linha, chamado de método de agrupamento com restrição de acessibilidade via AGM.

O método de agrupamento com restrição de acessibilidade via AGM utiliza a topologia de acessibilidade entre regiões como restrição para a formação de grupos. Dessa forma, os objetos de um mesmo grupo são semelhantes e acessíveis entre si.

A topologia de acessibilidade faz com que a chance de encontrarmos um resultado melhor que o método original aumente. Isso ocorre porque quanto mais arestas existirem no grafo, maiores são as possibilidades de alocação dos objetos nos grupos. Um estudo comparativo da qualidade dos resultados entre os dois métodos de agrupamento encontra-se no capítulo 4 desta dissertação.

Embora o uso da acessibilidade não se aplique a todos os casos de agrupamento de regiões, esperamos que seu uso permita a identificação de grupos mais homogêneos para os casos onde a acessibilidade é importante.

# Capítulo 4 – Avaliação da abordagem de agrupamento de regiões utilizando acessibilidade

## 4.1 – Introdução

Neste capítulo apresentamos a avaliação da abordagem de agrupamento proposta nesta dissertação. Essa avaliação consistiu em uma análise comparativa da qualidade dos resultados obtidos pelo método de agrupamento com restrição de acessibilidade via AGM e pelo método de agrupamento com restrição de contigüidade via AGM. Para isso, realizamos um experimento que consistiu em um estudo de caso, onde o objetivo foi utilizar os dois métodos de agrupamento em avaliação para classificar regiões de acordo com seus dados não-espaciais e suas relações topológicas.

O motivo para realizarmos essa comparação entre esses dois métodos se deve ao fato de que o método de agrupamento com restrição de acessibilidade é uma modificação do método de agrupamento com restrição de contigüidade via AGM. Portanto, essa comparação nos ajuda a avaliar se a abordagem de utilizar acessibilidade melhora ou não a qualidade dos grupos encontrados.

Além disso, NEVES (2003) apresenta uma análise comparativa entre o método de agrupamento com restrição de contigüidade via AGM e outros quatro métodos de agrupamento. Nessa análise, o método de agrupamento com restrição de contigüidade via AGM produziu resultados de melhor qualidade que os outros métodos. Portanto, encontrar resultados de melhor qualidade com a abordagem que utiliza acessibilidade significa que estamos superando também outros métodos de agrupamento.

## 4.2 – Ambiente de experimentação

A avaliação do método de agrupamento com restrição de acessibilidade via AGM foi realizada em um ambiente de experimentação. Esse ambiente consiste de um protótipo em linguagem Java onde os dois métodos a serem comparados foram implementados (MELLO *et al.* , 2007). Além desse protótipo, o ambiente também conta com o software Data Desk 6.0 (DATADESK, 2008), a planilha eletrônica Microsoft Excel, o servidor de mapas MapServer (MAPSERVER, 2008) e o banco de dados PostGreSql e sua extensão para dados espaciais PostGis (POSTGRESQL, 2008).

O protótipo que implementa os dois métodos de agrupamento se comunica com o banco de dados PostGreSql, busca os objetos, executa o algoritmo de agrupamento e grava o identificador do grupo para cada objeto no banco de dados. Assim, com uma simples consulta no PostGreSql é possível saber quais objetos estão alocados em quais grupos. O programa também é responsável por imprimir algumas medidas que são utilizadas para análise, descritas na seção 4.3.

O software Data Desk 6.0 é um ambiente estatístico onde é possível visualizar gráficos, calcular medidas estatísticas e manipular variáveis aleatórias (DATADESK, 2008). O Microsoft Excel também é utilizado para tratar e analisar os dados.

A visualização dos mapas nos experimentos é feita através de uma aplicação no MapServer, vide Figura 4.1. Essa aplicação permite visualizar o mapa com suas regiões coloridas de acordo com os grupos encontrados. Cada cor está associada a um grupo, portanto, regiões de mesma cor pertencem a um mesmo grupo.

Com esse ambiente foi possível tratar, analisar e processar os dados espaciais do experimento de maneira rápida e automatizada.

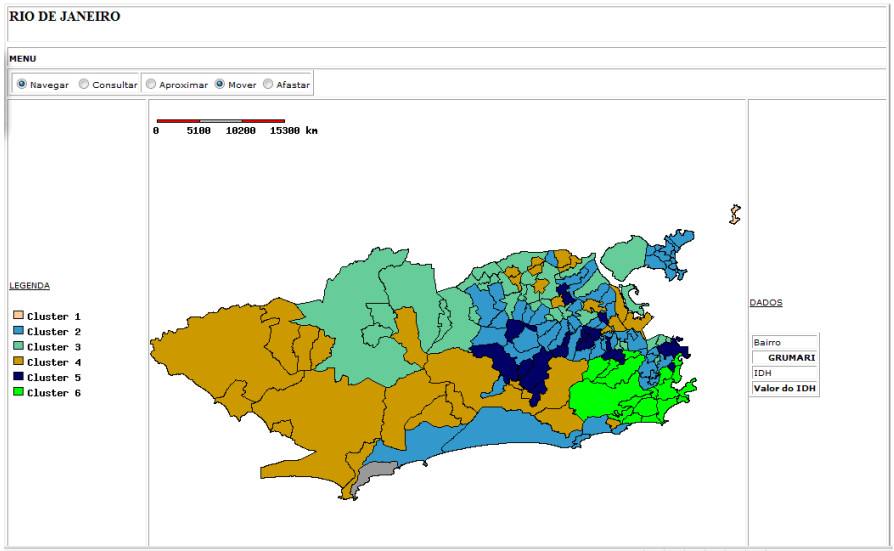


**Figura 4.1 - Aplicação no MapServer para visualização do grupos**

## 4.3 – Medidas de comparação

Nesta seção descrevemos as medidas adotadas para comparar e analisar os resultados obtidos no experimento.

Comparar a qualidade dos resultados entre métodos de agrupamento é uma tarefa difícil e existem várias maneiras de realizá-la. Nesta avaliação adotamos duas medidas para a comparação dos resultados.

### 4.3.1 – Soma das variações intragrupos

Segundo ALDENDERFER & BLASHFIELD (1984), a análise de agrupamentos tem como objetivo identificar grupos homogêneos minimizando a soma das diferenças dentro dos grupos – *intragrupos* – e maximizando a soma das diferenças entre grupos – *intergrupos*.

A soma das diferenças *intergrupos* é dada pela soma das distâncias dos centros de cada grupo. Esta não é uma boa medida de qualidade para comparar os dois métodos de agrupamento em questão. Esta medida parte do princípio que quanto mais distantes estão os centros dos grupos, melhor é o agrupamento. No entanto, nos dois métodos que estamos avaliando é possível que objetos semelhantes formem dois grupos com centros próximos ou até mesmo iguais. Isso ocorre devido à restrição espacial existente nos dados. A Figura 4.2 ilustra esse exemplo.

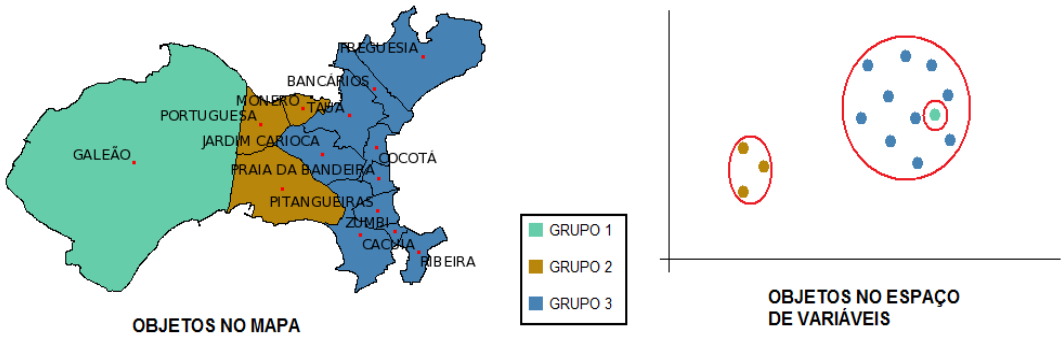


**Figura 4.2 - Exemplo de variações intragrupos**

Nessa figura podemos observar que os objetos do grupo 1 e do grupo 3 estão separados pela restrição geográfica de vizinhança. No entanto, no espaço de variáveis, os objetos formados por esses dois grupos estão muito próximos, fazendo com que a distância entre o centro do grupo 1 e do grupo 3 seja muito pequena. Então, devido a restrição espacial, a soma das diferenças *intergrupos* não representa um indicador de qualidade eficiente para o agrupamento.

Por outro lado, a soma das diferenças *intragrupos* é uma boa medida para avaliar a homogeneidade dos grupos formados e é amplamente utilizada na área de análise de agrupamentos. Essa medida avalia a posição dos objetos no espaço de variáveis dentro dos seus respectivos grupos. Além disso, o critério de agrupamento utilizado pelos dois métodos comparados é baseado nessa medida.

Portanto, um índice de qualidade de agrupamento baseado na soma das diferenças *intragrupos*, é dado por:

$$Q(\prod) = \sum_{i=0}^{k} SD ,$$

onde:

- $Q(\prod)$ é valor do índice de qualidade da partição;
- $\prod$ é a alocação de $n$ objetos em $k$ grupos;
- $SD = \sum_{j=0}^{m} \left\| x_j - \overline{X_i} \right\|$ é soma das distâncias (desvios) dos $m$ elementos do grupo $i$ em relação ao centro $\overline{X_i}$ (média) deste.

Portanto, ao executarmos os dois métodos com o mesmo número de grupos, o que obtiver o menor valor para $Q(\prod)$ será o que minimizou melhor as variações *intragrupos* e, conseqüentemente, o de melhor qualidade.

### 4.3.2 – Índice PBM

Além do problema de comparar a qualidade dos grupos entre os dois métodos, outro fator que afeta a qualidade do agrupamento é o número de grupos que se deseja encontrar. Isto é, descobrir em quantas classes os dados estão melhor distribuídos. Em nossos experimentos, não era conhecido o número de grupos que resulta na melhor estrutura de agrupamento para os dados. Portanto, foi necessário adotar uma medida para nos auxiliar nessa tarefa.

Existem várias medidas que podem ser adotadas para encontrar o número de grupos que resulte na melhor partição dos dados (XIE & BENI, 1991; BEZDEK & PAL, 1998; PAKHIRAA *et al*., 2004). A medida adotada foi o índice *PBM* (PAKHIRAA *et al*., 2004).

Dado um agrupamento, o índice *PBM* resulta em um valor de qualidade do resultado para os grupos formados. Quanto maior o valor obtido pelo índice, melhor será a qualidade dos resultados. O valor do índice *PBM* é dado por:

$$PBM(K) = \left( \frac{1}{K} \times \frac{E_1}{E_K} \times D_K \right)^2,$$

onde:

- $k$ é o número de grupos (ou classes);

- $E_K = \sum_{i=1}^{K} E_i$ ;

- $E_i = \sum_{j=1}^{n} u_{jk} \left\| x_j - z_k \right\|$;

- $n$ é o número total de objetos a serem agrupados;

- $u_{jk}$ é o elemento da matriz de partição dos dados $U(X) = \lfloor u_{jk} \rfloor_{n \times K}$;

- $x_j$ é o $j^{ésimo}$ objeto pertencente ao $k^{ésimo}$ grupo;

- $z_k$ é o centro do $k^{ésimo}$ grupo;

- $D_K = \max_{i,j=1}^{K} \left\| z_i - z_j \right\|$.

Portanto, o índice *PBM* é composto por três fatores: $\frac{1}{K}$, $\frac{E_1}{E_K}$ e $D_K$. O primeiro fator reduz o índice à medida que o número de grupos é aumentado. O segundo fator é composto pelo numerador $E_1$ e pelo denominador $E_K$. Este último é a soma dos desvios da posição de cada objeto no espaço de variáveis ao centro de seu respectivo grupo. O valor de $E_1$ é constante, sendo a soma dos desvios para todos os objetos alocados em uma única partição. Dessa maneira, $\frac{E_1}{E_K}$ é diretamente proporcional à homogeneidade dos grupos formados. Então, quanto menor $E_K$, maior a homogeneidade dos grupos e maior é o valor do índice. Por último, o fator $D_K$ é a distância máxima entre o centro de dois dos $k$ grupos formados. Quanto maior a distância entre os grupos, maior é o índice de qualidade.

