AGRUPAMENTO DE CONJUNTOS DE INSTÂNCIAS: UMA APLICAÇÃO AO ENEM

Victor Marinho Furtado

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Carlos Eduardo Pedreira

Rio de Janeiro

Julho de 2014

AGRUPAMENTO DE CONJUNTOS DE INSTÂNCIAS: UMA APLICAÇÃO AO ENEM

Victor Marinho Furtado

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) 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.

Examinada por:

Prof. Carlos Eduardo Pedreira, Ph.D.

Prof. Carlos Eduardo Ribeiro de Mello, D.Sc.

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Rodrigo Tosta Peres, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

JULHO DE 2014

Furtado, Victor Marinho

Agrupamento de Conjuntos de Instâncias: Uma Aplicação ao ENEM/ Victor Marinho Furtado. – Rio de Janeiro: UFRJ/COPPE, 2014.

XI, 95 p.: il; 29,7 cm.

Orientador: Carlos Eduardo Pedreira

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2014.

Referências Bibliográficas: p. 91-95.

1. Mineração de dados espaciais. 2. Análise de agrupamento. I. Pedreira, Carlos Eduardo. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Agradecimentos

Agradeço aos meus pais, Dalva e Carlos, pelo amor, apoio e insentivo que me deram durante toda a minha trajetória acadêmica.

À minha namorada Larissa, por todo seu amor, compreensão, apoio e também pelas palavras de carinho e motivação durante a realização deste trabalho.

À toda minha família e aos meus amigos pelo apoio, carinho e por compreenderem a minha ausência e dedicação ao trabalho.

Ao professor Carlos Pedreira pela excelente orientação, pelos ensinamentos, pelas palavras de motivação nos momentos oportunos e pela confiança no meu potencial.

Ao professor Carlos Mello pela excelente orientação, por todas horas de dedicação e pela confiança na minha capacidade.

Ao professor Geraldo Xexéo por aceitar participar da banca e pelas sugestões e conselhos utilizados no trabalho. Ao professor Rodrigo Peres por aceitar participar da banca e por contribuir para a concretização deste trabalho.

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 CONJUNTOS DE INSTÂNCIAS: UMA APLICAÇÃO AO ENEM

Victor Marinho Furtado

Julho/2014

Orientador: Carlos Eduardo Pedreira

Programa: Engenharia de Sistemas e Computação

O agrupamento de conjuntos de instâncias (observações) é um problema cujo objetivo é agrupar objetos que são representados por uma amostra. A abordagem adotada nesses casos é calcular alguma estatística dessas amostras, geralmente a média, e utilizá-la para representar o objeto. Assim, um algoritmo de agrupamento tradicional pode ser aplicado para resolver o problema, calculando a distância entre as estatísticas. Esta dissertação apresenta uma nova abordagem para este tipo de problema utilizando os conjuntos originais. A comparação entre os objetos para calcular a similaridade é feita a partir do teste de Kolmogorov-Smirnov para duas amostras. Este teste é utilizado quando se deseja decidir se duas amostras foram geradas da mesma população, a partir do cálculo do p-valor. Esta dissertação apresenta um estudo que indica ser viável a utilização do p-valor como uma medida de similaridade na aplicação de um método de agrupamento. Por fim, experimentos foram conduzidos para comparar os resultados obtidos entre o método proposto e a abordagem que calcula uma estatística das amostras. O problema abordado foi o agrupamento dos municípios do estado do Rio de Janeiro baseado nas notas de matemáticas do ENEM de 2011 e o experimento mostrou que o método proposto é viável e em alguns casos mais eficiente do que calcular alguma estatística.

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

SETS OF INSTANCES CLUSTERING: AN APPLICATION TO ENEM

Victor Marinho Furtado

July/2014

Advisor: Carlos Eduardo Pedreira

Department: Computer and Systems Engineering

Sets of instances (observations) clustering is a problem whose goal is to cluster objects that are represented by a sample. In these cases, the approach adopted is to calculate a statistical measure, usually the average, and use it to represent the object. Thus, a traditional clustering algorithm can be applied to solve the problem by calculating the distance between the statistical measures. This paper presents a new approach to solve this problem using the originals sets of instances (observations). We use the two-sample Kolmogorov-Smirnov test to estimate the similarity between the objects. This paper shows that the p-value from the Kolmogorov-Smirnov test can be used as a similarity measure in a clustering algorithm. Finally, an experiment was conducted to compare the results obtained by the proposed method and the statistical measure approach. The problem addressed was to cluster the cities of Rio de Janeiro, based on the math grades in ENEM of 2011. The result showed that the proposed method is feasible and in some cases more efficient than the statistical measure approach.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1 – Introdução

1.1 – Motivação

Atualmente, existem diversos métodos que propõe adquirir de forma automática algum tipo de conhecimento a partir de uma base de dados. Dentre eles, os métodos de agrupamento, que têm como objetivo classificar um conjunto de dados em grupos cujo significado não se conhece a priori. Neste caso, é necessário algum mecanismo de comparação que permita fazer estes agrupamentos.

Normalmente, existe uma grande de quantidade de características que descrevem os objetos de interesse no estudo. Por exemplo, se o objetivo for realizar uma análise sobre um conjunto de pessoas, diversos atributos poderiam ser utilizados para descrevê-las como altura, peso e idade. Assim, cada pessoa poderia ser representada por um ponto no espaço dos atributos (ALTURAxPESOxIDADE), tornando fácil verificar quais são as pessoas mais parecidas a partir da proximidade dos pontos.

Existem casos em que cada objeto é representado por um conjunto de instâncias (observações), fazendo com que ele seja mapeado em um conjunto de pontos no espaço dos atributos. Por exemplo, um problema cujo objetivo seja agrupar países que possuam pessoas com atributos parecidos. Neste problema, um país é representado por uma amostra de sua população, fazendo com que ele seja representado por um conjunto de pontos no espaço de atributos. Isto impossibilita que a solução de agrupar os pontos pela proximidade seja adotada diretamente.

Quando os objetos são representados por conjuntos de pontos, normalmente calcula-se alguma medida estatística para transformar os conjuntos em pontos e aplica-se algum método de agrupamento baseado em proximidade. Entretanto, a qualidade da solução apresentada por esta estratégia pode não ser boa devido a quantidade de informação que se perde quando se transforma um conjunto de pontos em apenas um objeto.

Supondo que, em um determinado, experimento o objetivo seja analisar alguns países com base na renda familiar. Caso seja calculado por exemplo a média das rendas familiares, perde-se a informação da distribuição da renda de cada país, se há ou não uma grande desigualdade, em que faixa de renda se concentra a maior parte das famílias, entre outras coisas. Além disso, algumas vezes é possível se obter uma grande amostra, mas suas observações acabam sendo desperdiçadas ao se utilizar medidas como a média.

Existem algumas técnicas que são utilizadas para comparar amostras de um determinado atributo (variável) baseadas nas distribuições das mesmas. Desta forma, é possível comparar o quão semelhantes são duas amostras baseando-se na frequência em que ocorrem os valores, considerando todas as informações provenientes das suas distribuições. Esse melhor aproveitamento dos dados disponíveis pode ser muito útil em diversos estudos, pois enriquece a comparação entre os objetos.

Tendo em vista o problema de como aplicar métodos de agrupamento em conjuntos de instâncias (observações), não foi encontrada na literatura nenhuma solução que fizesse uso de técnicas estatísticas de comparação das amostras.

1.2 – Objetivo

O objetivo desta dissertação é desenvolver uma abordagem não supervisionada para o agrupamento de conjuntos de instâncias (observações). Para servir de base para o experimento teste desta abordagem, foi definido o problema de agrupar os municípios do estado do Rio de Janeiro a partir das notas de matemática do Exame Nacional do Ensino Médio (ENEM) de 2011. A base do ENEM foi escolhida por conta da importância em avaliar as pessoas que desejam ingressar em uma universidade pública no Brasil. As notas de matemática foram escolhidas arbitrariamente.

Os municípios são agrupados por um algoritmo de agrupamento espacial. O algoritmo escolhido foi o SKATER, baseado em árvore geradora mínima, apresentado em ASSUNÇÃO *et al.* (2002).

Um conjunto de rotinas e padrões, conhecido como API (Application

Programming Interface), foi implementado em JAVA e utilizado em um experimento para testar a eficiência desta abordagem em relação as que mapeiam o conjunto de valores dos atributos aplicando a média.

1.3 – Estrutura

Esta dissertação está organizada em 5 capítulos, sendo esta introdução o primeiro.

O segundo capítulo consiste na revisão bibliográfica sobre análise de agrupamentos e testes de hipóteses estatísticos. Nesse capítulo, são apresentados e explicados diversos conceitos que são fundamentais para a leitura e compreensão da abordagem apresentada por esta dissertação.

No terceiro capítulo, é apresentada a metodologia utilizada para desenvolver a proposta desta dissertação e também a descrição detalhada da API.

O capítulo 4 apresenta a descrição do experimento realizado para comparação das abordagens, os resultados obtidos e, por fim, uma análise e discussão baseada nos resultados.

Por último, o capítulo 5 traz as conclusões tiradas da dissertação, algumas considerações finais e trabalhos futuros desta dissertação.

Capítulo 2 – Revisão da literatura e trabalhos relacionados

2.1 – Introdução

Neste capítulo, serão apresentados os conceitos básicos utilizados nesta dissertação, cuja proposta é criar um método de agrupamento espacial utilizando teste de Kolmogorov-Smirnov para duas amostras. O capítulo foi dividido em seções que descrevem diversos métodos e conceitos de análise de agrupamento e de testes de hipóteses, além de algumas definições necessárias para a compreenção da proposta.

2.2 – Análise de agrupamento

O volume de dados armazenados tem sido cada vez maior, devido a crescente necessidade das pessoas ampliarem seus conhecimento sobre os dados com os quais trabalham. A quantidade de dados cresceu tanto que passou a ser impossível para as pessoas extrairem maiores informações deles, sem o auxilio da computação. Essa excessiva busca por informações, levou ao desenvolvimento de diversos métodos computacionais de análise e processamento de dados. Dentre os métodos mais importantes, encontra-se a técnica de classificação, que tem como objetivo dividir os dados em grupos, cujo o significado pode ser conhecido a priori ou não.

O procedimento de classificação pode ser feito de forma supervisionada ou não-supervisionada (XU & WUNSCH II, 2005). A técnica de classificação supervisionada é feita mapeando os dados em classes (grupos), cujo o significado é conhecido no problema. Por exemplo, classificar um conjunto de filmes pelos seguintes gêneros: ação, comédia, drama, romance e terror. A classificação não-supervisionada, ou agrupamento, é feita quando não há classes pré-definidas (MELLO, 2008). O objetivo do agrupamento é particionar um conjunto finito de elementos em grupos, de forma que os elementos de um grupo sejam mais similares entre si do que se comparados aos de

qualquer outro grupo (XU & WUNSCH II, 2005). Por exemplo, dividir um conjunto de filmes em grupos de forma que filmes com gêneros parecidos fiquem juntos.

A análise de agrupamento é um procedimento que tem como objetivo obter algum conhecimento a partir de uma base de dados, utilizando um método de agrupamento. Este procedimento pode ser modelado de diferentes formas, como é possível ser ver em XU & WUNSCH II (2005) e JAIN *et al.* (1999). De um modo geral, os passos que devem ser executados são mostrados na Figura 2.1.

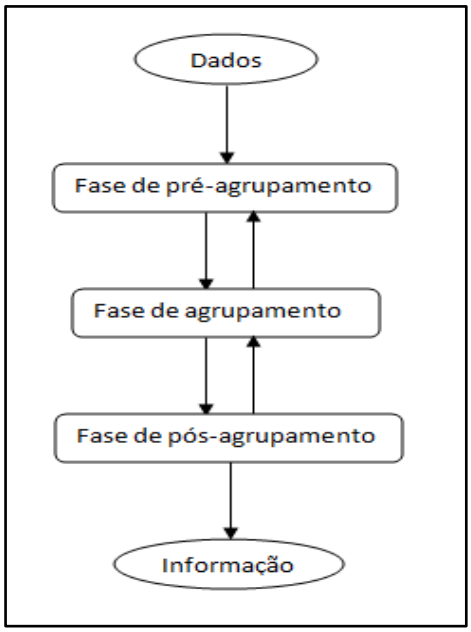


Figura 2.1 – Processo de análise de agrupamento

Fase de pré-agrupamento

Na fase de pré-agrupamento, os dados são preparados para execução do método de agrupamento. Nesta etapa, as principais atividades exercidas são a seleção e a extração de atributos. A seleção de atributos é a identificação dos atributos que sejam mais adequados para realizar o agrupamento (MELLO, 2008). A extração de atributos é a utilização de transformações nos atributos originais para produzir novos (JAIN et al., 1999, XU & WUNSCH II, 2005). Um exemplo de transformação é a Análise de Componentes Principais, do inglês Principal Component Analysis (PCA), que pode ser estudado com mais detalhes em JOLLIFFE (2005).

Fase de agrupamento

Na fase de agrupamento, é definida a medida de similaridade (ou dissimilaridade) que será utilizada como parâmetro para comparação entre os objetos. Existem diversas medidas possíveis que podem ser escolhidas, dependendo dos tipos de dados e do problema. Em seguida, um método de agrupamento deve ser aplicado nos dados. Este método tem por objetivo descobrir uma divisão natural da base de dados (JAIN, 2009). Muitos métodos foram criados e adaptados conforme a necessidade dos problemas em questão. XU & WUNSCH II (2005) diz que não há nenhum método de agrupamento que seja universal. Dessa forma, o método mais adequado somente poderá ser definido após se adquirir um profundo conhecimento do problema.