Para descobrirmos o número de grupos que maximiza a qualidade dos resultados, executamos o método de agrupamento para $K = 1,2,3,4,...$ grupos. Em seguida, calculamos os valores do índice $PBM(K)$ para cada valor de *K* e realizamos uma análise para escolhermos em quantos grupos devemos dividir os dados.

O índice *PBM* além de servir para encontrar o número de grupos também pode ser utilizado para comparar a qualidade dos resultados entre os dois métodos. Essa

comparação é realizada da mesma forma que utilizamos para encontrar o número de grupos em que devemos dividir os objetos. Calculamos o índice $PBM(K)$ para os dois métodos e comparamos os valores. Embora $K$ seja o mesmo nos dois resultados, os objetos podem estar alocados em grupos diferentes, fazendo com que $E_K$ e $D_K$ venham a ser diferentes. Assim, o método que obtiver o maior índice será o que formou agrupamentos de melhor qualidade.

## 4.4 – Metodologia

Nesta seção apresentamos a metodologia que definimos para realizar o experimento.

Para evitar equívocos e podermos repetir o experimento, optamos por definir e utilizar uma metodologia (ou processo) para realizá-lo. O objetivo dessa metodologia é conduzir o experimento de maneira que os resultados sejam confiáveis. Os passos da metodologia são:

- carga dos dados;
- análise exploratória dos dados;
- normalização de variáveis;
- execução do método de agrupamento;
- escolha do número de grupos;
- análise de agrupamento.

As próximas subseções descrevem cada passo dessa metodologia.

### 4.4.1 – Carga de dados

Os dados espaciais e os dados convencionais utilizados no estudo de caso foram obtidos de bases de dados externas. Estes, para serem utilizados nos experimentos, foram importados para o banco de dados PostGreSql com a extensão PostGis (POSTGRESQL, 2008). O PostGreSql é um banco de dados relacional que possui uma extensão, chamada PostGis, que suporta o gerenciamento de dados espaciais. O objetivo da carga dos dados é utilizar o suporte à consulta espacial do PostGis para auxiliar a construção dos grafos (ou matrizes) de vizinhança e de acessibilidade utilizados pelos métodos de agrupamento.

### 4.4.2 – Análise exploratória dos dados

A análise exploratória dos dados é realizada, principalmente, através da utilização de gráficos. Esta foi auxiliada através do software Data Desk 6.0 (DATADESK, 2008). Nosso objetivo nessa análise é identificar variáveis redundantes, valores extremos (*outliers*) e anormalidades nos dados que possam prejudicar os resultados do agrupamento.

Para realizar a análise dos dados, foram utilizadas técnicas de:

- análise univariada:
    - histograma (HAN & KAMBER, 2006);
    - diagrama de caixa ou *boxplot* (FRIGGE *et al.*, 1989);
    - sumário de medidas de tendência central e variabilidade (HAN & KAMBER, 2006);
- análise bivariada:
    - diagrama de espalhamento (HAN & KAMBER, 2006);
    - matriz de correlação (HAN & KAMBER, 2006);
- análise espacial;
    - visualização da distribuição das variáveis no mapa.

Os gráficos e medidas adotados têm como objetivo nos ajudar no entendimento da natureza dos dados utilizados como entrada para os algoritmos de agrupamento.

### 4.4.3 – Normalização

Nesta etapa, as variáveis envolvidas na análise de agrupamentos são normalizadas. A normalização tem como objetivo colocar os valores das variáveis em um intervalo, tipicamente, entre zero e um (SHALABI *et al.*, 2006, HAN & KAMBER, 2006). Isto faz com que medidas de distâncias utilizadas nos métodos de agrupamento possam estar na mesma escala. Dessa maneira, as variáveis envolvidas na análise de agrupamento possuem o mesmo peso, evitando tendências na análise.

### 4.4.4 – Execução do método de agrupamento

Neste passo, os dados normalizados, a matriz de vizinhança e a matriz de acessibilidade são buscados no banco de dados ou em arquivos de entrada de dados pelo programa que implementa os dois métodos de agrupamento. Os dois métodos são executados para o número de grupos $K = 1,2,3,4,\ldots$. As partições dos grupos e os valores calculados de $PBM(K)$ e $Q(\prod)$ são armazenados em uma tabela. O limite para o valor de $K$ é estabelecido a partir do problema em questão.

### 4.4.5 – Escolha do número de grupos

A escolha do número de grupos que resulta na melhor qualidade para os dois métodos é baseada na tabela gerada para os valores de $PBM(K)$ e $Q(\prod)$. Os valores de *PBM* e $Q(\prod)$ são analisados para cada *K* e, então, é decidido o número de grupos que queremos.

### 4.4.6 – Análise de agrupamento

Nesta etapa, analisamos os grupos gerados pelo método que oferecer resultados de melhor qualidade, isto é, $Q(\prod)$ menores e $PBM(K)$ maiores. O objetivo deste passo é elaborar rótulos para os grupos de maneira que a classificação das regiões se torne um instrumento para análise destas, de acordo com o estudo de caso.

Na análise de agrupamento também utilizamos gráficos de *bloxplot* e um sumário de medidas de tendência central e de variações.

## 4.5 – Estudo de caso – Distribuição espacial do IDH dos bairros da cidade do Rio de Janeiro

O Índice de Desenvolvimento Humano (IDH) é uma medida comparativa da riqueza, alfabetização, educação, esperança de vida, natalidade e outros fatores (PNUD, 2008). O objetivo do IDH é estabelecer uma maneira padronizada de avaliar e medir o bem-estar de uma população, especialmente a infantil. Essa medida vem sendo utilizada pela ONU (Organização das Nações Unidas) desde 1993 em seu relatório anual de programa de desenvolvimento (PNUD, 2008).

O IDH é a média aritmética de três medidas que consideram educação (E), longevidade (L) e renda (R). Estas também são chamadas de IDH-E, IDH-L e IDH-R, respectivamente (PNUD, 2008).

Portanto, para calcularmos o valor do IDH médio, temos a seguinte fórmula:

- $IDH = \frac{L + E + R}{3}$;
- onde:
  - $L = \frac{EV - 25}{60}$;
    - $EV =$ expectativa de vida,
  - $E = \frac{2TA + TE}{3}$,
    - $TA =$ taxa de alfabetização,
    - $TE =$ taxa de escolarização,
  - $R = \frac{\log_{10} PIBpc - 2}{2.60206}$, onde *PIBpc* é a renda per capita.

O índice varia em um intervalo real de 0 (nenhum desenvolvimento humano) até 1 (total desenvolvimento humano), sendo sua classificação considerada da seguinte forma:

- valores entre 0 e 0.499, são considerados baixos;
- valores entre 0.5 e 0.799, são considerados médios; e
- valores entre 0.8 e 1, são considerados altos.

**Problema**

Na administração pública, muitas vezes é necessário obter uma visão macro dos dados para que políticas públicas possam ser estabelecidas. Nesse contexto, identificar grupos de bairros na cidade do Rio de Janeiro que estão em situação de desenvolvimento humano semelhante é fundamental. Com isso, a administração da cidade pode atuar nesses bairros de acordo com suas características e necessidades, traçar políticas públicas semelhantes para bairros de mesmo grupo *etc*.

**Objetivo**

Neste estudo de caso, vamos realizar uma análise dos índices que compõem desenvolvimento humano dos bairros do Rio de Janeiro. Além disso, vamos agrupar e classificar os bairros, de maneira que seja possível ter uma visão geral do desenvolvimento humano da longevidade, educação e renda nos bairros da cidade.

**Justificativa da acessibilidade**

Como já mencionamos, o IDH é um índice que mede o desenvolvimento humano em uma região. No entanto, as pessoas podem se locomover para outras regiões ou bairros para utilizar serviços que não estão disponíveis em suas regiões. Por exemplo, quando um determinado bairro não tem um hospital público é possível que os moradores busquem esse serviço em outros bairros. Portanto, mesmo que um bairro não tenha um determinado serviço, não quer dizer que necessariamente as pessoas daquela região não usufruem disso em outros bairros.

Nesse contexto, acreditamos que uma análise de agrupamento dos bairros do Rio de Janeiro utilizando a informação espacial de acessibilidade entre os bairros através de rodovias possa trazer resultados interessantes.

### 4.5.1 – Carga dos dados

Os dados utilizados são referentes ao IDH do ano de 2000 e foram obtidos do Instituto Brasileiro de Geografia e Estatística (IBGE).

Os dados estão organizados em uma tabela com as seguintes colunas:

- bairro ou grupo de bairros;
- esperança de vida ao nascer (em anos);
- taxa de alfabetização de adultos (%);
- taxa bruta de freqüência escolar (%);
- renda per capita (em R$);
- índice de Longevidade – IDH-L;
- índice de Educação – IDH-E;
- índice de Renda – IDH-R; e
- índice de Desenvolvimento Humano – IDH.

Nessa tabela os dados referentes ao IDH podem corresponder a mais de um bairro. Há regiões que estão agrupadas em 2 ou 3 bairros. Por exemplo, Vidigal e São Conrado estão agrupados, portanto, os valores referentes ao IDH de Vidigal e de São Conrado são os mesmos, gerando um pouco de distorção.

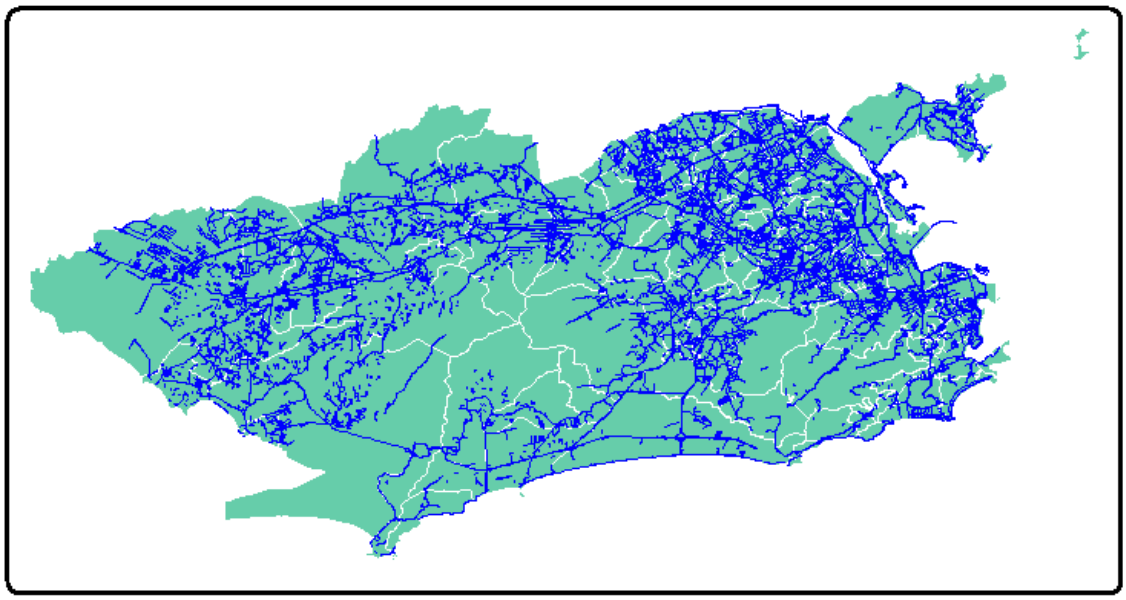

**Figura 4.3 – Mapa dos limites dos bairros e rodovias da cidade do Rio de Janeiro**

A Figura 4.3 apresenta os dados geográficos utilizados. Nessa figura temos o mapa dos limites dos 159 bairros do rio de janeiro e de 18168 rodovias. Para realizar a associação entre os bairros no mapa com as linhas da tabela do IDH, foi necessário desmembrar os grupos de bairros. Com isso, por exemplo, São Conrado e Vidigal ganharam cada um uma linha na tabela do IDH.

Os dados referentes ao IDH e os dados espaciais dos bairros e ruas foram importados para o banco de dados PostGreSql. Dessa maneira, podemos construir as matrizes de vizinhança e acessibilidade através de consultas espaciais no PostGis.

### 4.5.2 – Análise Exploratória dos dados e normalização

Na análise exploratória dos dados estamos interessados nos dados das colunas IDH-L (longevidade), IDH-E (educação) e IDH-R (renda) para cada bairro do Rio de Janeiro. Portanto, definimos as variáveis aleatórias IDH-L, IDH-E e IDH-R para análise. Note que não estamos interessados no valor do IDH médio, mas sim nos índices que o compõem. Assim, podemos obter uma avaliação de cada um dos aspectos considerados na formação do desenvolvimento humano.

### 4.5.2.1 – IDH-L

O IDH-L está diretamente associado à longevidade da população presente nos bairros. O sumário das medidas estatísticas, o histograma, o *boxplot* e o gráfico de probabilidade normal são apresentados na Figura 4.4.

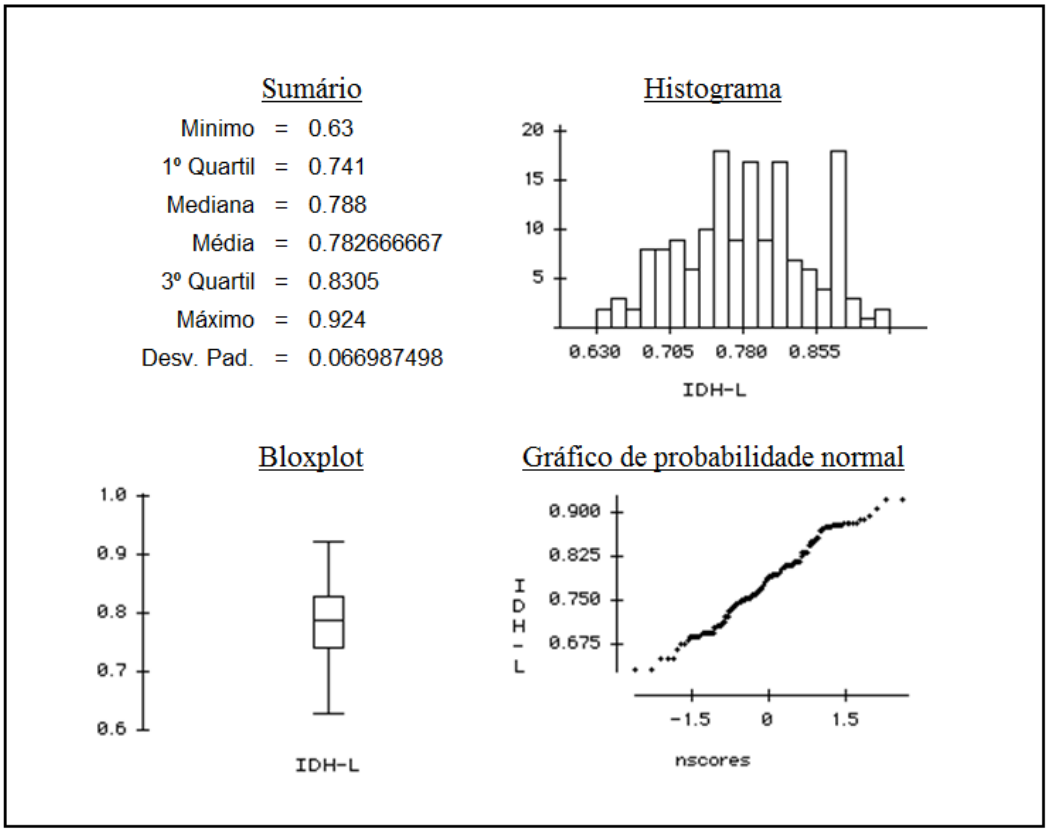


**Figura 4.4 - Gráficos e medidas do IDH-L**

O histograma dessa variável indica um formato aproximado ao da distribuição normal. Analisando o gráfico de probabilidade normal, vemos que os pontos tendem a uma reta e no *boxplot* constatamos um formato simétrico. Sendo assim, podemos afirmar que a variável tende a uma distribuição normal com média 0.78 e desvio padrão 0.066.

Portanto, o valor do IDH-L apresenta uma média de longevidade considerada mediana. Além disso, os valores de máximo e mínimo, 0.924 e 0.63, respectivamente, constatam que há grande discrepância em relação à longevidade. De acordo com o valor do 3º quartil, cerca de 25% dos bairros do Rio de Janeiro possuem o IDH-L maior que 0.83, isto é, com IDH-L considerado alto. A Figura 4.5 apresenta um mapa com a distribuição dos bairros de acordo com os intervalos interquartis. Na zona sul da cidade,

estão concentrados os bairros de maior índice de longevidade. Em parte da zona oeste e da zona norte estão os bairros de índice de longevidade mais baixos.

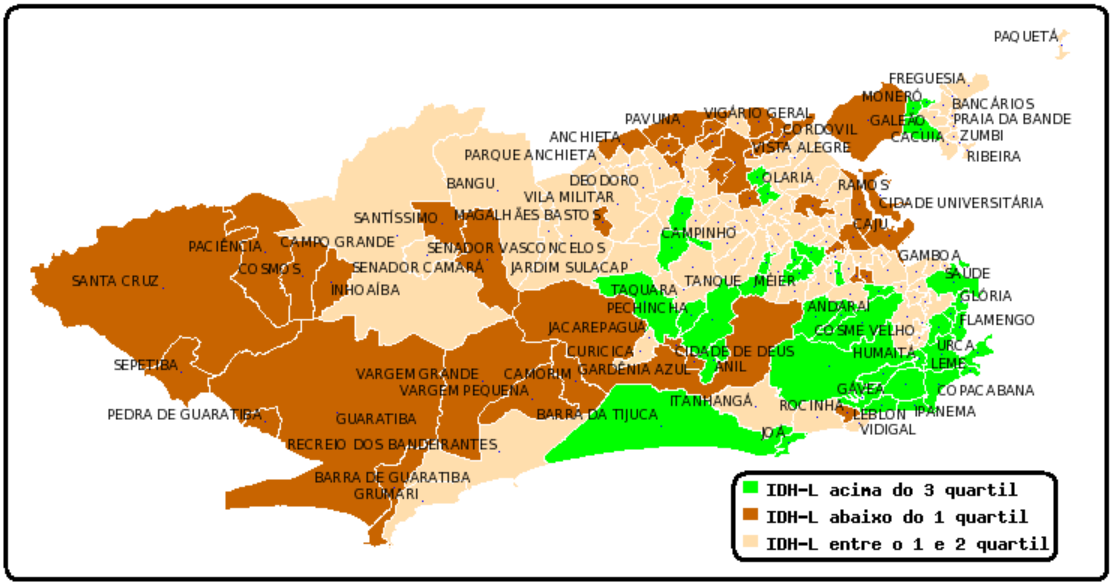


**Figura 4.5 - Mapa da distribuição do IDH-L**

## 4.5.2.2 – IDH-E

A Figura 4.6 apresenta o sumário das medidas estatísticas, o histograma, o *bloxplot* e o gráfico de probabilidade da variável IDH-E. O IDH-E indica o desenvolvimento humano voltado para a educação. Esse índice é formado pela taxa de alfabetização e pela taxa de freqüência escolar.

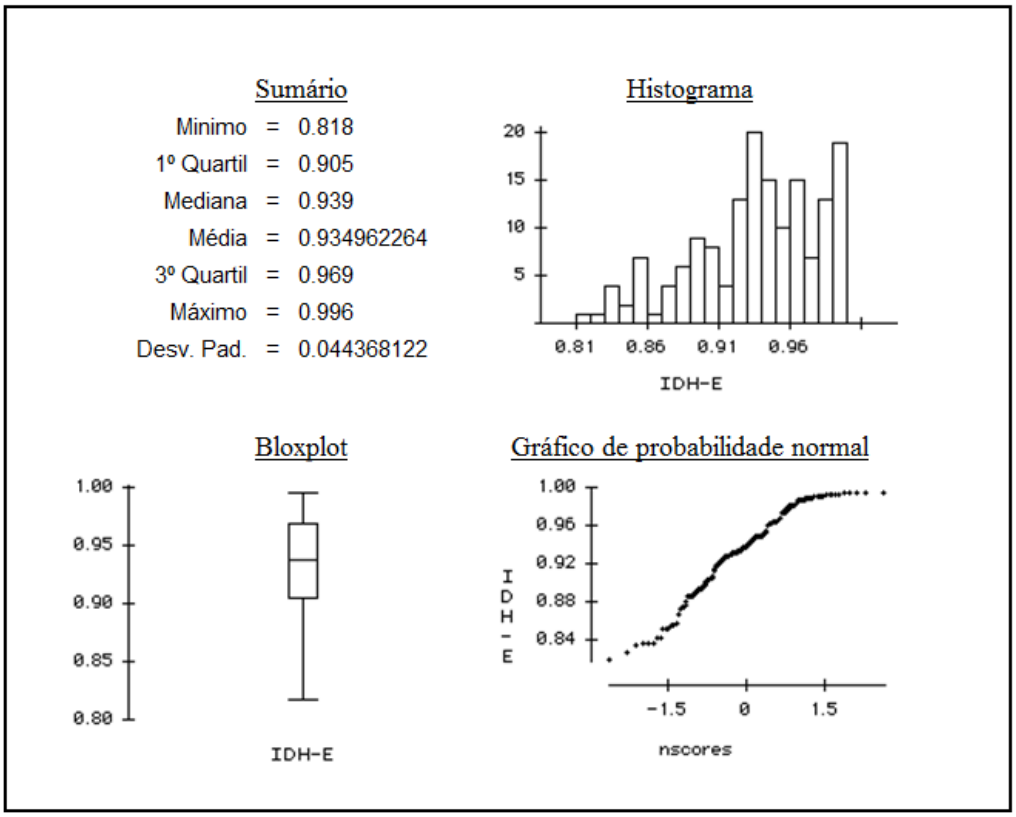


**Figura 4.6 - Gráficos e medidas do IDH-E**

O gráfico de probabilidade normal e histograma indicam que a variável IDH-E segue uma distribuição normal de média 0.93 e desvio padrão 0.044. O *boxplot* e o histograma mostram claramente que há menor variação do índice nos bairros com IDH-E mais alto que a mediana. Entretanto, no *boxplot*, a linha do limite inferior é maior, o que indica que há bairros com valores de IDH-E bem menores que o 1º quartil.

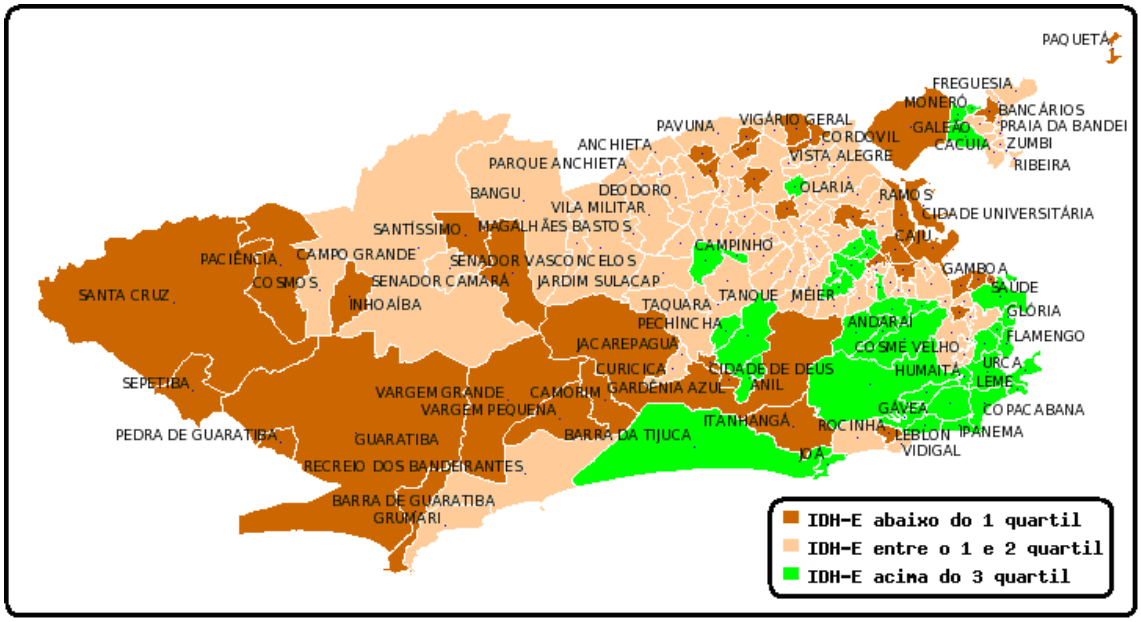


**Figura 4.7 - Mapa da distribuição do IDH-E**

No mapa dos bairros apresentado na Figura 4.7, podemos notar que a distribuição do IDH-E é semelhante a do IDH-L. Boa parte da zona oeste conta com os piores índices de IDH-E. Na zona sul encontra-se a maior concentração de bairros com o IDH-E mais altos.