Fase de pós-agrupamento

A fase de pós-agrupamento engloba tudo que deve ser feito após a execução do algoritmo de agrupamento até a obtenção das informações que satisfaçam a necessidade destacada pelo problema. Primeiramente, os resultados obtidos pelo método devem ser avaliados e validados para garantir que os resultados sejam satisfatórios. JAIN *et al.* (1999) afirna que alguns algoritmos de agrupamento podem obter resultados melhores do que outros. Portanto, é importante se obter bons métodos de validação de agrupamento. Um dos métodos de validação mais utilizados é o Silhouettes proposto por ROUSSEEUW (1986). Alguns outros podem ser vistos no estudo realizado em HALKIDI *et al.* (2001).

Além disso, após a validação dos grupos é realizada a interpretação dos resultados, que é quando os especialistas nas áreas relacionadas ao problema definem o significado dos grupos encontrados. Assim, o conhecimento é finalmente obtido e o processo é chegado ao fim. XU & WUNSCH II (2005) afirma que para garantir a confiança do conhecimento extraído pode ser necessários outros experimentos e análises.

Por fim, as setas de retorno apresentadas no modelo da Figura 2.1 representam as possíveis indicações de que o passo anterior deva ser repensado. Por exemplo, se os atributos ou as transformações aplicadas na fase pré-agrupamento não foram adequadas, a fase de agrupamento pode ser comprometida fazendo com que seja necessário um

retrocesso no procedimento.

2.3 – Medidas de comparação

A proposta dos métodos de agrupamento é dividir o conjunto de dados de forma que cada grupo só possua objetos similares. Para isso, é necessário algum meio de comparação que indique se dois objetos são mais similares do que dissimilares entre si. Medida de similaridade é qualquer métrica que avalie o quão semelhantes são dois objetos. Esta avaliação pode ser feita de diversas formas, dependendo dos tipos dos dados dos objetos, que podem ser quantitativos ou qualitativos, contínuos ou binários, nominais ou ordinais (XU & WUNSCH II, 2005). Normalmente, cada um desses objetos são representados por um vetor numérico ou um ponto num espaço multidimensional, onde os atributos do objeto são simbolizados, respectivamente, por cada posição do vetor ou por cada dimensão (JAIN *et al.*, 1999).

Tabela 2.1 – Tabela de cidades com suas temperaturas

Cidade	**Temperatura Máxima**	**Temperatura Mínima**	**Cidade**	**Temperatura Máxima**	**Temperatura Mínima**
Porto Alegre	27º	18º	Recife	31º	23º
São Paulo	27º	19º	Fortaleza	31º	24º
Brasília	28º	17º	Belo Horizonte	32º	21º
Campo Grande	28º	20º	Cuiabá	32º	22º
Salvador	30º	22º	Macapá	33º	23º
Manaus	30º	23º	Vitória	35º	24º
Goiânia	31º	19º	Rio de Janeiro	36º	21º
Palmas	31º	22º			

Na Tabela 2.1, são mostradas as temperaturas máxima e mínima de algumas das principais cidades brasileiras. Neste exemplo, os dados armazenados na tabela podem ser representados como pontos de um espaço bidimensional, sendo uma dimensão representante das temperaturas máximas e a outra das mínimas, como é possível ver na Figura 2.2.

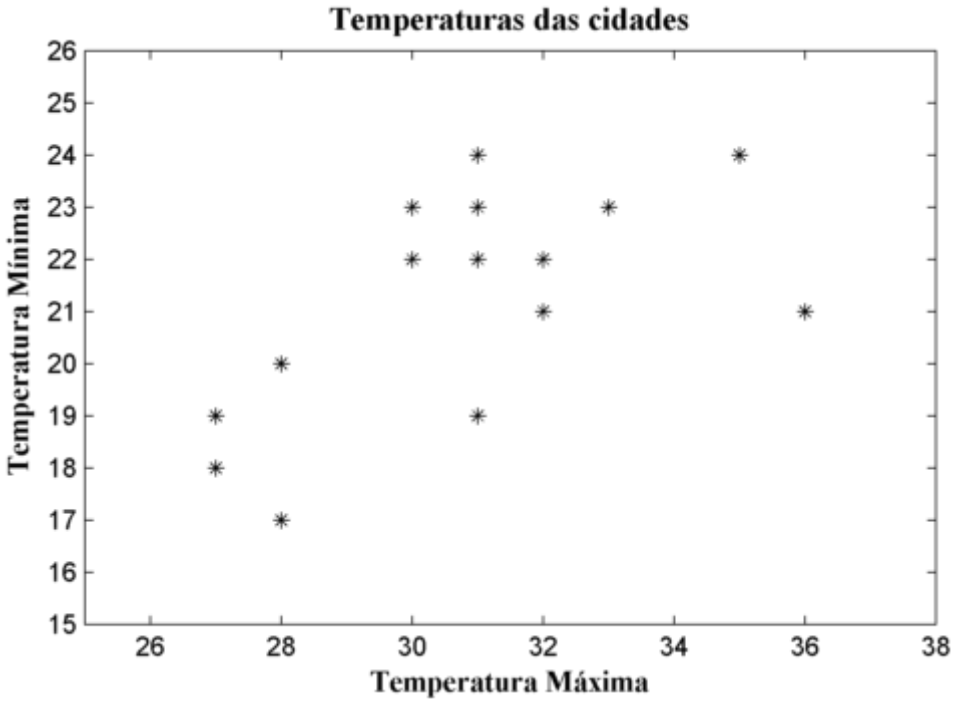


Figura 2.2 – Representação das cidades em pontos

A Figura 2.2 mostra a representação dos dados apresentados na Tabela 2.1 em pontos no espaço das temperaturas mínimas e máximas. Esta representação é utilizada pela maioria dos métodos de agrupamento, pois através dela é possível notar que existem regiões com maior concetração de pontos (JAIN *et al.*, 1999). A identificação dessas regiões é auxiliada por medidas de distância geométrica entre os pontos (HAN & KAMBER, 2001). Tais medidas têm como base a ideia de que quanto mais próximos estão os pontos, mais parecidos os objetos são. A Figura 2.3 mostra uma possível alocação dos dados da Tabela 2.1 em três grupos utilizando um algoritmo de agrupamento baseado em distância geométrica.

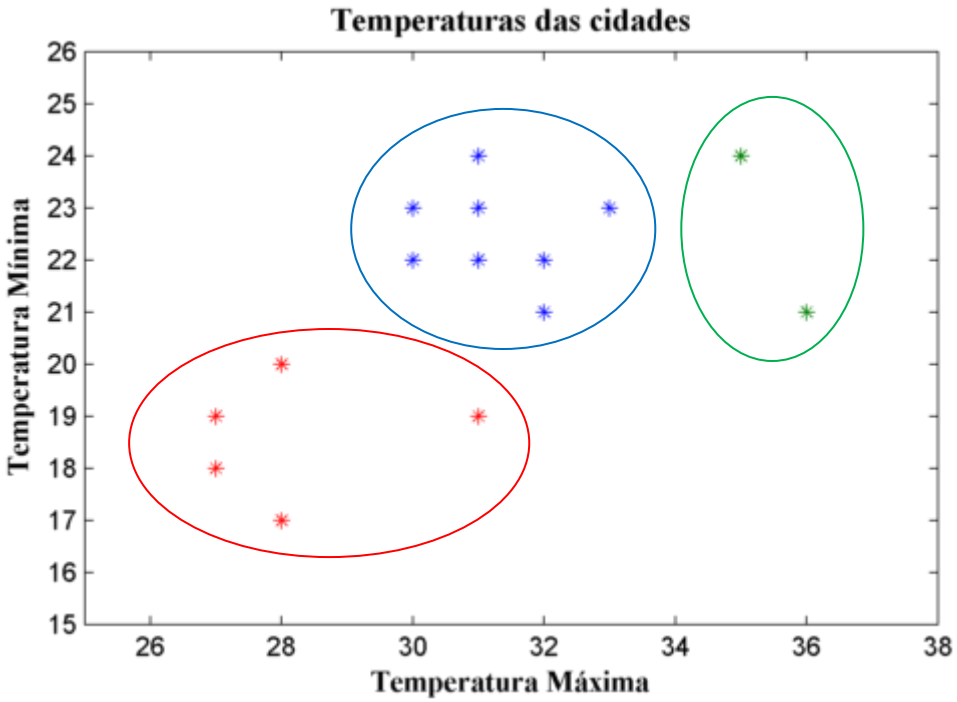


Figura 2.3 – Pontos alocados em 3 grupos

As cidades alocadas no grupo representado pela cor vermelha são Porto Alegre, São Paulo, Brasília, Campo Grande e Goiânia. É possível notar que nenhuma das cidades neste grupo possui temperatura mínima superior a 20ºC. A cidade de Goiânia apresenta uma temperatura máxima bem maior do que o restante, entretanto, a sua baixa temperatura mínima acabou fazendo com que ela pertencesse ao grupo vermelho. O grupo da cor verde possui as cidades do Rio de Janeiro e Vitória, com temperaturas máximas de 36ºC e 35ºC, respectivamente. Essas são as duas cidades com maior temperatura máxima, seguidas por Macapá, com 33ºC, que não foi alta o suficiente para colocá-la no grupo verde. Por último, o grupo azul com as cidades que possuem temperaturas mínimas e máximas nos intervalos de 21ºC à 24ºC e de 30ºC à 33ºC, respectivamente.

O exemplo anterior mostrou, na prática, que é possivel gerar grupos com objetos relativamente parecidos a partir de algoritmos de agrupamento que utilizam a distância geométrica como forma de calcular a semelhança entre os objetos. Em outras palavras, a distância geométrica entre os objetos no espaço gerado a partir de seus atributos pode ser utilizada como medida de similaridade ou dissimilaridade (XU & WUNSCH II, 2005). As medidas de distância mais conhecidas são a distância Euclidiana e a distância de Manhattan (ou de Quarteirão) (HAN & KAMBER, 2006).

A distância Euclidiana é calculada a partir da equação:

$$D_{ij} = \sqrt[2]{\sum_{l=1}^{L} |x_{il} - x_{jl}|^2}$$

Onde:

- D_{ij} é a distância Euclidiana entre os objetos i e j;

- l é o índice dos atributos dos objetos;

- L é o total de atributos dos objetos;

- x_{il} é o valor do l-ésimo atributo do objeto i;

Esta medida é amplamente utilizada, sendo aplicada principalmente no algoritmo de clusterização k-means. Ela tende a formar grupos com formatos hiperesféricos (XU & WUNSCH II, 2005). A Figura 2.4 mostra uma representação gráfica da distância Euclidiana.

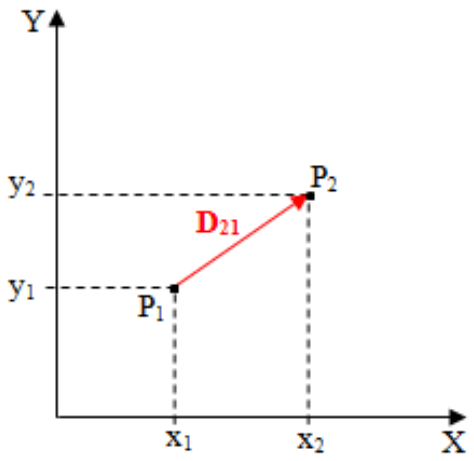


Figura 2.4 – Distância Euclidiana entre os pontos P1 e P2

A distância de Quarteirão é calculada a partir da equação:

$$D_{ij} = \sum_{l=1}^{L} |x_{il} - x_{jl}|$$

Onde as variáveis possuem o mesmo significado das descritas na distância Euclidiana.

Esta medida calcula a diferença entre cada atributo dos dois objetos e os soma para encontrar a distância. Ela tende a formar grupos com formatos hipercúbicos (MELLO, 2008). A Figura 2.5 ilustra a distância de Manhattan no $\mathbb{R}^2$.

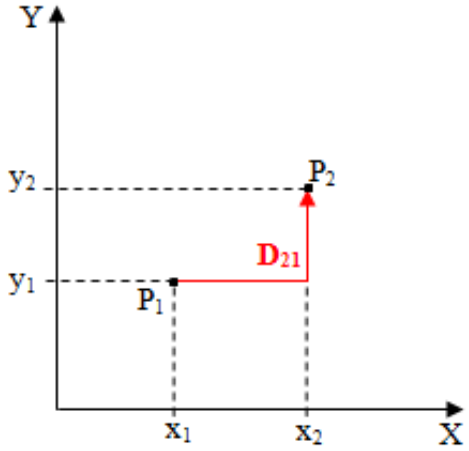


Figura 2.5 – Distância de Manhattan entre os pontos P1 e P2

Outra medida de distância importante é a distância de Minkowski, que é a generalização das distâncias Euclidiana e de Quarteirão (HAN & KAMBER, 2006). A distância de Minkowski é calculada a partir da seguinte equação:

$$D_{ij} = \sqrt[p]{\sum_{l=1}^{D} |x_{il} - x_{jl}|^p}$$

Onde o parâmetro p é um número inteiro positivo que determina a forma dos grupos encontrados (MELLO, 2008). Esta equação representa a distância de Quarteirão quando $p = 1$ e a distância Euclidiana quando $p = 2$ (HAN & KAMBER, 2006).

De acordo com HAN & KAMBER (2006), tanto a distância Euclidiana como a de Quarteirão respeitam as seguintes condições matemáticas de uma medida de distância:

1. $D_{ij} \geq 0$: A distância é sempre não negativa;
2. $D_{ii} = 0$: A distância de um objeto à ele mesmo é 0;

3. $D_{ij} = D_{ji}$: A distância é simétrica;
4. $D_{ij} \leq D_{ik} + D_{kj}$: Ir diretamente do objeto i ao objeto j, no espaço, não é mais distante do que partir de i, desviando-se por algum objeto k antes de chegar à j (desigualdade trianguar);

Estas medidas de distância são apenas algumas das muitas medidas de similaridade e dissimilaridade que existem. É possível encontrar medidas como a distância de Mahalanobis, a correlação de Pearson, distância de simetria de pontos e similaridade por cosseno em (XU & WUNSCH II, 2005).

Para que as medidas de similaridade baseadas em distância funcionem de maneira adequada, é necessário que todos os atributos estejam dentro do mesmo intervalo (MELLO, 2008). A normalização faz com que os dados dos atributos sejam escalados para específicos intervalos, como por exemplo de -1.0 a 1.0 ou de 0.0 a 1.0. Isto ajuda a previnir que atributos com intervalos maiores prevaleçam sobre atributos com intervalos menores (HAN & KAMBER, 2006). Existem diversos tipos de normalização de dados como Min-Max, Z-Score e escalonamento decimal, que podem ser encontrados em HAN & KAMBER (2001, 2006) e SHALABI *et al.* (2006).

Por fim, existem ainda medidas de similaridade para variáveis que não sejam contínuas. Em HAN & KAMBER (2006) e XU & WUNSCH II (2005) é possível ver alguns exemplos dessas medidas, embora definí-las ainda seja um desafio na área de análise de agrupamento (MELLO, 2008).

2.4 – Categorização dos métodos de agrupamento

Devido a grande quantidade de métodos de agrupamento existentes, organizá-los em categorias pode ser útil no momento de decidir o mais apropriado para o problema em questão. Entretanto, é possível encontrar diferentes formas de organização na literatura, como as que podem ser vistas em SONI & GANATRA (2012), HAN & KAMBER (2006) e BERKHIN (2006).

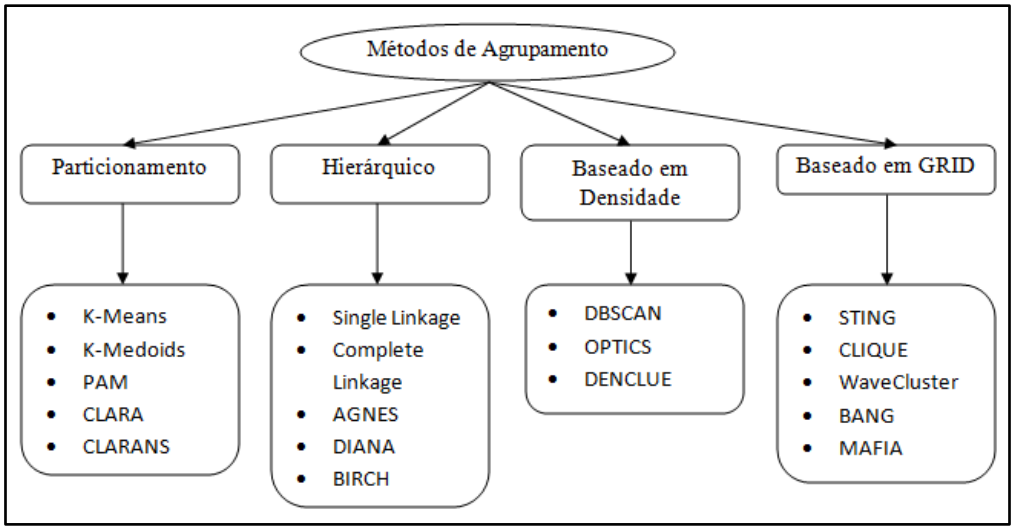


Figura 2.6 – Categorização dos métodos de agrupamento

De acordo com HAN *et al.* (2011), a divisão dos métodos em particionamento, hierárquicos, baseados em densidade e baseados em GRID, representa os métodos básicos ou fundamentais da análise de agrupamento. A Figura 2.6 ilustra essa organização dos métodos de agrupamento. Existem ainda métodos considerados avançados, como métodos baseados em modelos probabilísticos, métodos para dados com dimensões muito altas, métodos para grafos e redes e métodos com restrições, que podem ser explorados em (HAN *et al.*, 2011).

Nas próximas seções, serão apresentados as duas principais categorias de métodos de agrupamento, os métodos de particionamento e os métodos hierárquicos, além de um breve comentário sobre os métodos baseados em densidade e em GRID.

2.5 – Métodos de particionamento

Os métodos de particionamento são os mais simples e fundamentais dentro da análise de agrupamento (HAN *et al.*, 2011). Eles obtêm ao final do processo um simples particionamento e não uma estrutura de agrupamento do tipo dendrograma, como acontece nos métodos hierárquicos (JAIN *et al.*, 1999). Formalmente, dado um conjunto de n objetos e uma quantidade de grupos k, um método de particionamento organiza os n objetos dentro dos k grupos (onde $k \leq n$), de maneira que os objetos dentro de um

mesmo grupo sejam similares entre si e objetos de organizados em grupos diferentes sejam dissimilares (HAN *et al.*, 2011).

A seguir, serão apresentados alguns dos principais métodos de particionamento, como o k-means, k-medoids, PAM, CLARA e CLARANS.

2.5.1 – K-Means

Apesar de ter sido proposto há mais de 50 anos, o algoritmo k-means é o método de agrupamento mais utilizado nas aplicações científicas e industriais (JAIN, 2009, BERKHIN, 2006). Isto se deve ao fato de que o algoritmo k-means pode ser facilmente implementado e o custo computacional é de $O(n)$ (JAIN *et al.*, 1999). Inicialmente, o k-means foi desenvolvido para ser um método computacionalmente capaz alcançar o particionamento ótimo, entretanto, em grande parte dos casos, isso não acontece (MACQUEEN, 1967).

O k-means usa o ponto médio de uma partição como forma de representá-la. Cada ponto, também chamado de centróide, é definido pelos valores médios de todos os objetos que compõem o respectivo grupo. O método inicia escolhendo aleatoriamente k objetos para representar os centróides dos k grupos. Em seguida, cada um dos demais objetos é adicionado ao grupo que tiver o centróide mais próximo dele. Depois de formados os grupos, os novos centróides são calculados e novamente os objetos devem ser alocados nos grupos dos centróides mais próximos. Este processo é repetido até que o critério de agrupamento convirja ou, em outras palavras, quando não haja mais alterações nos grupos.

Normalmente, a qualidade do particionamento gerado pelo k-means é dado pela variação intragrupo, que é a soma dos erros quadráticos entre os objetos e seus respectivos centróides, dado pela seguinte equação (HAN *et al.*, 2011):

$$E = \sum_{i=1}^{k} \sum_{p \in G_i} |p - c_i|^2$$

Onde:

- E é a soma dos erros quadráticos;

- c_i é o centróide do grupo G_i;

- p é um ponto no espaço que representa um dado objeto;

Para cada objeto dentro de cada grupo, a distância entre ele e o centróide de seu grupo é elevada ao quadrado e em seguida somada às demais distâncias. Ao tentar minimizar esta função, o k-means tende a criar grupos tão compactos e tão afastados entre si quanto possível (HAN *et al.*, 2011).

A Figura 2.7 mostra o passo a passo de como ocorre a execução do algoritmo.

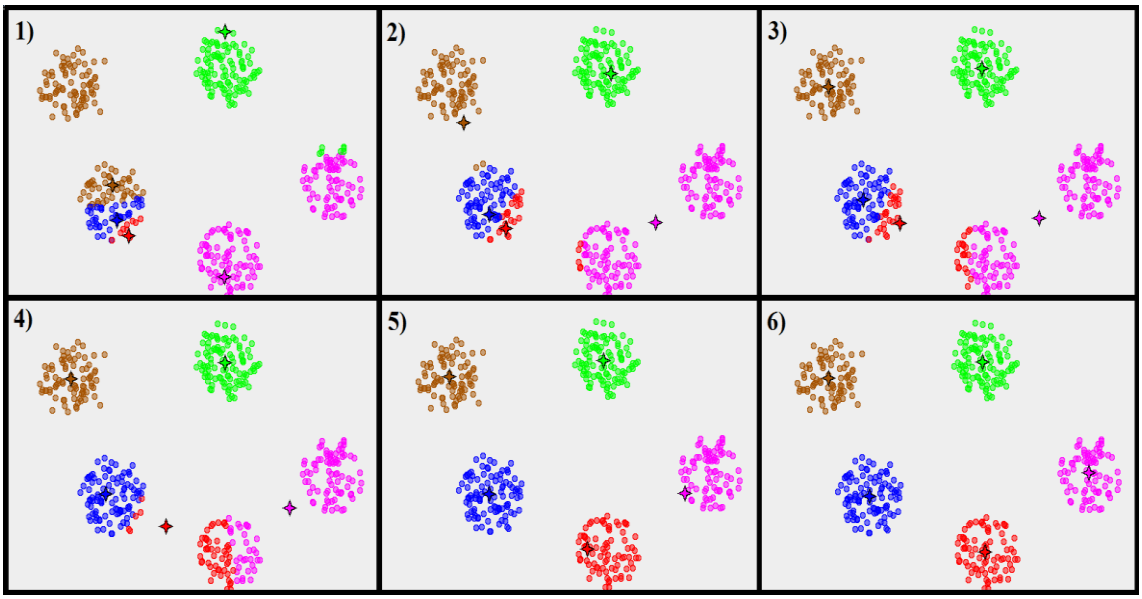


Figura 2.7 – Execução do K-Means (MIRKES, 2011)

As principais vantagens de se usar o método k-means é a facilidade de implementação e o custo computacional. Além disso, o k-means também é um bom método para lidar com grande quantidade de dados (MACQUEEN, 1967). Entretanto, o k-means só pode ser aplicado em objetos cuja média seja calculável, o que pode não ocorrer caso os objetos possuam atributos nominais (MELLO, 2008). Outra desvantagem do método, é o fato de não ser adequado para encontrar grupos com formatos não-convexos e de tamanhos muito diferentes (HAN *et al.*, 2011). Além disso, existe a necessidade de se definir o parâmetro k, o que normalmente é um desafio por não ser possível ter uma forma de se vizualizar a organização dos dados no espaço.

O k-mean é sensível a ruídos e valores atípicos (outliers), pois apenas uma pequena quantidade de dados extremos já é capaz de influenciar fortemente na média do grupo, afetando erroneamente na posição do centróide (MELLO, 2008).

Apesar das desvantagens apresentadas, o k-means é bastante utilizado, inclusive como base para métodos mais robustos ou mesmo específicos para outras aplicações. CHATURVEDI *et al.* (2001) apresentou um método chamado k-modes, que é uma variação do k-means, para agrupar dados nominais. PELLEG & MOORE (2000) apresentam uma variação, chamada de x-means, que oferece uma solução para três das maiores deficiências do k-means, que são a demora para processar cada iteração do método, a necessidade de se decidir a quantidade de grupos (parâmetro k) e a tendência em cair em mínimos locais.

2.5.2 – K-Medoids

O método k-medoids é mais um dos métodos que pode ser considerado uma variação do k-means. Entretanto, no k-medoids, ao invés de calcular o ponto médio dos objetos dentro de cada grupo para encontrar os centróides, um dos objetos é escolhido para ser o centróide (ou medoid) do grupo (REYNALDS *et al.*, 2004). Por causa disso, o k-medoids é menos sensível a ruídos e valores atípicos, o que o faz demandar mais tempo de processamento (HAN *et al.*, 2001, KAUFMAN & ROUSSEEUW, 2005).

O k-medoids funciona da seguinde maneira, k objetos são escolhidos ao acaso para ser os medoids iniciais (REYNALDS *et al.*, 2004). Cada um dos objetos restantes deverá ser alocado no grupo que tiver o medoid mais similar a ele. O algoritmo prossegue baseado no princípio de minimização da soma das distâncias entre cada objeto de um grupo e seu respectivo medoid (HAN *et al.*, 2011). O k-medoids minimiza a chamada soma dos erros absolutos, que é calculado a partir da seguinte equação (MELLO, 2008):

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

Onde:

- E é a soma dos erros absolutos;

- m_i é o objeto que representa o medoid do grupo G_i;

- p é um ponto no espaço que representa um dado objeto;

Estes passos, de alocar os objetos em um grupo e recalcular os medoids, são repetidos até que não haja mais alterações nos objetos definidos como medoids. A Figura 2.8 ilustra a execução do método.