Embora existam discrepâncias nos valores do IDH-E entre os bairros, o desenvolvimento humano educacional nos bairros do Rio de Janeiro é classificado como alto. O valor mínimo do IDH-E é 0.818, considerado alto. Há menos desigualdade educacional nos bairros com IDH-E maior que a mediana e maior desigualdade nos bairros com IDH-E menores que a mediana.

### 4.5.2.3 – IDH-R

O IDH-R representa uma medida da distribuição de renda da população carioca. O sumário das medidas estatísticas, o histograma, o *boxplot* e o gráfico de probabilidade normal são apresentados na Figura 4.8.

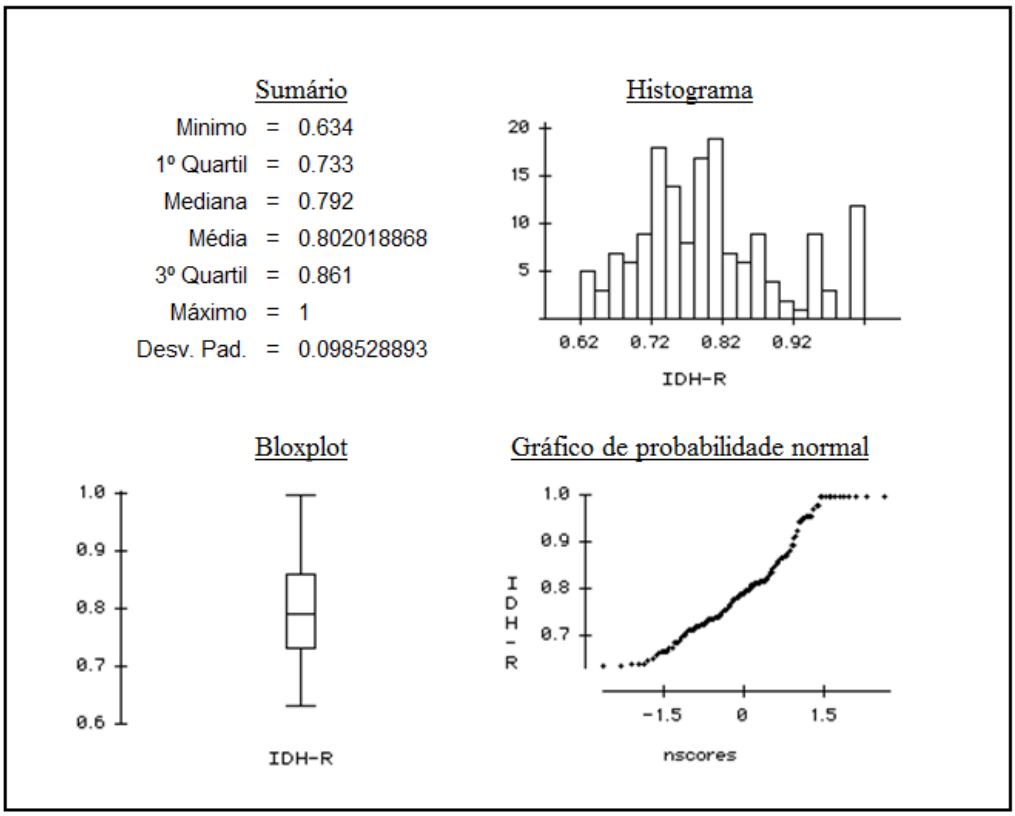


**Figura 4.8 - Gráficos e medidas do IDH-R**

O gráfico de probabilidade normal, o *boxplot* e o histograma sugerem que o IDH-R obedece a uma distribuição normal. Entretanto, podemos constatar pela mediana que mais de 50% dos bairros possuem índices menores que a média. Além disso, no histograma, podemos notar que há poucos bairros com IDH-R muito altos. O limite superior do *boxplot* confirma isso.

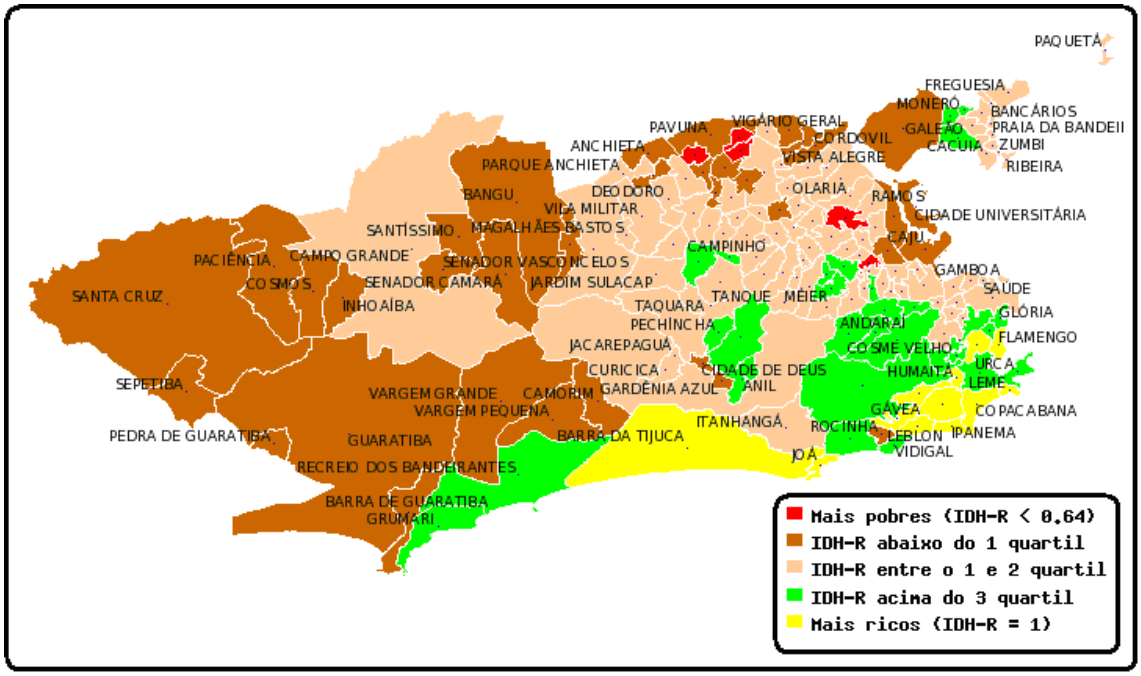


**Figura 4.9 - Mapa da distribuição do IDH-R**

Na Figura 4.9, o mapa indica que os bairros mais ricos estão concentrados na zona sul e Barra da Tijuca. Por outro lado, os bairros mais pobres se encontram na zona norte, nos limites da cidade com a baixada fluminense.

O índice de desenvolvimento humano de renda indica que, no Rio de Janeiro, há bairros onde a concentração de renda é bem maior que em todos os outros. Além disso, também há bairros onde o IDH-R é bem mais baixo. Outro ponto importante na distribuição do índice, é que 50% dos bairros estão classificados de acordo com IDH-R como médios e os restantes como altos.

### 4.5.2.4 – Análise conjunta das variáveis

Nesta seção, realizamos uma análise comparativa das distribuições das três variáveis: IDH-L, IDH-E e IDH-R. Essa análise comparativa é possível porque as três variáveis estão dentro da mesma escala de valores.

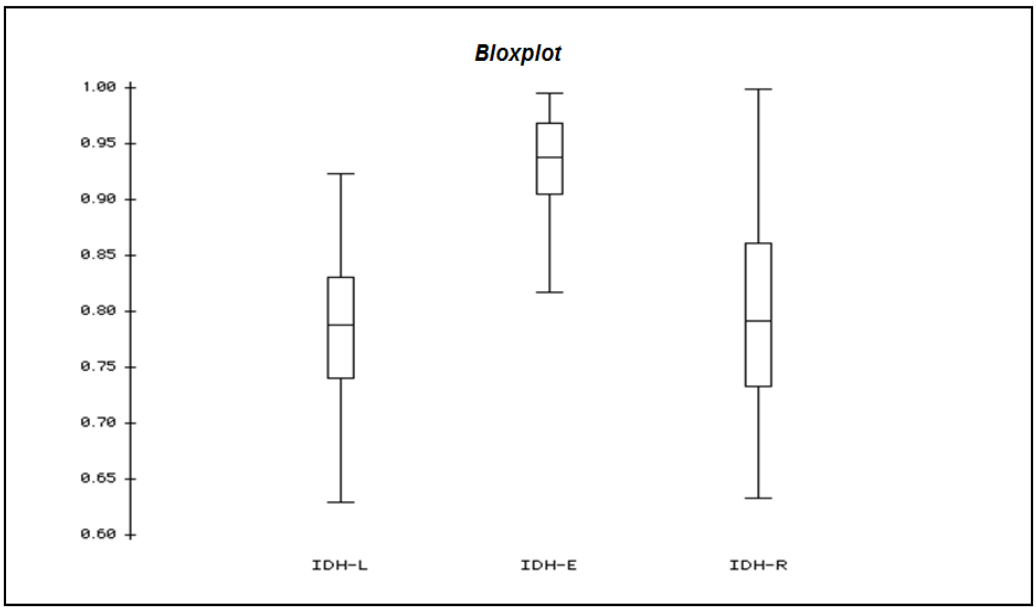


**Figura 4.10 - BoxPlot IDH-L, IDH-E e IDH-R**

Na Figura 4.10, temos o *boxplot* das três variáveis: IDH-L, IDH-E e IDH-R. Podemos notar que o índice de renda (IDH-R) é o índice que apresenta maior desigualdade. Nesse índice, os limites, superior e inferior, estão mais distantes do 3º e 1º quartil, respectivamente. Isso indica que há poucos bairros com renda muito maior que o restante e poucos com renda muito menor. A caixa do *boxplot* está mais deslocada para baixo, o que distancia ainda mais os bairros mais ricos do restante dos bairros. O índice de desenvolvimento de longevidade (IDH-L) acompanha a desigualdade social do IDH-R, embora suas distorções sejam um pouco menores. O índice de educação (IDH-E) é o índice que apresenta menor desigualdade entre os bairros cariocas e ainda possui todos os seus valores acima de 0.8, isto é, valores considerados altos.

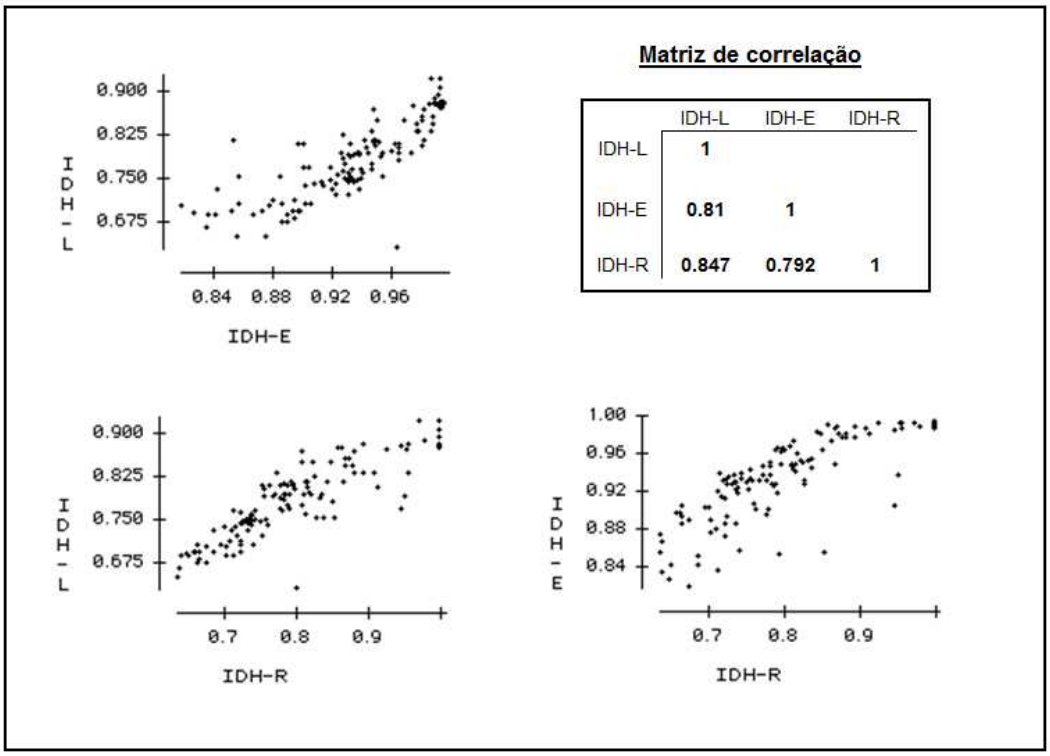


**Figura 4.11 - Gráficos e medidas de dispersão par a par do IDH-L, IDH-E e IDH-R**

A Figura 4.11 apresenta os gráficos de dispersão par a par das variáveis em análise e também a matriz de correlação entre estas. De acordo com a matriz de correlação apresentada, as variáveis são positivamente correlatas. Isto é, à medida que um dos índices aumenta, os outros tendem a aumentar também. Essa correlação pode ser vista nos gráficos de dispersão. Notamos pelo gráfico de dispersão IDH-L x IDH-R e pelo valor da correlação entre essas variáveis que estas são fortemente relacionadas, o que justifica seus *boxplots* semelhantes. Embora o índice de desenvolvimento educacional tenha alto valor de correlação com as outras variáveis, no gráfico de dispersão há bairros que não acompanham essa tendência, possuindo IDH-E baixo e IDH-R alto.