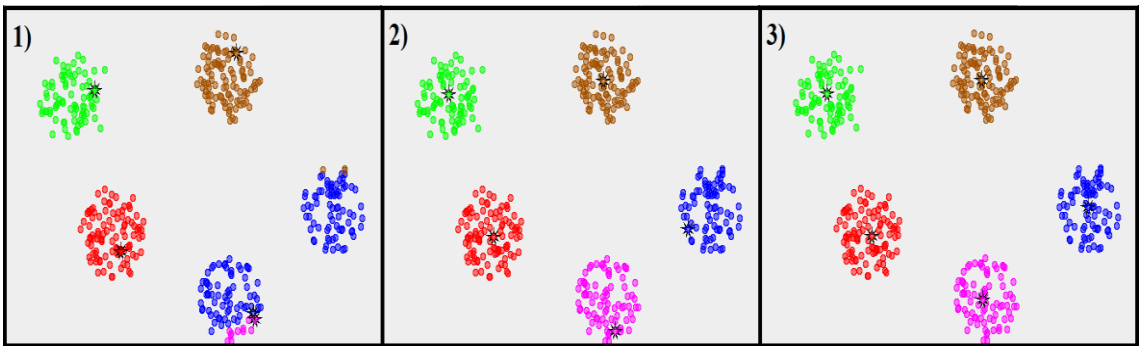


Figura 2.8 – Execução do K-Medoids (MIRKES, 2011)

2.5.3 – PAM (Partitioning Around Medoids)

O método PAM, desenvolvido e apresentado em KAUFMAN & ROUSSEEUW (2005), foi um dos primeiros algoritmos a utilizar a abordagem do k-medoids (HAN & KAMBER, 2006). Este algoritmo aborda o problema de uma forma iterativa e gulosa (HAN *et al.*, 2011).

O algoritmo PAM consiste em duas etapas, a primeira é chamada de construção e a segunda de troca (KAUFMAN & ROUSSEEUW, 2005). Na primeira etapa, um particionamento inicial é obtido definindo os k objetos que serão os medoids. Em seguida, na etapa de troca, cada um dos representantes dos grupos (medoids) é comparado com cada um dos objetos restantes, para testar se a qualidade do particionamento pode ser melhorada (KAUFMAN & ROUSSEEUW, 2005). Essa comparação é feita trocando o medoid pelo objeto que está sendo testado e verificando o que acontece com a qualidade do agrupamento.

A complexidade do método PAM é $O(k(n-k)^2)$, onde k é a quantidade de partições (grupos) a serem criadas e n é a quantidade de objetos. Para valores muito elevados de k e n, o custo computacional pode se tornar muito alto (HAN *et al.*, 2011).

2.5.4 – CLARA (Clustering LARge Applications)

O método CLARA é outro método baseado em k-medoids, desenvolvido por Kaufman e Rousseeuw (KAUFMAN & ROUSSEEUW, 2005). O PAM é um algoritmo eficiente para conjunto de poucos dados, mas que não trabalha bem com muitos dados (HAN *et al.*, 2011). Por esse motivo, o método CLARA foi desenvolvido para lidar com grande quantidade de dados de forma mais eficiente (KAUFMAN & ROUSSEEUW, 2005). Para isso, o CLARA trabalha com amostras aleatórias ao invés de utilizar todos os pontos (HAN *et al.*, 2011).

O método funciona em duas etapas. Primeiro uma amostra é retirada do conjunto de objetos e os k medoids são calculados dentro dessa amostra, utilizando algum método baseado em k-medoids. Em seguida, cada um dos objetos que não pertencem à amostra são associados ao grupo que tiver o medoid mais próximo. Desta forma, é gerado um particionamento utilizando-se todos objetos. A qualidade do particionamento é calculada da mesma forma que em um algoritmo baseado em k-medoids. Esse processo é repetido algumas vezes e o melhor particionamento é selecionado como resultado.

O custo computacional de gerar um particionamento a partir de uma amostra dos dados é $O\left(ks^2 + k(n-k)\right)$, onde s é o tamanho da amostra, k a quantidade de grupos e n o total de objetos (HAN *et al.*, 2011). O CLARA tem um bom desempenho para grandes quantidade de dados (NG & HAN, 1994). Entretanto, a sua eficácia depende do tamanho e da qualidade da amostra (HAN *et al.*, 2011).

2.5.5 – CLARANS (Clustering Large Applications based on RANdomized Search)

O método CLARANS foi desenvolvido e apresentado em NG & HAN (1994). Ele apresenta um meio termo entre as soluções de custo computacional e eficácia apresentadas pelos métodos PAM e CLARA (HAN *et al.*, 2011). O método não utiliza a abordagem gulosa do PAM, nem restringe o universo de busca à uma amostra do conjunto de dados (HAN & KAMBER, 2006).

Como normalmente acontece com métodos baseados em k-medoids, inicialmente o algoritmo seleciona k objetos aleatoriamente para formarem os k medoids. Em seguida, o algoritmo seleciona aleatoriamente um dos medoids x e um dos objetos que não seja um medoid y e calcula a qualidade do particionamento gerado a partir da troca do medoid x por y. Caso a qualidade tenha melhorado, a troca é efetivada, caso contrário o particionamento continua como estava. Essas tentativas de trocas são repetidas l vezes e o resultado é considerado um ótimo local. O processo todo é repetido m vezes e o melhor resultado é selecionado como o particionamento final.

NG & HAN (1994) mostrou a partir de experimentos que o CLARANS é mais eficiente que o PAM e o CLARA para conjuntos de dados pequenos ou grandes. A grande desvantagem do CLARANS é que para executá-lo é necessário a definição de mais dois parâmetros, além da quantidade de partições k, que são a quantidade de tentativas de trocar um medoid l e a quantidade de ótimos locais m que devem ser encontrados (MELLO, 2008).

2.6 – Métodos hierárquicos

Enquanto os métodos de particionamento trabalham com a ideia básica de criar uma forma de organizar todos os objetos em uma determinada quantidade de grupos, às vezes pode ser necessário organizar o particionamento em níveis, como uma hierárquia (HAN *et al.*, 2011). Os métodos hierárquicos trabalham com todas as quantidades de grupos possíveis, ou seja, variando $k = 1$ (todos os objetos no mesmo grupo) até $k = n$ (cada objeto em um grupo diferente) (KAUFMAN & ROUSSEEUW, 2005). Em outras palavras, esses métodos constroem uma árvore de particionamentos, também conhecida como dendrograma (JAIN *et al.*, 1999, BERKHIN, 2006).

Os métodos hierárquicos podem ser de dois tipos diferentes, os aglomerativos e os divisivos (KAUFMAN & ROUSSEEUW, 2005, EVERITT *et al.*, 2001). Os métodos aglomerativos inicializam o processo com cada objeto pertencendo a um grupo diferente. Em seguida, esses grupos vão se unindo até que no fim todos os objetos estejam no mesmo grupo ou algum critério de parada seja satisfeito (HAN *et al.*, 2011). Nos métodos divisivos ocorre justamente o oposto, o processo inicializa com apenas um

grupo com todos os objetos. Este grupo sofre sucessivas divisões até que cada objeto esteja alocado em um grupo diferente (BERKHIN, 2006).

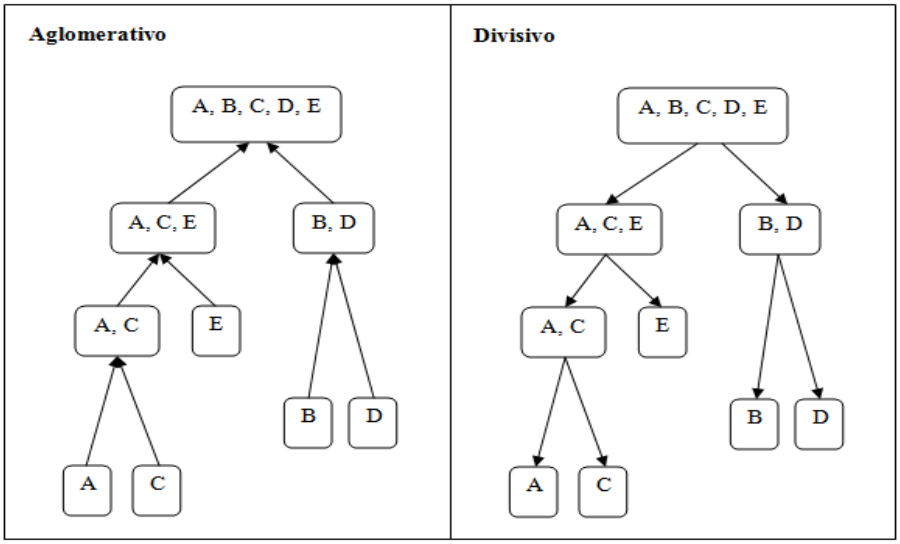


Figura 2.9 – Diferença entre métodos aglomerativos e divisivos

A Figura 2.9 mostra a diferença entre as abordagens dos métodos aglomerativos e divisivos. Nesta seção serão apresentados alguns dos principais algoritmos baseados na abordagem dos métodos hierárquicos, tanto aglomerativos como divisivos.

2.6.1 – AGNES (Agglomerative NESting)

O método AGNES foi desenvolvido e apresentado por KAUFMAN & ROUSSEEUW (2005). Este método é baseado na ideia dos métodos aglomerativos, que os grupos vão se unindo baseado em algum crtério até que exista apenas um grupo (KAUFMAN & ROUSSEEUW, 2005).

O AGNES funciona da seguinte forma, o algoritmo inicializa com cada objeto em um grupo diferente. Em cada passo, os dois grupos mais similares (ou menos dissimilares) se unem virando apenas um grupo com os objetos existentes nos dois. A forma como deve calculada a dissimilaridade entre dois grupos é um ponto crucial, pois a maioria dos métodos aglomerativos se diferenciam nesse ponto (KAUFMAN & ROUSSEEUW, 2005). O AGNES calcula a dissimilaridade ou apenas distância entre dois grupos a partir do método chamado distância média, que é dada pela seguinte

equação:

$$D_{G_i G_j} = \frac{1}{n_i n_j} \sum_{p \in G_i} \sum_{p' \in G_j} |p - p'|$$

Onde:

- $D_{G_i G_j}$ é a distância média entre os grupos G_i e G_j;

- n_i é a quantidade de objetos que fazem parte do grupo G_i;

- p e p' são os pontos no espaço que representam objetos dos grupos G_i e G_j, respectivamente;

A Figura 2.10 mostra um exemplo de um dendrograma gerado a partir da execução do método AGNES em um conjunto de 15 objetos.

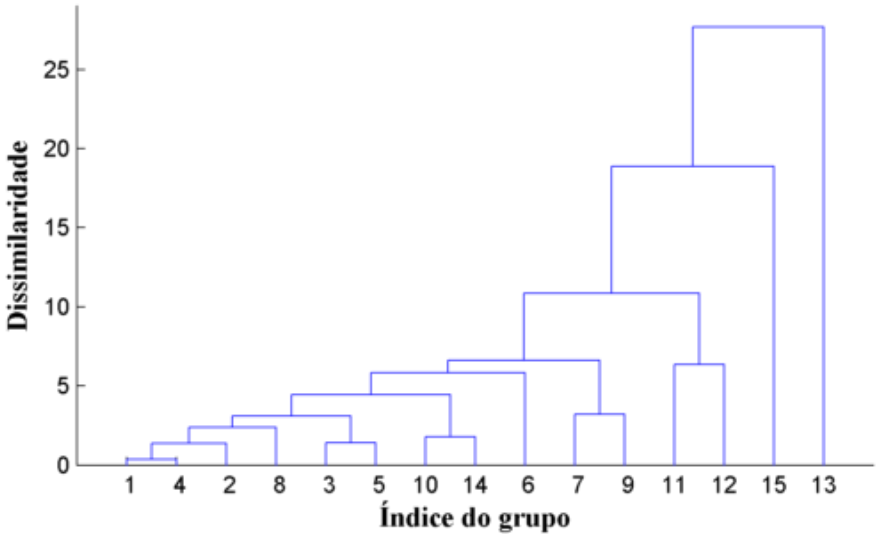


Figura 2.10 – Dendrograma gerado pelo AGNES

2.6.2 – Single linkage e complete linkage

Existem diversos métodos que abordam a ideia de agrupamento hierárquico aglomerativo (KAUFMAN & ROUSSEEUW, 2005). Em grande parte das vezes, a diferença entre os métodos é apenas a forma como a dissimilaridade é calculada. O método chamado single linkage, que é o algoritmo aglomerativo mais antigo e mais simples, funciona exatamente como o AGNES, exceto pela forma como a dissimilaridade é calculada (KAUFMAN & ROUSSEEUW, 2005). A dissimilaridade do single linkage é calculada pela distância mínima, que é dada pela equação:

$D_{G_iG_j} = min_{p \in G_i, p' \in G_j}(|p - p'|)$

Onde:

- $D_{G_iG_j}$ é a distância mínima entre os grupos G_i e G_j;

- p e p' são os pontos no espaço que representam objetos dos grupos G_i e G_j, respectivamente;

Isso faz com que a dissimilaridade entre os dois grupos seja dada pela distância entre os dois objetos mais próximos, tal que cada um pertença a um dos grupo.

Em contrapartida ao single linkage, existe o método chamado complete linkage (KAUFMAN & ROUSSEEUW, 2005). Ele trabalha da mesma forma que o single linkage, entretanto a dissimilaridade utilizada é baseada na distância máxima. O cálculo da distância máxima é dado pela equação:

$D_{G_iG_j} = max_{p \in G_i, p' \in G_j}(|p - p'|)$

onde os parâmetros são os mesmos da equação da distância mínima.