Apesar das três variáveis estarem fortemente correlacionadas entre si, vamos utilizar todas elas no algoritmo de agrupamento. Isso não prejudica nossa análise, pois as três são correlacionadas entre si, isto é, carregam a mesma informação. Portanto, não temos problemas em atribuir pesos diferentes para informações diferentes. Em contrapartida, o uso das três variáveis nos permite agrupar os bairros utilizando os três critérios que compõem o IDH.

### 4.5.3 – Execução dos métodos de agrupamento e comparação da qualidade dos resultados

Nesta seção apresentamos os resultados da execução dos métodos de agrupamento com restrição de contigüidade e com restrição de acessibilidade. Os resultados dos dois métodos são comparados através das medidas de qualidade $PBM(K)$ e $Q(\prod)$.

#### 4.5.3.1 – Construção da matriz de vizinhança ou contigüidade

Para executarmos o método de agrupamento com restrição de contigüidade é preciso construir a matriz de vizinhança. O programa que implementa esse método constrói a matriz de vizinhança automaticamente. Entretanto, apenas com a vizinhança, o grafo que representa essas relações entre os bairros fica desconexo. Isso ocorre devido à existência de ilhas como a Cidade Universitária, Ilha do Governador e Paquetá. Dessa forma, inicialmente o método começa com quatro grupos, o que é ruim para nossa análise.

Portanto, para resolvermos esse problema, colocamos manualmente ligações da Cidade Universitária para a Maré e da Cidade Universitária para o Galeão na matriz de vizinhança. Essas ligações são baseadas na existência de rodovias interligando esses bairros. Uma ligação entre o Centro e Paquetá também foi adicionada. Esta ligação está baseada na existência do acesso de embarcações entre esses dois bairros. A adição dessas ligações faz com que o grafo de relações de vizinhança entre os bairros seja conexo.

#### 4.5.3.2 – Construção da matriz de acessibilidade e definição dos coeficientes de acessibilidade

O programa que implementa o método de agrupamento com restrição de acessibilidade via AGM constrói a matriz de acessibilidade automaticamente.

A matriz de acessibilidade é baseada no acesso dos bairros por rodovias. Quaisquer dois bairros são acessíveis, se existir uma rodovia interligando-os.

Entretanto, devido à ausência de rodovias entre Paquetá e o restante dos bairros, o grafo de acessibilidade entre os bairros fica desconexo. Para solução desse problema, utilizamos a mesma abordagem para a matriz de vizinhança. Adicionamos, manualmente, uma ligação entre Paquetá e o Centro na matriz de acessibilidade, tornando o grafo de acessibilidade conexo.

Os coeficientes de acessibilidade foram definidos em função da distância geográfica entre os bairros, pois não encontramos dados sobre fluxo de pessoas ou veículos para todas as rodovias utilizadas na análise.

Os bairros que são acessíveis e distam geograficamente em até 20 quilômetros possuem coeficiente de acessibilidade 1. Os bairros que são distantes geograficamente um do outro mais que 20 quilômetros possuem coeficiente de acessibilidade zero. Definimos o coeficiente dessa maneira, porque a extensão dos bairros inviabiliza o uso de um coeficiente de acessibilidade proporcional a distância geográfica.

O valor de 20 quilômetros foi escolhido experimentalmente. Executamos o método de agrupamento utilizando acessibilidade e verificamos que mesmo aumentando o valor da distância entre os bairros, os resultados não se alteravam significativamente.

#### 4.5.3.3 – Normalização dos dados

Os dados do IDH-L, IDH-E e IDH-R estão expressos na mesma unidade de medida. Entretanto, o intervalo de variação de cada um deles é diferente. Por isso, foi necessário normalizá-los.

Os dados foram transformados utilizando a fórmula de normalização MIN-MAX (SHALABI *et al.*, 2006):

$$z = \frac{x - \min(X)}{\max(X) - \min(X)}.$$

Com essa normalização, os valores das variáveis ficam no mesmo intervalo de variação, entre 0 e 1.

#### 4.5.3.4 – Execução e avaliação da qualidade dos métodos de agrupamento

As matrizes de acessibilidade e de vizinhança construídas, os coeficientes de acessibilidade e os dados normalizados foram utilizados para a execução dos métodos

de agrupamento. Inicialmente, os métodos são executados para encontrar de 2 até 15 grupos.

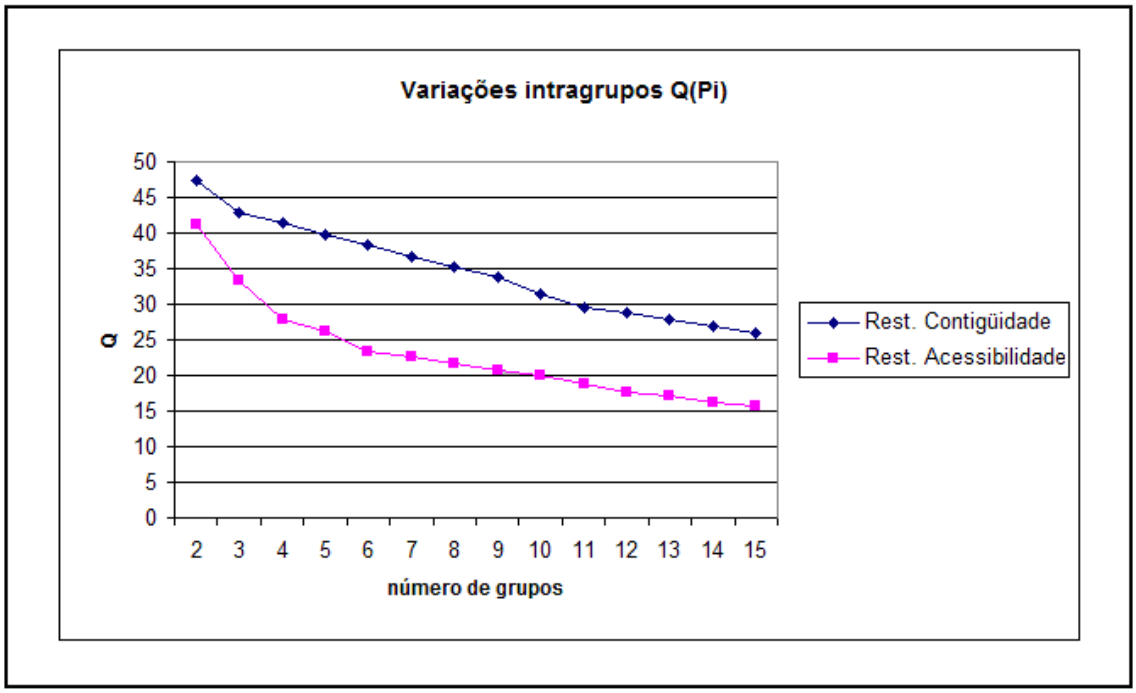


**Figura 4.12 - Gráfico da variação intragrupos para IDH dos Bairros**

A Figura 4.12 apresenta a variação do índice $Q(\prod)$ para os dois métodos. Nesta, podemos notar que o método com restrição de acessibilidade produz agrupamentos com valores de $Q(\prod)$ menores que o método com restrição de contigüidade.

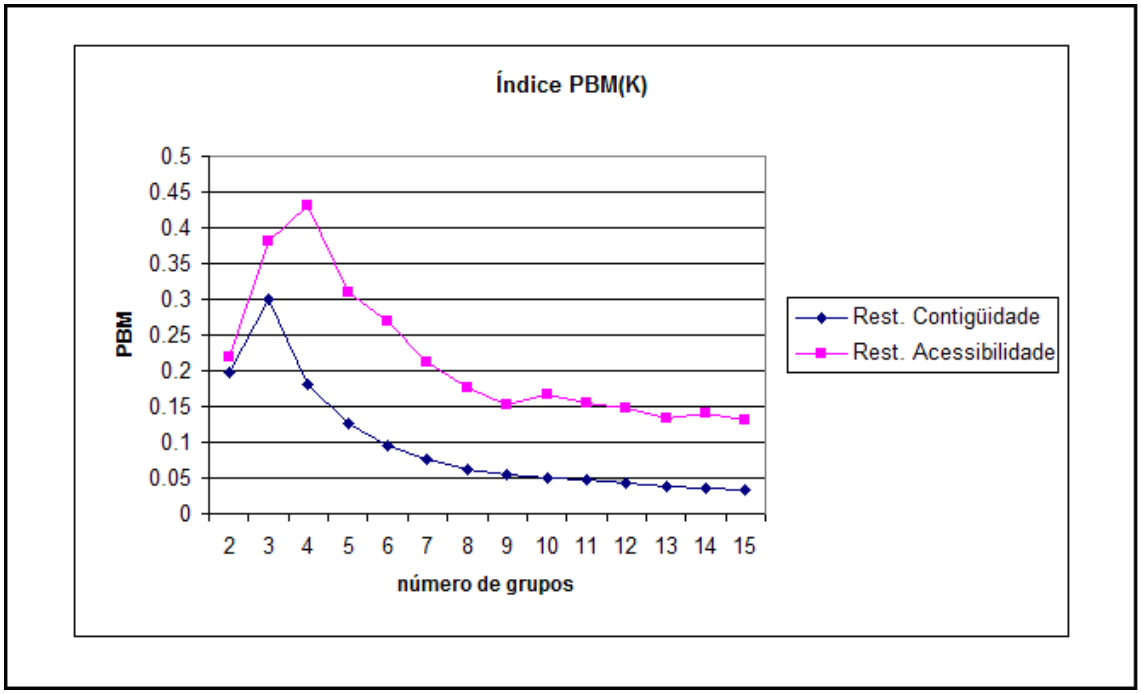


**Figura 4.13 - Gráfico do índice PBM para IDH dos Bairros**

Na Figura 4.13, temos a comparação do valor do índice PBM para o método com restrição de contigüidade e com restrição de acessibilidade. Os valores de PBM para o método de restrição de acessibilidade apresentaram-se maiores em todos os agrupamentos.

De acordo com as duas medidas de qualidade adotadas, o método de agrupamento com restrição de acessibilidade apresentou agrupamentos de melhor qualidade. Para o agrupamento com 2 grupos, as medidas de qualidade estão relativamente próximas, entretanto, para 3 grupos essa distância aumenta e é mantida à medida que o numero de grupos aumenta.

### 4.5.4 – Escolha do número de grupos

Na Figura 4.14, o maior valor de PBM ocorre para K=4, no método de agrupamento com restrição de acessibilidade. No método de agrupamento com restrição de contigüidade, o maior valor ocorre para K=3.

Esse valor pode ser confirmado no gráfico da Figura 4.15. Nesse gráfico, a maior queda no valor de $Q(\Pi)$ ocorre para 4 grupos no método de agrupamento com restrição acessibilidade e para 3 grupos no método de agrupamento com restrição de contigüidade.