De forma oposta ao que acontece com o single linkage, este método adota como dissimilaridade entre dois grupos a distância entre os dois objetos mais distantes, cada um pertencendo a um dos dois grupos.

2.6.3 – DIANA (DIvisive ANAlysis)

O DIANA (KAUFMAN & ROUSSEEUW, 2005) é um método divisivo e segue a linha oposta dos métodos aglomerativos, começando com apenas um grupo e o dividindo até que cada objeto esteja em grupos diferentes. Na literatura existe muito mais métodos aglomerativos do que divisivos (HAN *et al.*, 2011). No geral, quando as pessoas falam sobre métodos hierárquicos elas estão se referindo a métodos aglomerativos (KAUFMAN & ROUSSEEUW, 2005). O principal motivo para isso é o custo computacional envolvido nos métodos divisivos. Existem $2^{n-1} - 1$ formas possíveis de se dividir um grupo com n objetos em dois grupos (HAN *et al.*, 2011). É fácil ver que, mesmo para valores não tão grandes de n, se torna computacionamente inviável analisar todas as possíveis combinações e por isso os métodos divisivos

normalmente fazem uso de alguma heurística, podendo fazer com que o resultado não seja tão bom. Por outro lado, existem $\frac{n(n-1)}{2}$ possíveis combinações para unir dois dos n grupos, tornando os métodos aglomerativos viáveis mesmo para valores altos de n (KAUFMAN & ROUSSEEUW, 2005).

O método DIANA não analisa todas as possíveis formas de dividir um grupo com todos os objetos em dois outros grupos. Ao invés disso, ele utiliza um tipo de procedimento iterativo para decidir como deve ser feita a divisão. Este procedimento funciona da seguinte maneira, o grupo com maior diâmetro é selecionado para ser dividido. Diâmetro é definido como a maior distância ou dissimilaridade entre dois objetos dentro de um mesmo grupo. Selecionado o grupo G, a distância média (utilizada no AGNES) entre cada um dos seus objetos e o restante do grupo é calculada. O objeto que estiver mais distante é selecionado para sair do grupo e criar um novo grupo G'. O método executa uma iteração para montar o grupo G'. A iteração segue da seguinte forma, para cada objeto $o \in G$ o algoritmo calcula a diferença entre as distâncias médias D_{oG}, entre o e os demais objetos de G, e $D_{oG'}$, entre o e G'. Caso não haja nenhum objeto o tal que $D_{oG} - D_{oG'} > 0$, então o passo iterativo termina. Caso contrário, seleciona-se o objeto o com maior $D_{oG} - D_{oG'}$ para sair do grupo G e passar para o grupo G' e continua-se o passo iterativo. Quando o passo iterativo termina, um grupo é dividido em dois. Este processo deve ser repetido até que não seja mais possível dividir nenhum grupo, ou seja, até que cada objeto esteja em um grupo diferente.

2.6.4 – Outros métodos hierárquicos

Os métodos aglomerativos normalmente são mais utilizados que os divisivos por conta do custo computacional envolvido. Entretanto, os métodos divisivos são considerados mais seguros, pois escolhas ruins no início dos métodos aglomerativos não podem ser corrigidas mais tarde (STEINBACH *et al.*, 2000). Assim como ocorre com o AGNES, existem diversos métodos divisivos muito parecidos com o DIANA, mudando alguns detalhes, como a forma de calcular distância entre dois grupos.

Na literatura, é possível encontrar muitos outros métodos hierárquicos, como o BIRCH, ROCK, CURE e CHAMELEON, vistos em ZHANG *et al.* (1996), GUHA *et al.* (1999), GUHA *et al.* (1998) e KARYPIS *et al.* (1999), respectivamente.

2.7 – Outras categorias de métodos de agrupamento

Existem diversas formas de organizar os métodos de agrupamento em categorias. Além dos métodos de particionamento e hierárquicos, pode-se destacar os métodos baseados em densidade e GRID. Nessa seção, essas duas abordagens serão apresentadas e brevemente discutidas.

Métodos baseados em densidade: A maioria dos métodos de particionamento agrupam os objetos baseada na distância. Estes métodos encontram grupos com formatos hiperesféricos e têm dificuldade em encontrar grupos com formatos arbitrários. Assim, alguns métodos de agrupamento foram criados baseados na noção de densidade. A ideia destes métodos é que cada grupo deve continuar se expandindo, enquanto sua vizinhança tiver uma quantidade de objetos que exceda um limiar predefinido. Estes métodos podem ser utilizados para filtrar os valores atípicos (outliers) e descobrir grupos com formatos arbitrários. Três dos principais métodos dessa categoria são DBSCAN, OPTICS e DENCLUE, que podem ser encontrados respectivamente em ESTER *et al.* (1996), ANKERST *et al.* (1999) e HINNERBURG & KEIM (1998).

Métodos baseados em GRID: Estes métodos dividem o espaço de variáveis em uma quantidade finita de células formando uma estrutura de grade. Todas as operações de agrupamento são aplicadas nessa estrutura de grade. A principal vantagem desta abordagem é o rápido tempo de processamento, pois normalmente não depende da quantidade de objetos e sim da quantidade de células que formam a estrutura de grade. Dois exemplos típicos de métodos baseados em GRID são o STING e o CLIQUE, que podem ser encontrados em WANG *et al.* (1997) e AGRAWAL *et al.* (1998). Além disso, existem também os método WaveCluster, MAFIA e BANG, encontrados em SHEIKHOLESLAMI *et al.* (1998), GOIL *et al.* (1999) e SCHIKUTA & ERHART (1997).

2.8 – Sistema de gerenciamento de banco de dados espacial

Os dados podem ser de vários tipos, sendos mais comuns os dados numéricos, textuais e lógicos. Entretanto, em diversos problemas existe a necessidade de armazenar e manipular dados geométricos, geográficos ou espaciais. Identificar centros de atividades urbanas, de atividades criminais, de epidemias, são exemplos desses problemas (CHANDRA & ANURADHA, 2011). Em SHEKHAR *et al.* (1999) é possível ver que dados espaciais geralmente estão relacionados a:

- Espaço como o mundo físico (abordado por exemplo nas áreas da geografia, do planejamento urbano, da astronomia etc);
- Mapeamento de partes de organismos vivos (anatomia humana, por exemplo);
- Projetos de engenharia (circuitos integrados de grande escala, projeto de um automóvel, estrutura molecular de uma droga farmacêutica);
- Espaço de informação conceitual (um sistema multidimensional de suporte a decisão, fluxo de fluidos, campo eletromagnético).

Dentre esses, o exemplo de aplicação que mais se destaca é o mapeamento em duas dimensões de parte ou de toda superfície da Terra. Neste tipo de aplicação, geralmente o interesse é armazenar informações como a extensão e o limite de continentes, países ou regiões, hidrografia, vegetação, vias etc (MELLO, 2008).

Para modelar simples objetos espaciais, são necessários os três tipos de dados espaciais, que são ponto, linha e região (também chamado de polígono) (GUTING, 1994).

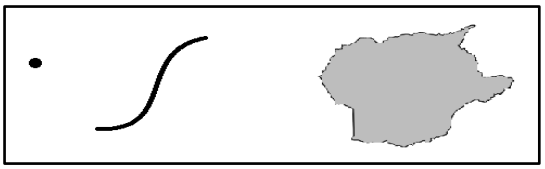


Figura 2.11 – Exemplos de objetos dos tipos ponto, linha e polígono

A Figura 2.11 ilustra os três tipos de dados espaciais. O ponto representa um objeto que apenas a sua localização no espaço é relevante, mas não sua extensão. A linha representa curvas, rotas ou conexões no espaço, como por exemplo rodovias, rios, cabos telefônicos, eletricidade etc. A região ou polígono representa algum objeto que

tenha extensão definida no espaço, como países, estados, oceanos etc.

Os três tipos de dados espaciais fornecem a base para a modelagem da estrutura espacial, sendo possível assim gerenciar suas relações (verificar se dois objetos espaciais se intersetam), suas propriedades (calcular a área de um objeto espacial) e executar operações entre objetos espaciais (encontrar a região de interseção entre dois objetos espaciais) (GUTING, 1994).

GUTING (1994) define os dois tipos mais importantes de coleções de objetos espacialmente relacionados, que são as partições e as redes. Uma partição pode ser entendida como um conjunto de objetos do tipo região, que são disjuntas no espaço, ou seja, não há sobreposição entre os objetos. Nessa coleção existem pares de regiões que são adjacentes, ou seja, que formam uma fronteira. Essa formação é uma topologia de vizinhança entre as regiões dessa coleção. As partições podem ser utilizadas para representar mapas temáticos, por exemplo. As redes podem ser entendidas como um conjunto de conexões, onde as linhas representam o caminho entre duas posições no espaço. Este tipo de coleção pode ser utilizado para representar rodovias, rios, rotas de transportes públicos etc. A Figura 2.12 ilustra os dois tipos de coleções.

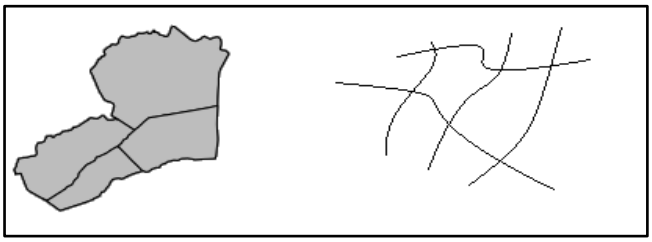


Figura 2.12 – Exemplos de coleções dos tipos partição e rede

Os sistemas de banco de dados espaciais funcionam como os banco de dados tradicionais, acrescentando o fato de trabalharem com dados espaciais (ESTER *et al.*, 2000). Esse sistema é geralmente utilizado em Sistemas de Informação Geográfica (SIG), auxiliando no armazenamento e no processamento dos atributos espaciais.

Os dados espaciais são codificados para que sejam armazenados em tabelas, junto com os demais dados. Como se trata de um novo tipo de dado, é necessário que um sistema de banco de dados espacial forneça suporte às consultas que utilizem operações sobre relações espaciais entre os objetos espaciais. De acordo com GUTING (1994) e ESTER *et al.* (2001) existem três tipos de relações espaciais entre objetos espaciais, que são as topológicas, de direção e de medida.

Relações topológicas são aquelas que descrevem como dois objetos se relacionam no espaço. Existem diversas relações topológicas como adjacente, dentro, disjunto e sobrepõe. Essas relações são independentes de transformações nos objetos, como translação, escala ou rotação.

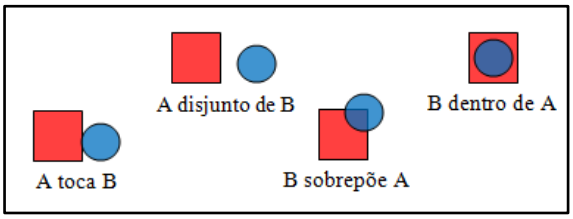


Figura 2.13 – Relações topológicas

Relações de direção indicam a direção em que um objeto se encontra em relação ao outro, como acima, abaixo, norte de, sul de etc.

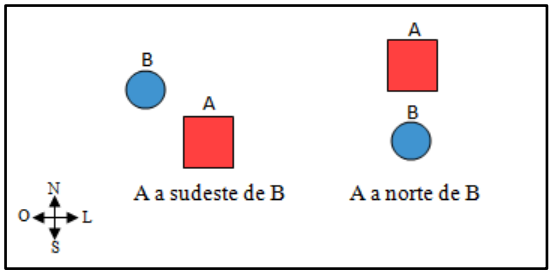


Figura 2.14 – Relações de distância

Relações de medida são aquelas que são baseadas em alguma métrica, como "$distância > 100$".

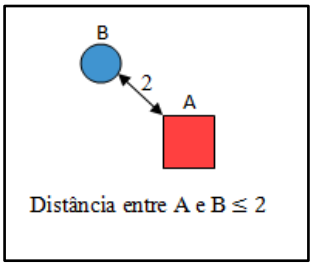


Figura 2.15 – Relação de medida

A partir dessas relações, é possível criar consultas complexas envolvendo dados espacais em um banco de dados espacial. Atualmente, existem diversos sistemas de banco de dados espaciais, como o PostgreSQL (POSTGRESQL, 2014) com a extensão PostGIS (POSTGIS, 2014).

2.9 – Agrupamento espacial

Na seção anterior, foram apresentados os conceitos de dado espacial e banco de dados espacial. Normalmente, quando dados espaciais são analisados, o interesse é buscar características que fogem do esperado, como por exemplo responder questões como "Existe uma estranha concentração de casos de leucemia perto da estação de energia nuclear?" (LAWSON & DENISON, 2002). Nesse caso, o foco é encontrar regiões no espaço (normalmente em duas dimensões) onde há uma concentração maior do que a esperada.

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). AMBROISE & DANG (2009) diz que um método de agrupamento espacial geralmente tem dois objetivos:

- Obter um particionamento de forma que cada grupo seja o mais homogêneo possível, ou seja, com maior similaridade possível entre os objetos do mesmo grupo;
- Fazer com que objetos contíguos dentro do espaço geográfico tenham uma probabilidade maior de pertencerem ao mesmo grupo.

A escolha da relação espacial no problema de agrupar objetos espaciais pode variar (MELLO, 2008). Enquanto relações de medida baseadas na distância já pode ser considerada uma noção natural para objetos espaciais do tipo ponto, pode ser mais apropriado utilizar relações topológicas baseadas em adjacência ou sobreposição para objetos do tipo região ou polígono com grandes diferenças de tamanho (ESTER *et al.*, 2001).