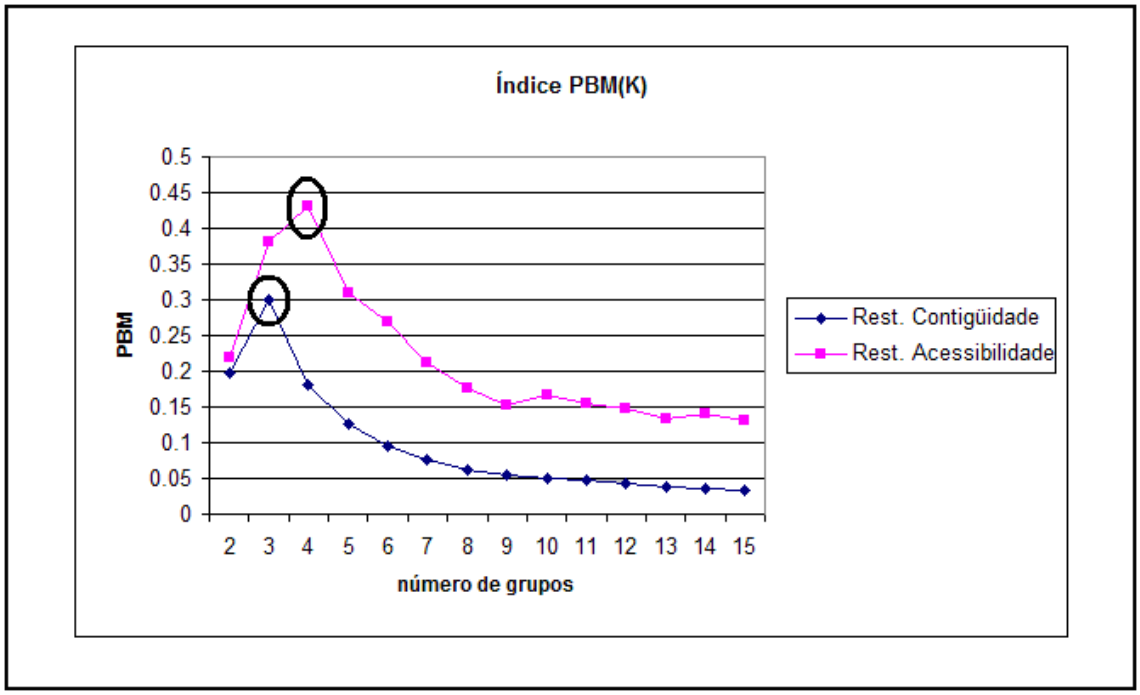


**Figura 4.14 – Maiores valores de PBM para o agrupamento dos bairros por IDH**

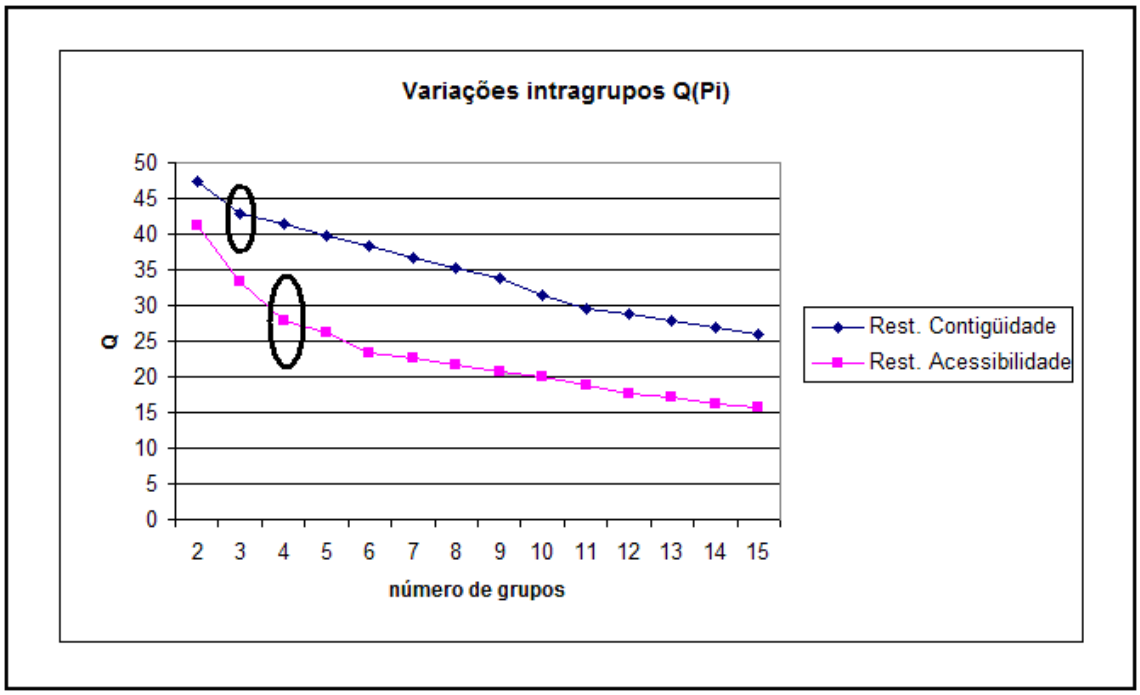


**Figura 4.15 – Queda das variações intragrupos para o agrupamento de bairros por IDH**

### 4.5.5 – Análise de agrupamento

Os dois métodos foram executados novamente para 4 grupos no método de agrupamento com restrição de acessibilidade e para 3 grupos no método de agrupamento com restrição de contigüidade. A Figura 4.16 e a Figura 4.17 ilustram os mapas com os grupos coloridos pelos dois métodos de agrupamento para os seus respectivos números de grupos que maximizam a qualidade do agrupamento.

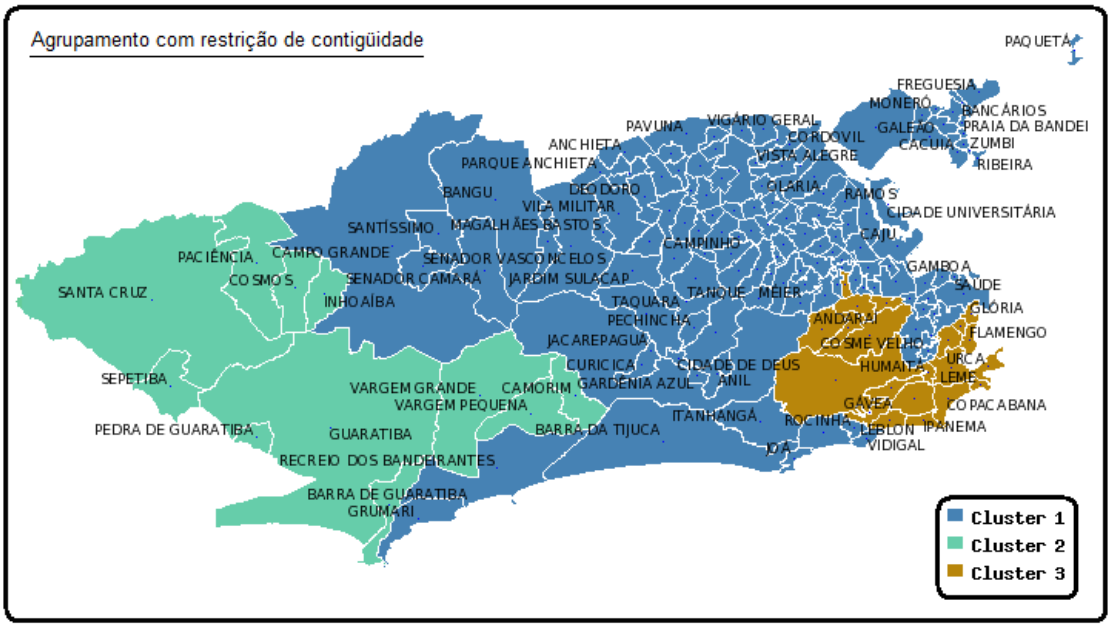


**Figura 4.16 - Agrupamento com restrição de contigüidade dos bairros por IDH**

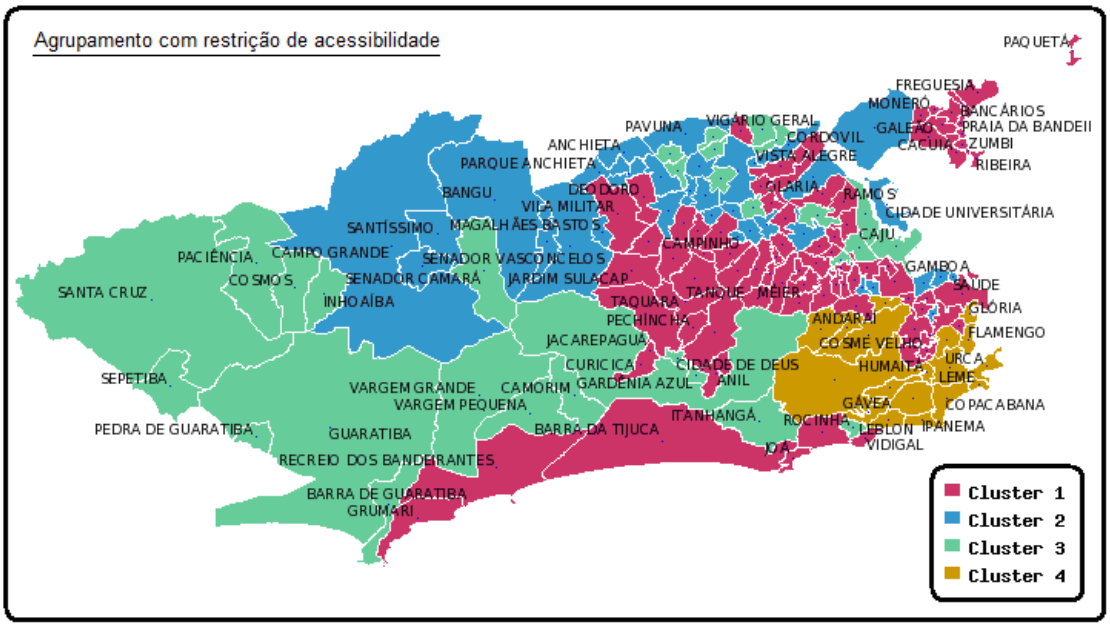


**Figura 4.17 - Agrupamento com restrição de acessibilidade dos bairros por IDH**

### 4.5.5.1 – Análise dos grupos

Nesta etapa, vamos analisar os 4 grupos formados pelo método de agrupamento com restrição de acessibilidade, pois este foi o método que resultou melhor qualidade no agrupamento.

***Sumário dos grupos formados***

| | | Média | | | Desvio padrão | | |
|---|---|---|---|---|---|---|---|
| **Grupo** | **Núm. Bairros** | IDH-L | IDH-E | IDH-R | IDH-L | IDH-E | IDH-R |
| 1 | 77 | 0.814 | 0.954 | 0.834 | 0.036 | 0.026 | 0.057 |
| 2 | 34 | 0.740 | 0.924 | 0.739 | 0.033 | 0.019 | 0.021 |
| 3 | 30 | 0.692 | 0.867 | 0.684 | 0.023 | 0.025 | 0.044 |
| 4 | 18 | 0.878 | 0.991 | 0.980 | 0.024 | 0.003 | 0.027 |

**Tabela 4.1 – Sumário dos grupos formados por bairros**

A Tabela 4.1 apresenta o número de bairros, a média e o desvio padrão para cada um dos grupos. Os valores do desvio padrão de cada grupo são menores que 10% da média, o que caracteriza grupos bem definidos. Entretanto, uma análise de dispersão de cada uma das variáveis nos permitirá investigar melhor a distribuição destas.

Na Figura 4.18, onde temos os gráficos *boxplots* das variáveis para cada grupo, notamos que, para os grupos 2, 3 e 4, as caixas dos gráficos são pequenas, indicando que há pouca dispersão entre os elementos dos grupos. No grupo 1, as caixas são maiores, isto é, bairros mais heterogêneos. No entanto, isso é justificado pela grande quantidade de bairros contidos nesse grupo.