O agrupamento espacial pode ter duas abordagens diferentes com relação ao espaço de atributos. A primeira abordagem é agrupar os objetos espaciais baseando-se apenas nos atributos espaciais. Diversos métodos tradicionais de agrupamento são utilizados em agrupamento espacial utilizando-se dessa abordagem. Os métodos tradicionais normalmente utilizam a distância entre os objetos no espaço dos atributos não-espaciais como medida de dissimilaridade. Utilizando a relação espacial de medida

baseada em distância como dissimilaridade entre os objetos, os métodos passam a formar grupos cujos objetos são próximos uns dos outros. Em outras palavras, esses métodos definem áreas no espaço geográfico com alta densidade de objetos espaciais (ESTER *et al.*, 2001). A outra abordagem é agrupar os objetos espaciais considerando tanto os atributos espaciais como os não-espaciais.

Os métodos de particionamento mais encontrados na literatura utilizados em agrupamento espacial são o PAM, o CLARA e o CLARANS (MELLO, 2008). NG & HAN (1994) criaram duas variações do CLARANS, o (SD)CLARANS e o (NSD)CLARANS, que utilizam tanto os atributos não-espaciais como os espaciais para agrupar os objeos espaciais.

Existem, ainda, métodos em que o relacionamento de vizinhança entre os objetos é utilizado no algoritmo por meio de dispositivos auxiliares, como uma matriz, um grafo ou listas de objetos vizinhos (NEVES, 2003). Dois exemplos de métodos que seguem essa abordagem são o AZP, que usa lista de objetos vizinhos (OPENSHAW & RAO, 1994), e o método da árvore geradora mínima, que utiliza um grafo para guardar a topologia dos objetos (ASSUNÇÃO *et al.*, 2002).

2.9.1 – Métodos baseados em atributos espaciais

Os métodos convencionais de agrupamento podem ser utilizados em agrupamento espacial. Para isso, basta considerar que a medidade de dissimilaridade entre dois objetos espaciais é a distância entre eles no espaço geográfico, ao invés de utilizar a distância no espaço dos atributos não-espaciais. Existem diversas formas de calcular a distância entre dois objetos espaciais (NG & HAN, 2002). Entretanto, a maneira mais simples é representar os objetos espaciais como objetos do tipo ponto e utilizar medidas de distância, como as discutidas na seção 2.2.1, para calcular a distância entre dois objetos.

NG & HAN (1994,2002) apresentam experimentos comparando o desempenho dos métodos de particionamento PAM, CLARA e CLARANS agrupando dados espaciais do tipo ponto. A Figura 2.16 ilustra o agrupamento de pontos próximos no espaço de duas dimensões.

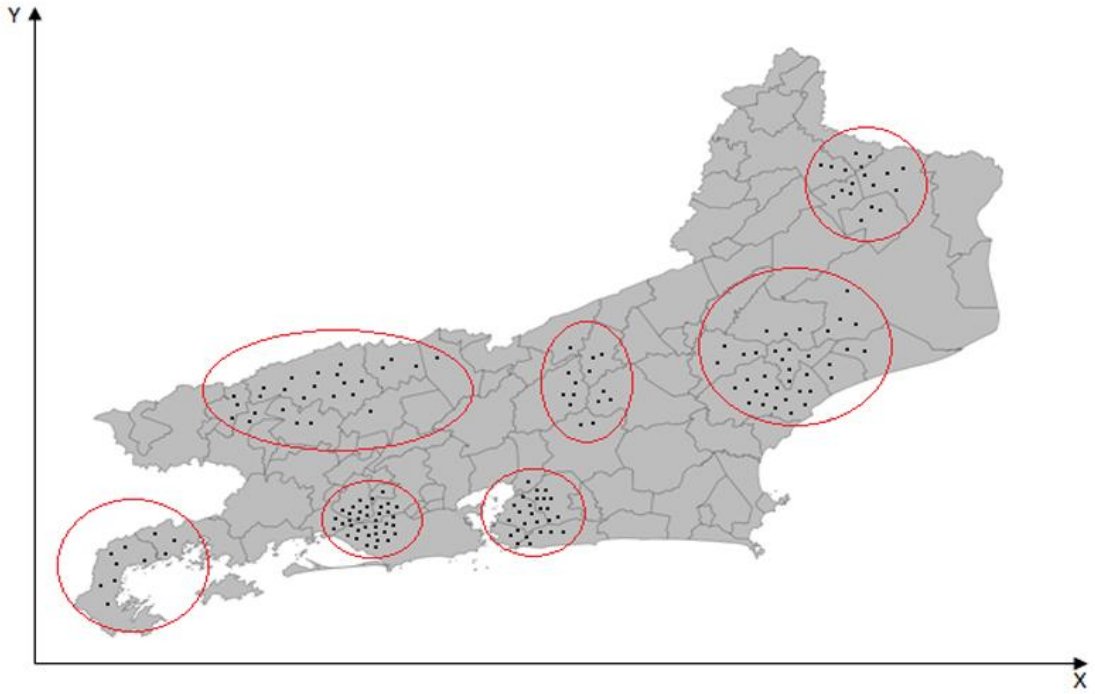


Figura 2.16 – Agrupamento de pontos no espaço

Representar objetos espaciais por pontos funciona bem nos casos que informações espaciais como extensão e forma não são relevantes (MELLO, 2008). Entretanto, normalmente os objetos espaciais podem possuir uma grande variedade de tamanhos e formas, e representar os objetos por um ponto, pode facilmente produzir agrupamentos de baixa qualidade (NG & HAN, 2002). A solução para isso seria utilizar objetos do tipo polígono.

Os métodos de agrupamento espacial, baseados apenas em atributos espaciais, normalmente requerem calcular a distância entre os objetos como forma de dissimilaridade. NG & HAN (2002) apresentam três formas de calcular distância entre dois polígonos convexos. A primeira é calculando a distância exata entre dois polígonos. A segunda é calculando a distância mínima entre os vértices dos dois polígonos. A terceira é encontrando o menor retângulo possível que contém por completo cada um dos polígonos e calcular a distância exata entre os retângulos encontrados.

Em NG & HAN (2002), é possível encontrar experimentos com as três formas de calcular a distância entre polígonos sendo utilizadas pelos métodos PAM, CLARA e

CLARANS.

2.9.2 – Métodos baseados em atributos espaciais e não-espaciais

Existem diversos tipos de abordagens para utilização de atributos espaciais junto com não-espaciais em agrupamento espacial. Uma delas é representar os objetos espaciais como pontos e adicionar as coordenadas dos pontos aos atributos não-espaciais dos objetos. Nesse caso, normalmente os atributos não-espaciais e os espaciais possuem pesos diferentes no cálculo da dissimilaridade. O problema dessa abordagem é que se o peso dos atributos espaciais não for alto o suficiente, os grupos resultantes podem não ser contíguos (NEVES, 2003). O método de agrupamento espacial utilizado pelo sistema SAGE segue essa abordagem, como pode ser visto em (NEVES, 2003).

Outras duas abordagens são encontradas em NG & HAN (1994), são elas dominante espacial e dominante não-espacial. Nelas, o método de particionamento CLARANS é utilizado em conjunto com o DBLEARN (HAN *et al.*, 1992), que é uma ferramenta que contrói regras a partir da generalização de atributos não espaciais num banco de dados.

Na abordagem do dominante espacial, o método, chamado SD(CLARANS), encontra um particionamento baseado nos atributos espaciais, utilizando o CLARANS, e em seguida divide cada grupo baseado nos atributos não-espaciais, utilizando o DBLEARN. Por outro lado, na abordagem dominante não-espacial, o método chamado NSD(CLARANS) funciona de forma inversa, aplicando primeiro o DBLEARN e depois o CLARANS. Nada impede que outros métodos de particionamento espacial sejam usados, gerando assim SD(CLARA) ou NSD(PAM), por exemplo (NG & HAN, 1994).

2.9.3 – Métodos baseados em relações topológicas

Os métodos espaciais descritos até agora baseiam-se na relação espacial de medida, calculando a distância entre os objetos como medida de dissimilaridade entre eles. Por vezes, quando os métodos de agrupamento espacial são utilizados, a distância entre os objetos espaciais dentro de um mesmo grupo não são relevantes, mas sim se os grupos são contíguos. ASSUNÇÃO *et al.* (2002) definiram que regionalização é um processo de classificação aplicado a um conjunto de objetos espaciais de forma que os

grupos sejam homogêneos e contíguos. Portanto, normalmente quando o objetivo é aplicar uma regionalização, a informação da distância entre os objetos espaciais não é relevante, mas sim a topologia que os envolve.

Um método que utiliza a relação topológica entre os objetos é o AZP (Automatic Zoning Procedure). Este algoritmo executa os sete passos a seguir (OPENSHAW & RAO, 1994):

1. Comece gerando uma partição aleatória dos n objetos espaciais em k grupos contíguos;
2. Crie uma lista com os k grupos;
3. Selecione aleatoriamente um grupo G_i e o remova da lista;
4. Crie uma lista de todos os objetos vizinhos do grupo G_i e que possam ser removidos de seus respectivos grupos, sem que eles deixem de ser contíguos;
5. Selecione aleatoriamente objetos da lista de vizinhos até que algum deles faça com que a qualidade do particionamento melhore ao ingressar no grupo G_i (volte para o passo 4) ou a lista termine (siga para o passo 6);
6. Quando não houver mais vizinhos na lista, retorne até o passo 3;
7. Repita os passos de 2 a 6 até que não haja mais nenhuma realocação que resulte em melhora na qualidade do particionamento;

Métodos baseados em tentativa e erro normalmente não apresentam um bom tempo de processamento. Entretanto, existem na literatura algumas melhorias para este método (ASSUNÇÃO *et al.*, 2002).

Outro método baseado em relações topológicas é método da árvore geradora mínima. Este método pode ser chamado de SKATER (Spatial ‘K’luster Analysis by Tree Edge Removal) (ASSUNÇÃO *et al.*, 2002). O SKATER é um método de agrupamento espacial desenvolvido para solucionar problemas de regionalização.

O SKATER usa um grafo de conectividade para armazenar a relação topológica entre os objetos espaciais. Em grafo de conectividade, cada objeto é representado por um vértice do grafo e, quando os dois objetos são vizinhos, existe uma aresta ligando os vértices que os representam (NEVES *et al.*, 2002). No SKATER, o custo de cada aresta é dado por um valor proporcional à dissimilaridade entre os objetos vizinhios, que é calculada a partir dos atributos não-espaciais. Cortando o grafo nos lugares adequados, o resultado são grupos contíguos. Dessa forma, o problema de regionalização é

transformado em um problema de particionamento ótimo de grafos. O particionamento ótimo de grafos é um problema NP-Difícil – detalhes sobre complexidade computacional podem ser encontrados em ARORA & BARAK (2009) – e, por conta disso, seria necessário uma heurística aplicável a um grande conjunto de dados espaciais para alcançar soluções subótimas em um tempo aceitável. Para limitar a complexidade do grafo, o método encontra uma árvore geradora mínima (AGM) do grafo de conectividade. Depois de criar a árvore geradora mínima, o método particiona a árvore encontrando grupos de regiões.

O método de agrupamento com restrição de contigüidade via AGM está implementado na ferramenta SKATER (SKATER, 2014). 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.

2.10 – Teste de hipótese

Geralmente, não é possível se obter o conjunto completo dos dados para fazer análises. Por exemplo, para analisar dados referentes às pessoas, seria necessário recolher informações das mais de 7 bilhões de pessoas que existem no planeta atualmente, sendo necessário criar mecanismos para fazer estimativas com base em uma parte do conjunto completo dos dados. A Inferência Estatística é um ramo da Estatística que estuda formas de fazer inferências sobre a população (conjunto completo dos dados) com base em amostras (subconjunto da população) (GOEL & SHANKAR, 2012).

Um dos principais métodos de inferência estatística é o teste de hipótese. CASELLA & BERGER (1990) definem que hipótese é uma afirmação sobre algum parâmetro da população e que o objetivo de um teste de hipótese é decidir, a partir de uma amostra da população, qual das duas hipóteses complementares é a verdadeira. As hipóteses complementares num problema de teste de hipótese são chamadas de hipótese nula e hipótese alternativa, que são denotadas respectivamente por H_0 e H_1. Assumindo que s é um medida da população, o formato padrão das hipóteses é $H_0: s \in P_0$ e $H_1: s \in P_0^c$, onde P_0 é um subconjunto de todos os possíveis valores de s e P_0^c é o complemento de P_0. Por exemplo, supondo que s represente a média de altura dos

homens do estado do Rio de Janeiro e o problema seja tentar descobrir se os homens possuem altura média diferente de 1,80 metros. Nesse caso, as hipóteses são denotadas por $H_0: s \neq 1{,}80$ e $H_1: s = 1{,}80$. O teste de hipótese é uma regra que especifica:

1. Para quais amostras a decisão é aceitar H_0 como verdadeira;
2. Para quais amostras H_0 é rejeitada e H_1 é aceita como verdadeira.

O subconjunto do espaço de amostras em que H_0 é rejeitada é chamado de região de rejeição ou região crítica. O complemento da região de rejeição é chamada de região de aceitação. No exemplo anterior, supondo que H_0 deva ser rejeitada caso a altura média da amostra da população seja $\leq 1{,}70$ ou $\geq 1{,}90$, as regiões de rejeição e aceitação serão como mostrado na Figura 2.17:

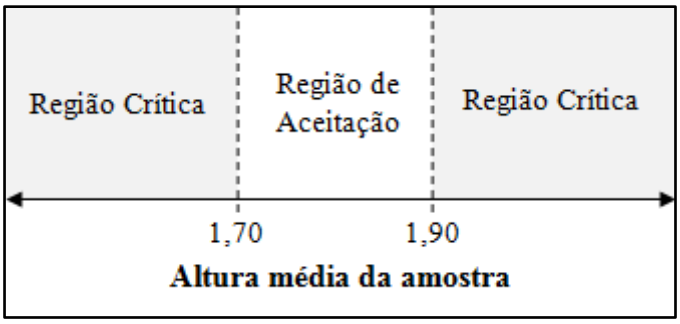


Figura 2.17 – Região Crítica

Quando um teste é realizado ou a decisão correta foi alcançada ou um dos seguintes erros pode ter sido cometido: rejeitar a hipótese nula quando ela é verdadeira (erro do tipo I) ou aceitar a hipótese nula quando ela é falsa (erro do tipo II) (LEHMANN & ROMANO, 2006). As consequências dos erros geralmente são bastante diferentes. Por exemplo, se alguém testando a presença de alguma doença incorretamente decidir que existe a necessidade de um tratamento, pode causar desconforto no paciente e gastos desnecessários. Por outro lado, se houver falha em diagnosticar a presença da doença isso pode levar a morte do paciente.

O ideal seria executar o teste de forma a manter a probabilidade dos dois tipos de erros tão baixas quanto possível. Entretanto, dado uma amostra da população, não é possível controlar as duas probabilidades ao mesmo tempo. Dessa forma, é habitual atribuir um limite para probabilidade de rejeitar incorretamente a hipótese nula. Assim, um nível de significância α, que é um número real entre 0 e 1, é definido de forma que

a probabilidade da hipótese nula ser rejeitada incorretamente seja menor do que α (LEHMANN & ROMANO, 2006).

A escolha do nível de significância é de certa forma arbitrária, já que na maioria das situações não há como precisar um limite da probabilidade de um erro do tipo I que seja tolerável (LEHMANN & ROMANO, 2006). Existem alguns valores padrões, como 0.01 ou 0.05, que são escolhidos para evitar a necessidade de tabelas na execução de diversos testes.

Depois que o teste de hipótese é criado, é necessário algum procedimento para tomar a decisão, se a hipótese nula deve ou não ser rejeitada. Uma forma comum de tomar essa decisão, é utilizando o p-valor, que pode ser comparado diretamente com o nível de significância. O p-valor é a probabilidade de, dado que a hipótese nula é verdadeira, se obter uma amostra cuja medida seja no mínimo tão distânte do intervalo definido pela hipótese nula quanto a atual. Por exemplo, supondo que em um teste de hipótese deseja-se verificar se uma moeda é viciada ou não, a hipótese nula seria que a probabilidade de cair cara é de $1/2$ e a alternativa é que a probabilidade seria diferente disso. Se a moeda foi lançada 10 vezes e só uma vez caiu cara, então o p-valor seria a probabilidade da moeda cair no máximo uma vez cara (1 cara ou 0 caras em 10 lançamentos), dado que a moeda não é viciada. Nesse caso, o p-valor seria aproximadamente 0.011 ou 1.1%.

Após calcular o p-valor, o próximo passo para decidir se a hipótese nula deve ou não ser rejeitada é fazer uma comparação entre ele e o nível de significância escolhido. De maneira simples, se o p-valor for menor do que o nível de significância, então a hipótese nula deve ser rejeitada, caso contrário, a hipótese nula não é rejeitada.

Na literatura, existem diversos testes de hipóteses que se diferem pela medida que é utilizada como base para criar as hipóteses que são testadas. Entretanto, como a dissertação se baseou apenas no teste de Kolmogorov-Smirnov, este será o alvo da próxima seção.

2.10.1 – Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov é um teste de hipótese não-paramêtrico, tendo em vista que não depende de nenhum parâmetro da distribuição da população. A

estatística de Kolmogorov-Smirnov é calculada a partir da distância entre as funções de distribuição acumulada (fda) de duas distribuições diferentes (LEHMANN & ROMANO, 2006), como mostrado a seguir:

$$D = max_X(|F_1(X) - F_2(X)|)$$

Onde:

- D é a distância entre as duas fda's;

- $F_1(X)$ e $F_2(X)$ são as duas fda's diferentes;

No teste de Kolmogorov-Smirnov, a hipótese nula é que $F_1 = F_2$ e a hipótese alternativa é que $F_1 \neq F_2$. Este teste pode ser de dois tipos, de uma amostra ou de duas amostras. No teste de uma amostra, F_1 é a fda empírica da amostra em questão e F_2 é a fda de uma distribuíção já conhecida. De maneira simples, este teste tem por objetivo decidir se a amostra foi gerada de uma população cuja fda é F_2.

A Figura 2.18 mostra um exemplo de um teste que se deseja decidir se uma amostra foi retirada de uma população com distribuição normal. Nela é possível ver a fda empírica da amostra e a fda da distribuição normal.

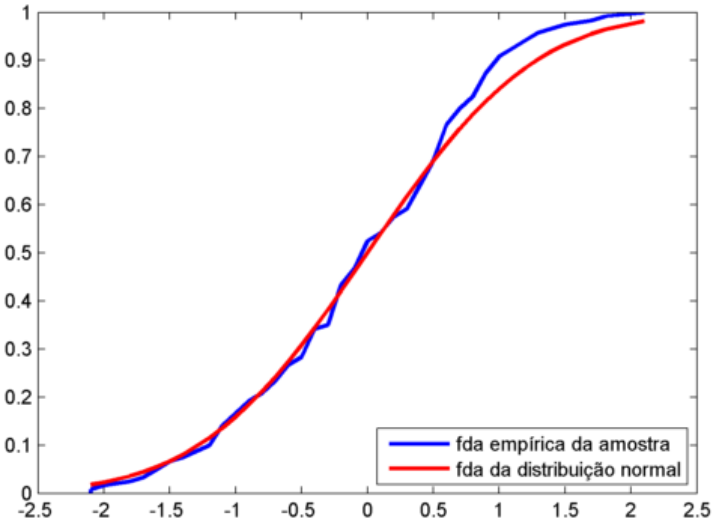


Figura 2.18 – Comparação entre as fda's de uma amostra e da distribuição normal

No teste de duas amostras, o objetivo é testar se as duas amostras foram geradas da mesma população ou de populações que possuam a mesma distribuição. Nesse caso, F_1 e F_2 são fda's empíricas de cada uma das amostras. A Figura 2.19 mostra as fda's de

cada uma das amostras que são base do teste.

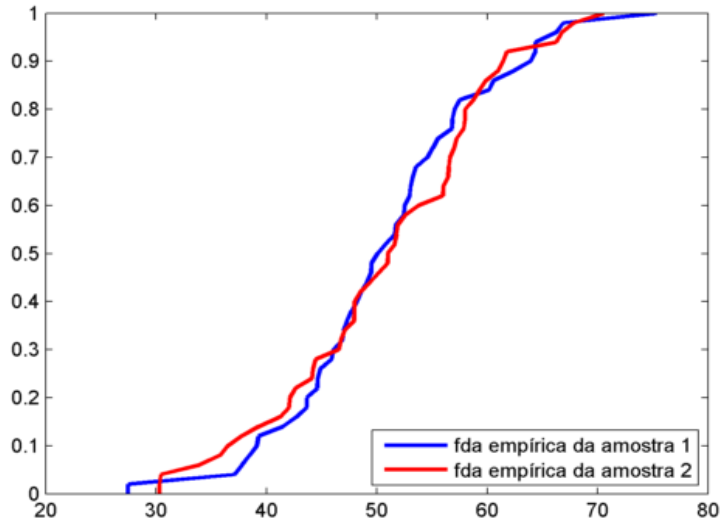


Figura 2.19 – Comparação entre as fda's de uma amostra e da distribuição normal

O p-valor do teste de Kolmogorov-Smirnov de uma amostra pode ser interpretado da seguinte maneira. Dado que A é a amostra e F é a distribuição que são base do teste, o p-valor seria a probabilidade de se tirar uma amostra A^* de uma população que tenha distribuição F, com distância de Kolmogorov-Smirnov maior ou igual a calculada para A. Para o teste de duas amostras, o p-valor pode ser interpretado como a probabilidade de se obter outras duas amostras com distância de Kolmogorov-Smirnov maior ou igual ao valor atual, dado que as duas tenham sido retiradas de populações com a mesma distribuição.

Alguns métodos computacionais, que implementam os testes de uma ou duas amostras, calculam o p-valor a partir da distribuição assintótica a seguir (STATA, 2014):

$$\lim_{m,n\to\infty} Pr\left\{\sqrt{mn/(m+n)}\, D_{m,n} \leq z\right\} = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} exp(-2i^2z^2)$$

No geral usa-se os primeiros termos do somatório para alcançar o nível de aproximação desejado. É possível utilizar métodos númericos de aproximação para chegar bem perto do valor exato com um custo computacional acessível, como pode ser visto em (STATA, 2014).

Entretanto, de forma simples, utiliza-se os valores da tabela de Kolmogorov-

Smirnov (Tabela 2.2) para decidir se a hipótese nula deve ser aceita ou não. Caso a distância D seja menor do que o valor obtido pela tabela, a hipótese nula deve ser aceita e, caso contrário, ela deve ser recusada (RAZALI & WAH, 2011).

Tabela 2.2 – Tabela de Kolmogorov-Smirnov

	$\boldsymbol{\alpha = 0.05}$	$\boldsymbol{\alpha = 0.01}$		$\boldsymbol{\alpha = 0.05}$	$\boldsymbol{\alpha = 0.01}$
$\boldsymbol{n = 5}$	0.565	0.669	$\boldsymbol{n = 30}$	0.240	0.290
$\boldsymbol{n = 10}$	0.410	0.490	$\boldsymbol{n = 40}$	0.210	0.250
$\boldsymbol{n = 15}$	0.338	0.404	$\boldsymbol{n = 50}$	0.190	0.230
$\boldsymbol{n = 20}$	0.294	0.356	$\boldsymbol{n} > 50$	$1.36/\sqrt[2]{n}$	$1.63/\sqrt[2]{n}$

2.11 – Considerações finais

Neste capítulo, foi realizada uma revisão bibliográfica sobre análise de agrupamento e testes de hipóteses. Dentro de análise de agrupamento, foi apresentado o problema do agrupamento aplicado a uma base de dados espaciais. Os métodos para resolver esse tipo de problema, chamados de métodos de agrupamento espacial, na maioria da vezes utilizam a relação espacial de distância entre os objetos. Entretanto, existem métodos que utilizam a relação topológica para criar o agrupamento. Estes fazem uso de estruturas, como árvores ou matrizes, para armazenar as relações topológicas entre os objetos. O SKATER é um exemplo desse tipo de algoritmo, que utiliza uma estrutura de árvore para armazenar as relações topológicas. Este método foi apresentado e explicado na seção 2.9 e será mais explorado no próximo capítulo, por se tratar do método utilizado no experimento realizado e descrito nesta dissertação.

Na revisão feita sobre testes de hipóteses, foram apresentados os principais conceitos sobre essa técnica estatística de inferência. A estatística de Kolmogorov-Smirnov foi apresentada, assim como o teste de hipótese baseado nela. Esse teste pode ser de duas amostras, de maneira que a hipótese a ser testada é se as duas amostras foram obtidas de populações com a mesma distribuição. O teste de Kolmogorov-Smirnov, que foi apresentado na seção 2.10, será base de uma nova medida de similaridade apresentada no próximo capítulo.

Capítulo 3 – Agrupamento espacial de amostras utilizando teste de Kolmogorov-Smirnov para duas amostras

3.1 – Introdução

Neste capítulo, o algoritmo do agrupamento espacial de amostras utilizando teste de Kolmogorov-Smirnov para duas amostras será descrito. O objetivo desta dissertação é obter um método capaz de agrupar elementos comparando suas amostras a partir de um teste de hipóteses.

No capítulo anterior, foi explicado que, em problemas nos quais se deseja agrupar regiões que possuem amostras de variáveis aleatórias como atributos, costuma-se aplicar o algoritmo de agrupamento levando em conta a média amostral. Este tipo de problema poderia ser resolvido de forma mais precisa, caso as informações provenientes das amostras fossem melhor aproveitadas. É possível notar que, ao analisar apenas a média, todo o conhecimento sobre distribuição é perdido, como pode ser visto na Figura 3.1.