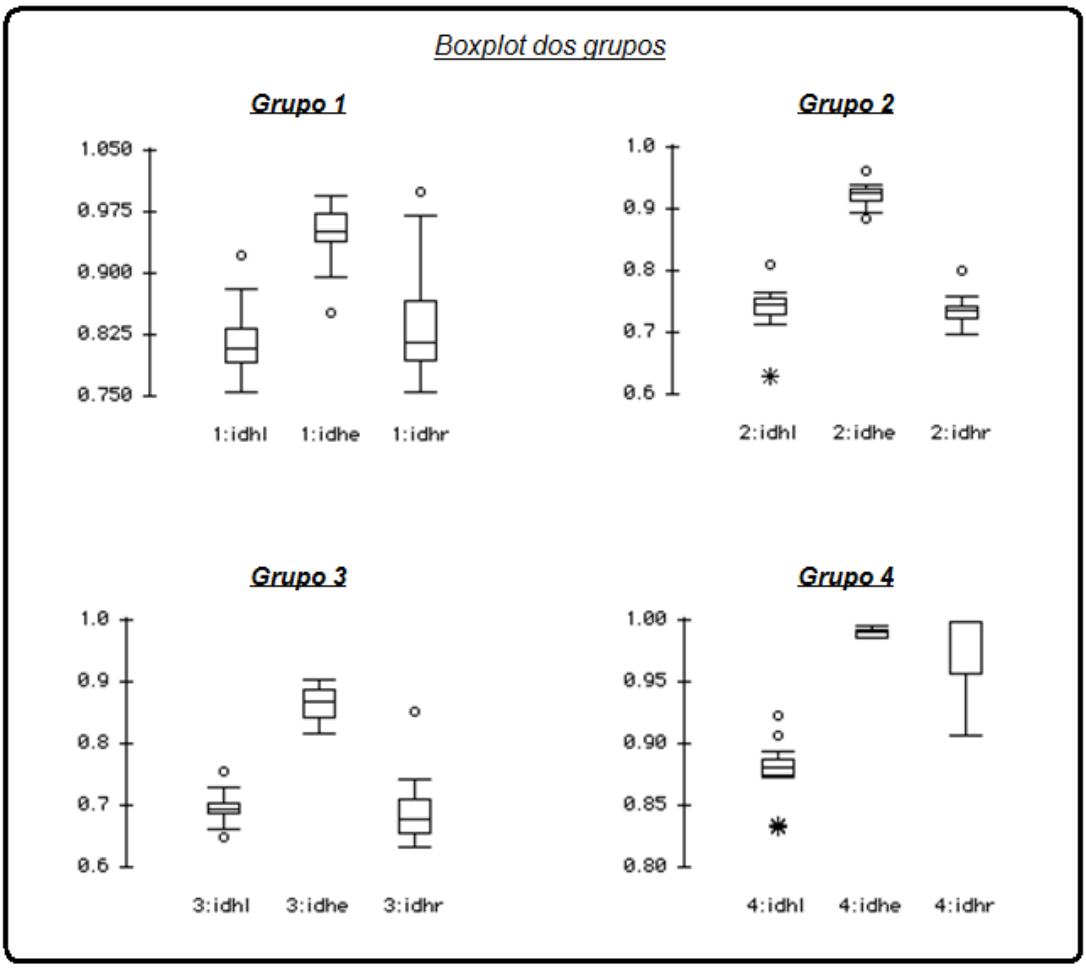


**Figura 4.18: BoxPlot dos grupos formados**

***Grupo 1***

O grupo 1 é o mais diversificado geograficamente (vide Figura 4.17), sendo composto da maior parte de bairros da zona norte, bairros próximos ao Centro, Ilha do Governador (exceto Galeão), os bairros da zona oeste que estão no litoral (Grumari, Recreio dos Bandeirantes e Barra da Tijuca), Joá, São Conrado e Vidigal.

Os *boxplots* de suas variáveis indicam que esse é o grupo que possui maior heterogeneidade de bairros. No entanto, isso é explicado pela enorme quantidade de bairros presentes nele, 77 dos 159.

O índice de desenvolvimento humano de longevidade médio deste grupo é considerado alto. Entretanto, como podemos notar no *boxplot* dessa variável, existem alguns bairros com IDH-L abaixo de 0.8. Mesmo assim, em relação aos outros grupos, o IDH-L é o segundo maior. O índice de desenvolvimento humano educacional médio é 0.954, considerado alto. Nesse grupo, o IDH-E, exceto por um bairro (valor extremo),

está entre 0.9 e 0.975. Assim como o IDH-L, o IDH-E desse grupo é o segundo maior, comparado aos outros 3 grupos. O IDH-R, índice que considera a renda, também segue as outras variáveis, sendo o segundo maior entre os grupos.

***Grupo 2***

O grupo 2 é formado por bairros da zona norte próximos a fronteira da cidade com a Baixada Fluminense e por bairros da zona oeste localizados mais ao norte da cidade. Esse grupo possui 34 bairros e, de acordo com os gráficos *boxplot* de suas variáveis, seus bairros são bastante homogêneos.

Os valores médios do IDH-L, IDH-E e IDH-R podem ser considerados como os terceiro melhores de todos os grupos. Essas medidas apresentaram pouquíssima dispersão, portanto, podemos avaliar essas variáveis utilizando os seus valores médios.

***Grupo 3***

Os grupo 3 é formado por 30 bairros da zona oeste que não estão no litoral carioca. Esse grupo possui alto grau de homogeneidade em seus bairros, pois suas variáveis apresentam pouca dispersão.

Os valores do IDH-L, IDH-E e IDH-R registram os mais baixos entre os 4 grupos. Chamamos atenção pela discrepância dos valores médios do IDH-L e do IDH-R em relação aos valores nos outros grupos. Esse grupo é formado pelos bairros com os piores indicadores de desenvolvimento humano da cidade do Rio de Janeiro.

***Grupo 4***

O grupo 4 é formado por 18 bairros pertencentes à zona sul e alguns à zona norte (Tijuca, Maracanã, Alto da Boa Vista, Andaraí e Grajaú). Embora, o IDH-R apresente uma maior dispersão que os outros índices, podemos afirmar que os bairros deste grupo têm alto grau de homogeneidade.

Esse grupo possui, com larga vantagem, os melhores indicadores de desenvolvimento humano da cidade do Rio de Janeiro. Seu IDH-E e IDH-R médios apresentam valores muito próximos a 1, teto da escala do IDH. Além disso, há muitas rodovias que interligam esses bairros, garantindo o acesso da população a vários serviços em outros bairros. Outro fator interessante é que esse grupo é formado por bairros relativamente pequenos, o que influência a acessibilidade para outros locais.

## 4.6 – Considerações finais

O experimento do estudo de caso apresentou resultados interessantes. O método de agrupamento com restrição de acessibilidade via AGM obteve grupos de melhor qualidade que o método de agrupamento com restrição de contigüidade. Através dos valores obtidos pelo índice PBM, os bairros foram agrupados pelo método de agrupamento com restrição de acessibilidade em 4 grupos.

Os 4 grupos encontrados foram analisados de acordo com os valores do IDH-L, IDH-E e IDH-R. Concluímos que os grupos estão definidos da seguinte forma:

- Grupo 1 apresenta a maior quantidade de bairros no Rio de Janeiro. Os valores dos índices são mais dispersos que os dos outros grupos. Apesar dessa dispersão, podemos afirmar que os valores dos índices desses bairros são os mais altos da cidade, perdendo apenas para os bairros do grupo 4.
- Grupo 2 é formado pelos bairros com os piores índices de desenvolvimento humano, ganhando apenas dos bairros do grupo 3.
- Grupo 3 é composto pelos bairros com os piores índices de desenvolvimento da cidade. Esses bairros todos localizados na zona oeste, possuem pouca acessibilidade ao restante da cidade e estão entre os maiores bairros em extensão.
- Grupo 4 é composto pelo bairros com os melhores índices de desenvolvimento humano. A maioria dos bairros desse grupo encontra-se na zona sul. Os valores desses índices são discrepantes em relação ao resto dos bairros da cidade. Outro fator interessante é que esse grupo possui bairros com muitas rodovias, isto é, alta acessibilidade.

Os grupos encontrados não são formados por regiões contiguas, mas por regiões que são acessadas através da malha rodoviária. A Figura 4.19 ilustra o exemplo em que a Barra da Tijuca está interligada aos bairros da zona norte pela Linha Amarela.

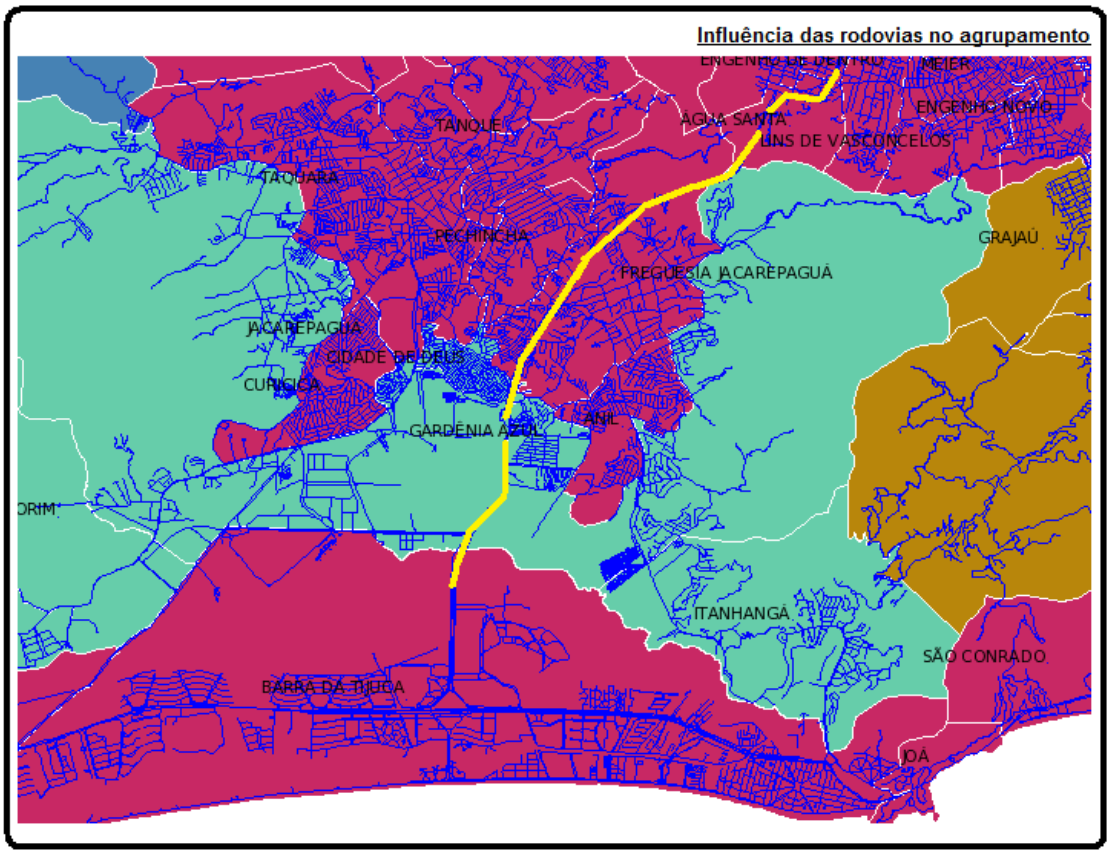


**Figura 4.19 - Linha Amarela interligando Barra da Tijuca à Zona Norte**

A qualidade dos resultados obtidos nos dois métodos confirma que as rodovias exercem grande influência na distribuição do IDH. O IDH-L está associado à qualidade de vida e acesso a hospitais. O IDH-E está associado à proximidade e acesso a escolas. Portanto, os resultados obtidos indicam que a abordagem utilizando acessibilidade para o agrupamento de regiões através dos dados do IDH é mais eficiente, em termos de qualidade, do que a abordagem por vizinhança.

# Capítulo 5 – Conclusão

Nesta dissertação, atentamos para a importância de agrupar regiões utilizando a informação de acessibilidade. Os métodos de agrupamento encontrados na literatura utilizam relações espaciais de distância ou relações topológicas de vizinhança como critério espacial para a formação de grupos. No entanto, há casos em que a informação sobre o acesso entre as regiões pode ser de grande valor para o agrupamento.

Uma pesquisa bibliográfica foi realizada com o intuito de identificar as principais abordagens de agrupamento espacial existentes. Nessa pesquisa, não foram encontrados métodos de agrupamento que utilizassem a informação de acessibilidade entre regiões. Portanto, propomos uma abordagem para o agrupamento de regiões através da informação de acessibilidade. Para isso, conceitos como relação de acessibilidade e coeficiente de acessibilidade foram definidos. Também desenvolvemos o método de agrupamento com restrição de acessibilidade via AGM, que utiliza nossa abordagem de agrupamento. Este foi baseado no método de agrupamento com restrição de contigüidade via AGM, onde algumas alterações foram realizadas para que pudéssemos trabalhar com a topologia de acessibilidade e com os coeficientes de acessibilidade.

Uma análise comparativa entre o método proposto e o método original foi realizada, onde seu objetivo foi avaliar a qualidade dos grupos encontrados através da abordagem por vizinhança e por acessibilidade. Essa análise consistiu de um estudo de caso sobre a distribuição do IDH nos bairros da cidade do Rio de Janeiro. Nesse estudo de caso, a abordagem de agrupamento por acessibilidade que propomos identificou grupos de bairros mais homogêneos que a abordagem por vizinhança. Embora esses grupos não sejam contíguos, os bairros de um mesmo grupo são acessíveis entre si. Esse relaxamento na restrição de contigüidade permite que grupos mais homogêneos sejam identificados.

Portanto, concluímos que a abordagem de agrupamento de regiões utilizando acessibilidade pode gerar resultados de melhor qualidade que a abordagem por vizinhança.

## 5.1 – Trabalhos futuros

Como possíveis trabalhos futuros desta dissertação, destacamos a realização de novos estudos de caso com o método de agrupamento com restrição de acessibilidade via AGM. Estes podem nos fornecer mais avaliações sobre o desempenho do método proposto. Além disso, a execução do método proposto com dados de tráfego de veículos em rodovias, tráfego de pessoas entre cidades *etc* pode trazer resultados mais interessantes para o agrupamento. Outro trabalho futuro que destacamos é a implementação do método de agrupamento com restrição de acessibilidade via AGM em ferramentas *open-source* como o WEKA, conforme proposto em MELLO *et al.* (2007).

# Referências


ALDENDERFER, M. S., BLASHFIELD, R. K., 1984, *Cluster analysis*, Beverly Hills, CA: Sage.

ANKERST, M., BREUNIG, M., KRIEGEL, H.-P., SANDER, J., 1999, "OPTICS: Ordering points to identify the clustering structure.", *In: Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pp. 49–60, Philadelphia, June.

ASSUNÇÃO, R. M., NEVES, M. C., CÂMARA, G., DA COSTA FREITAS, C., 2006, "Efficient regionalization techniques for socio-economic geographical units using minimal spanning tree", *International Journal of Geographical Information Science*, v. 20, n. 7 (Aug), pp. 797-811.

BEZDEK, C. J., PAL, R. N., 1998, "Some New Indexes of Cluster Validity", *IEEE Transations on systems, man, and cybernetics – Part B: Cybernetics*, v. 28, n. 3 (Jun).

BOLSTAD, P., 2005, *GIS Fundamentals: a first text on Geographic Information Systems*, 2 ed., USA, Eider Press.

CHATURVEDI, A., GREEN, P., CARROLL, J., 1994, "K-means, k-medians and k-modes: Special cases of partitioning multiway data.", *In The Classification Society of North America (CSNA)Meeting Presentation*, Houston.

CHATURVEDI, A., GREEN, P., CARROLL, J., 2001, "K-modes clustering.", *J. Classification*, v. 18, pp. 35–55.

CHEN, M., HAN, J., YU, P. S., 1996, "Data Mining: An Overview from Database Perspective", *IEEE Transactions on Knowledge and Data Eng.*, v. 8, n. 6 (Dec), pp. 866-883.

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., STEIN, C., 2001, *Introduction to Algorithms*, 2 ed., USA, MIT Press and McGraw-Hill.

DATADESK, 2008, "DataDesk 6.0", url: http://www.datadesk.com/products/data_analysis/datadesk/, acesso em Janeiro 2008.

DEMPSTER, A., LAIRD, N., RUBIN, D., 1977, "Maximum likelihood from incomplete data via the EM algorithm.", *Journal Royal Statistical Society*, v. 39, pp. 1–38.

DUBES, R. C., 1987, "How many clusters are best? – an experiment.", *Pattern Recognition*, v. 20, n. 6 (Nov), pp. 645-663.

ESTER, M., FROMMELT, A., KRIEGEL, H. P., SANDER, J., 2000, "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support". *Data Mining and Knowledge Discovery*, v. 4, n. 2 (Jul), pp. 193-216.

ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X., 1996, "A density-based algorithm for discovering clusters in large spatial databases.", *In Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pp. 226–231, Portland, Aug.

ESTER, M., KRIEGEL, H. P., SANDER, J., 2001, "Algorithms and Applications for Spatial Data Mining", In: Miller, H. J., Han, J., *Geographic Data Mining and Knowledge Discovery*, 1 ed., chapter 7, London and New York, UK and USA, Taylor and Francis.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., 1996, "Knowledge discovery and data mining toward a unifying framework", In*: Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining*, pp. 82-88.

FISHER, D, 1987, "Improving inference through conceptual clustering.", *In Proc. 1987 Nat. Conf. Artificial Intelligence (AAAI'87)*, pp. 461–465, Seattle, July.

FRIGGE, M., HOAGLIN, D. C., IGLEWICZ, B., 1989, "Some Implementations of the Boxplot", *The American Statistician*, v. 43, n. 1. (Feb), pp. 50-54.

GANTI, V., GEHRKE, J. E., RAMAKRISHNAN, R., 1999, "CACTUS—clustering categorical data using summaries.", *In Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, pp. 73–83, San Diego.

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

GUHA, S., RASTOGI, R., SHIM, K., 1998, "Cure: An efficient clustering algorithm for large databases", *In: Proc. Int. Conf. Management of Data (SIGMOD'98)*, pp. 73-84, Seattle, WA, June.

GUHA, S., RASTOGI, R., SHIM, K., 1999, “ROCK: A robust clustering algorithm for categorical attributes.”, *In Proc. 1999 Int. Conf. Data Engineering (ICDE’99)*, pp. 512–521, Sydney, Australia, March.

GUTING, R. H., 1994, “An Introduction to Spatial Database Systems”, *VLDB Journal*, v. 3, n. 4 (Out), pp. 357-399.

HAN, J., CAI, Y., CERCONE N., 1992, “Knowledge Discovery in Databases: an Attribute-Oriented Approach”, *In: Proc. 18th VLDB*, pp. 547-559, Vancouver, Canada.

HAN, J., KAMBER, M., TUNG, A. K. H., 2001, “Spatial clustering methods in data mining: A survey”, In: Miller, H. and Han, J. (eds.), *Geographic Data Mining and Knowledge Discovery*, 1 ed., chapter 8, London and New York, UK and USA, Taylor and Francis.

HAN, J., KAMBER, M., 2001, *Data Mining: Concepts and techniques*, 1 ed., USA, Morgan Kaufmann Publishers.

HAN, J., KAMBER, M., 2006, *Data Mining: Concepts and techniques*, 2 ed., USA, Morgan Kaufmann Publishers.

HINNEBURG, A., KEIM, D. A., 1998, “An efficient approach to clustering in large multimedia databases with noise.”, *In: Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD’98)*, pp. 58–65, New York, NY, Aug.

JAIN, A. K., MURTY, M. N., FLYNN, P. J., 1999, “Data clustering: a review”, *ACM Computing Surveys*, v. 31, n. 3 (Set), pp. 264-323.

KARYPIS, G., HAN, E. - H., KUMAR, V., 1999, “CHAMELEON: A hierarchical clustering algorithm using dynamic modeling”, *COMPUTER*, v. 32, pp.68-75.

KAUFMAN, L., ROUSSEEUW, P. J., 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, JohnWiley & Sons, 1990.

KOPERSKI, K., ADHIKARY, J., HAN, J., 1996, “Spatial data mining: Progress and challenges survey paper”, *In: Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada.

KREMPI, A. P., BRONDINO, N. C. M., SILVA, A.N.R., 2002, “Evaluating transportation accessibility with spatial statistics tools in a GIS environment”, *In:*

*Proc. of International Conference on Design and Decision Support Systems in Architecture and Urban Planning*, p.132-144, Eindhoven, Holland.

MACQUEEN, J., 1967, "Some methods for classification and analysis of multivariate observations.", *In Proc. 5th Berkeley Symp. Math. Statist. Prob*., v. 1, pp. 281–297, Berkeley.

MAPSERVER, 2008 "MapServer v.5.0", url: http://mapserver.gis.umn.edu/download/current/, acessodo em Janeiro 2008.

MELLO, C., SILVA, G. Z., SOUZA, J. M., 2007, "Extensão do WEKA para Métodos de Agrupamento com Restrição de Contigüidade", *In: XI Simpósio Brasileiro de Geoinformática*, São José dos Campos, Brasil, 2007.

NEVES, M. C., 2003, *Procedimento eficientes para regionalização de unidades sócio-ecônomicas em bancos de dados geográficos*, Tese de D.Sc., INPE, São José dos Campos, SP, Brasil.

NEVES, M.C., CÂMARA, G., ASSUNÇÃO, R.M., FREITAS, C.C., 2002, "Procedimentos Automáticos e Semi-automáticos de Regionalização por Árvore Geradora Mínima.", *In: Proc. Simpósio Brasileiro de Geoinformática*, Campos do Jordão, Brasil.

NG, R., HAN, J., 1994, "Efficient and Effective Clustering Methods for Spatial Data Mining", *In: Proceedings of 20th International Conference on Very Large DataBases*, pp 144-155.

NG, R., HAN, J., 2002, "CLARANS: A Method for Clustering Objects for Spatial Data Mining", *IEEE Trans. Knowledge & Data Engineering* , v.14, n. 5 (Set), pp. 1003-1016.

OPENSHAW, S., 1977, "A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling", *Transactions of the Institute of British Geographers* (New Series), v. 2, n.4, pp. 459-472.

PAKHIRAA, M. K., BANDYOPADHYAYB, S., MAULIK, U., 2004,"Validity index for crisp and fuzzy clusters", *Pattern Recognition*, v. 37, n.3 (Mar), pp: 487-501.

PNUD, 2008, "Programa das Nações Unidas para o Desenvolvimento". Disponível em http://www.pnud.org.br/idh, acessado em fevereiro de 2008.

POSTGRESQL, 2008, "PostgreSQL v.8.2", Disponível em http://www.postgresql.org/download/, acessado em Janeiro 2008.

SHALABI, L. A., SHAABAN, Z., KASASBEH, B., 2006, "Data mining: A preprocessing engine", *Journal of Computer Science*, v. 2, n. 9, pp. 735– 739.

SHEIKHOLESLAMI, G., CHATTERJEE, S., ZHANG, A., 1998, "WaveCluster: A multi-resolution clustering approach for very large spatial databases.", *In: Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pp. 428–439, New York, NY, Aug.

SILVA, A. N. R., LIMA, R. S., RAIA JR, A. A., VAN DER WAERDEN, P., 1998, "Urban transportation accessibility and social inequity in a developing country.", *In: Freeman, P. e C. jamet (eds.) Urban transport policy – A sustainable development tool*, Rotterdam, Balkema. p.709-714.

SKATER, 2008, "SKATER (Spatial 'K'luster Analysis by Tree Edge Removal)", url: http://www.est.ufmg.br/leste/skater.htm, acessada em Janeiro de 2008.

TUNG, A. K. H., HOU, J., HAN. J., 2001, "Spatial clustering in the presence of obstacles", *In: Proc. 17th International Conference on Data Engineering*, pp. 359-367, Heidelberg, Germany.

WANG, W., YANG, J., MUNTZ, R., 1997, "STING: A statistical information grid approach to spatial data mining.", *In Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97)*, pp. 186–195, Athens, Greece, Aug.

XIE, X.L., BENI, G., 1991, "A validity measure for fuzzy clustering", *IEEE Transations on Pattern Analisys and Machine Intelligence*, v. 12, n. 8 (Aug), pp. 841-847.

XU, R., WUNSCH II, D., 2005, "Survey of Clustering Algorithms", *IEEE Transactions on Neural Networks*, v. 16, n. 3 (may), pp. 645-678.

ZHANG, T., RAMAKRISHNAN, R., LIVNY, M., 1996, "BIRCH: an efficient data clustering method for very large databases", *In: Proc. of Int. Conf, Management of Data (SIGMOD'96)*, pp. 103-114, Montreal, Canada, June.