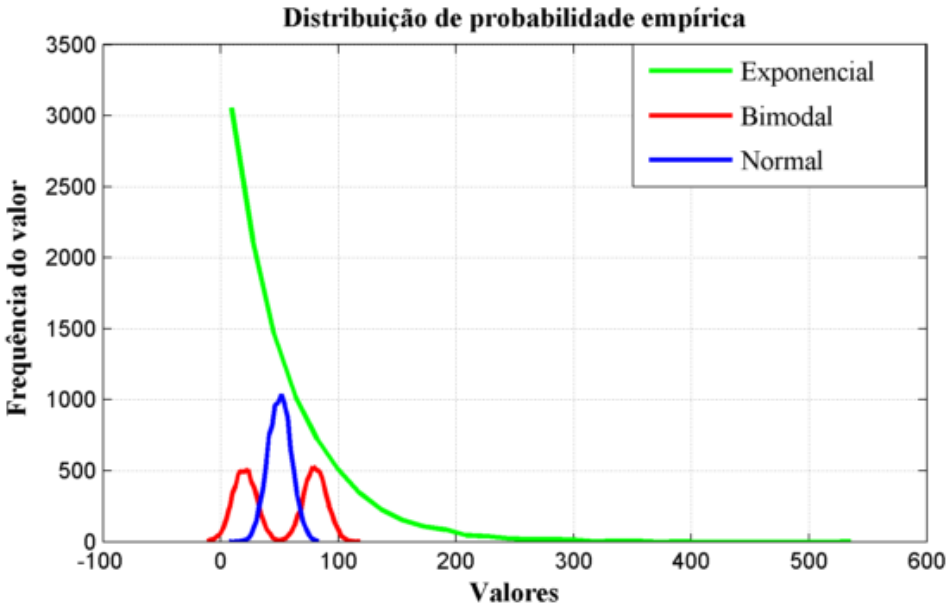


Figura 3.1 – Três distribuições diferentes com a mesma média

A Figura 3.1 mostra três distribuições bem distintas, mas com médias bem próximas. A curva verde é referente a uma amostra de tamanho 10000 gerada a partir de uma distribuição exponencial com média 50, a vermelha é de uma amostra de uma bimodal com modas em 20 e 80 e a azul foi gerada de uma normal com média 50 e desvio padrão 10. As três amostras descritas possuem médias bem parecidas, porém foram retiradas de populações bem distintas. Num problema em que se quer agrupar elementos com amostras semelhantes às descritas acima, estes ficariam em um mesmo grupo por conta da proximidade de suas médias. Entretanto, as diferenças encontradas em suas distribuições indicam que suas características provavelmente diferem, portanto os elementos não deveriam ficar no mesmo grupo.

Isso mostra que, se as comparações fossem feitas através das distribuições ao invés das médias amostrais, o resultado seria muito mais preciso. Nas próximas seções, será apresentada a proposta desta dissertação para tentar melhorar a forma como essas comparações são feitas, utilizando o teste de Kolmogorov-Smirnov para duas amostras. Além disso, será explicado como foi feita a implementação da API em Java, ferramenta utilizada na execução dos experimentos de validação da proposta.

3.2 – Teste de Kolmogorov-Smirnov como medida de similaridade entre duas amostras

Na seção anterior, foi possível observar que, seria mais eficiente comparar duas amostras utilizando a informação da distribuição ao invés da média. Para isso, seria necessário utilizar um método que, a partir de duas amostras, retorne uma medida que indique o grau de semelhança entre a distribuição de cada uma delas.

O teste de Kolmogorov-Smirnov para duas amostras é um método que auxilia na decisão de considerar se ambas foram extraídas da mesma população ou não. Neste teste, a hipótese nula diz que as amostras pertencem à mesma população. Portanto, o p-valor é a probabilidade de duas amostras retiradas da mesma população serem, no mínimo, tão diferentes quanto as que estão sendo testadas. Na prática, isto significa que se o p-valor calculado é muito próximo de zero, pode-se considerar que as duas amostras foram geradas de populações diferentes. Analogamente, se o p-valor é muito

perto de um, então maior é a chance de que pertençam a mesma população.

Neste contexto, o p-valor calculado a partir do teste de Kolmogorov-Smirnov para duas amostras pode ser utilizado como uma medida de similaridade. Em algoritmos de agrupamento, para comparar dois elementos que se deseja agrupar, utiliza-se uma medida ou função de similaridade. Como o próprio nome já diz, quanto maior a similaridade entre os elementos, mais parecidos eles serão. Por outro lado, se a similaridade for menor, mais distintos eles serão. Desta forma, é possível identificar quais objetos devem ser alocados em um mesmo grupo.

A seguir será aplicado o teste de Kolmogorov-Smirnov para as amostras exemplificadas na seção anterior.

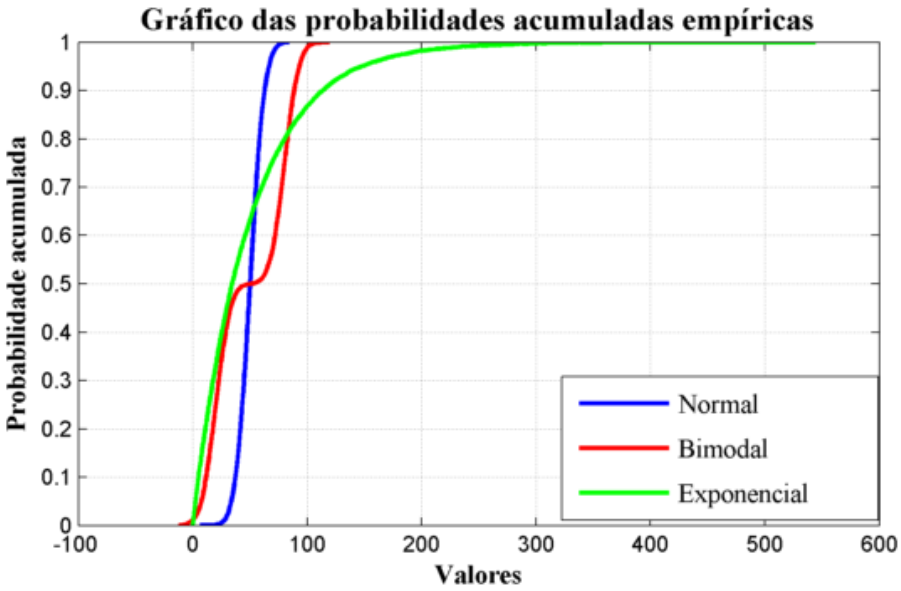


Figura 3.2 – Função de distribuição empírica de três amostras diferentes com a mesma média

A Figura 3.2 mostra o gráfico das funções de distribuição empírica (FDE) das três amostras geradas no exemplo da Figura 3.1. Como foi dito anteriormente, o teste de Kolmogorov-Smirnov considera a diferença entre as FDEs para calcular os p-valores, o que leva a crer que eles serão bem baixos, devido a grande diferença apresentada.

Tabela 3.1 – Matriz de p-valores

	Exponencial	Normal	Bimodal
Exponencial	1	0	$2.3740 * e^{-168}$
Normal	0	1	0
Bimodal	$2.3740 * e^{-168}$	0	1

A Tabela 3.1 mostra os p-valores resultantes da aplicação do teste de hipóteses de Kolmogorov-Smirnov nas amostras geradas anteriormente. É fácil notar que o p-valor para as mesmas amostras deve ser igual a 1, pois não há distância entre as FDEs. Mesmo com curvas bem diferentes, o p-valor das amostras exponencial e bimodal não é zero, porém a probabilidade é tão baixa que, mesmo com nível de significância de 0.5% (normalmente utiliza-se entre 1% e 5%), a hipótese nula não seria aceita, indicando que as amostras não poderiam ser considerados da mesma população. Os demais resultados são nulos, pois as distribuições das três amostras são bem diferentes, apesar de terem médias parecidas.

3.3 – Proposta

A proposta desta dissertação é a criação de um método capaz de fazer um agrupamento de objetos espaciais utilizando o p-valor resultante do teste de Kolmogorov-Smirnov como medida de similaridade. O algoritmo de agrupamento espacial utilizado na implementação foi o SKATER (ASSUNÇÃO *et al.*, 2002). A seguir serão descritas as etapas deste método.

3.3.1 – Matriz de Similaridade

O primeiro passo que o método deve executar é calcular a matriz de similaridade, responsável por armazenar a proximidade entre dois elementos quaisquer, a partir dos dados de entrada do problema. A partir desta informação, é possível identificar quais são os objetos que são mais parecidos e os mais distintos. No geral, o valor máximo da similaridade é 1 e o mínimo é 0.

A matriz de similaridade é uma matriz quadrada, o número de linhas é o mesmo que o de colunas, e também simétrica, pois o elemento 1 é tão similar ao elemento 2 quanto o 2 é do 1, não importando a ordem que a comparação é feita. A diagonal principal sempre é preenchida com 1, já que representa a comparação entre cada elemento com ele próprio.

A medida ou função de similaridade definida na implementação é utilizada para preencher a matriz de similaridade. A matriz nada mais é do que uma representação da medida de similaridade no domínio do problema, armazenando a comparação de cada elemento com todos os outros.

Como foi dito anteriormente, o objetivo desta dissertação é obter um método capaz de comparar objetos a partir de suas variáveis aleatórias de interesse, utilizando a distribuição da amostra ao invés da média. Para isso, foi definido que a medida de similaridade utilizada para preencher a matriz de similaridade é o p-valor calculado a partir do teste de Kolmogorov-Smirnov. Assim, a matriz é preenchida com probabilidades, fazendo com que o maior valor possível de similaridade seja 1 e o menor, 0.

Para o exemplo dado na seção 3.1, considerando que cada uma das amostras pertence a elementos diferentes, a matriz de similaridade é justamente a Tabela 3.1.

3.3.2 – Matriz Topológica

Em problemas de agrupamento espacial, normalmente, deseja-se levar em conta as restrições geradas a partir da disposição espacial dos elementos. Por exemplo, muitas vezes não é interessante ter, em um mesmo agrupamento, elementos que estejam espacialmente muito distantes, mesmo que eles sejam parecidos. A matriz topológia é responsável por armazenar as informações relativas às disposições espaciais dos elementos.

A matriz topológica geralmente é uma matriz binária, quadrada e simétrica, na qual cada linha e cada coluna representa um dos n elementos do problema. Cada célula com 1 indica que existe uma ligação entre os elementos representados pela linha e pela coluna e com 0, indica a ausência de ligação entre eles.

Na implementação desta proposta, a matriz topológica representa a contiguidade entre dois elementos, ou seja, caso os elementos sejam vizinhos haverá um 1 na célula que representa a ligação entre os dois e, caso contrário, haverá um 0.

Voltando novamente ao exemplo da seção 3.1, supondo que as amostras pertençam à três elementos diferentes e que estes estejam organizados geograficamente como mostra a Figura 3.3 – onde o elemento com a distribuição exponencial é a região em verde, o com distribuição normal é a região azul e a bimodal é a vermelha – a matriz topológica com base na contiguidade entre os elementos é representada na Tabela 3.2.

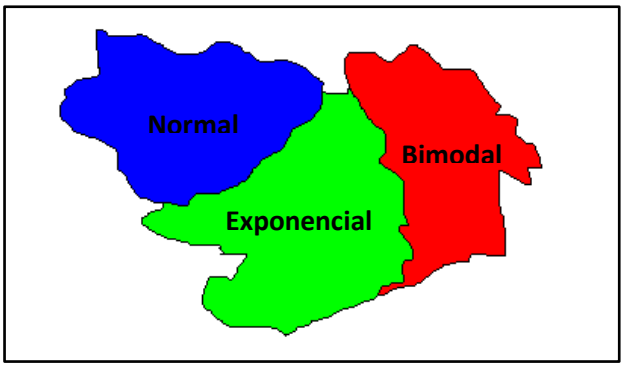


Figura 3.3 – Organização geográfica do exemplo da seção 3.1

Tabela 3.2 – Matriz topológica do exemplo da seção 3.1

	Exponencial	Normal	Bimodal
Exponencial	1	1	1
Normal	1	1	0
Bimodal	1	0	1

3.3.3 – Algoritmo de Agrupamento espacial

O algortimo de agrupamento espacial utilizado na implementação foi uma adaptação do SKATER (ASSUNÇÃO *et al.*, 2002). Entretanto, como a proposta é fundamentada na aplicação do teste de Kolmogorov-Smirnov como medida de similaridade, qualquer algoritmo de agrupamento espacial pode ser utilizado.

O SKATER executa os três seguintes passos para realizar o agrupamento espacial:

1. Gerar o grafo representativo do problema;
2. Calcular a árvore geradora mínima;
3. Fazer cortes na árvore até que haja K componentes conexas;

3.3.3.1 – Gerar o grafo representativo do problema

Um grafo é definido por um conjunto de objetos, chamados vértices, e um conjunto de arestas, que representa a forma como esses objetos se relacionam.

O primeiro passo do algoritmo é gerar um grafo que represente o problema com as informações espaciais e não-espaciais. Neste grafo, cada vértice simboliza um dos elementos de interesse, e as arestas representam a forma como esses elementos se relacionam entre si. As arestas não possuem orientação, e possuem pesos que representam a distância entre os vértices. Existe uma aresta entre dois vértices se, e somente se, existir uma relação entre os elementos representados por estes vértices na matriz topológica.

Uma vez que a similaridade entre dois elementos pode ser traduzida como a proximidade entre eles, a dissimilaridade pode ser considerada a distância. Desta forma, o peso da aresta que os liga pode ser dado pela dissimilaridade. A dissimilaridade entre dois elementos pode ser calculada pela dissimilaridade máximo menos a similaridade entre eles. Caso a similaridade seja 0.7 e a dissimilaridade máxima seja 1, a aresta terá peso $1 – 0.7 = 0.3$.

A seguir será mostrado um exemplo de como deve ser gerado o grafo representativo do problema, supondo que no problema os elementos estejam organizados geograficamente conforme mostrado na Figura 3.4.

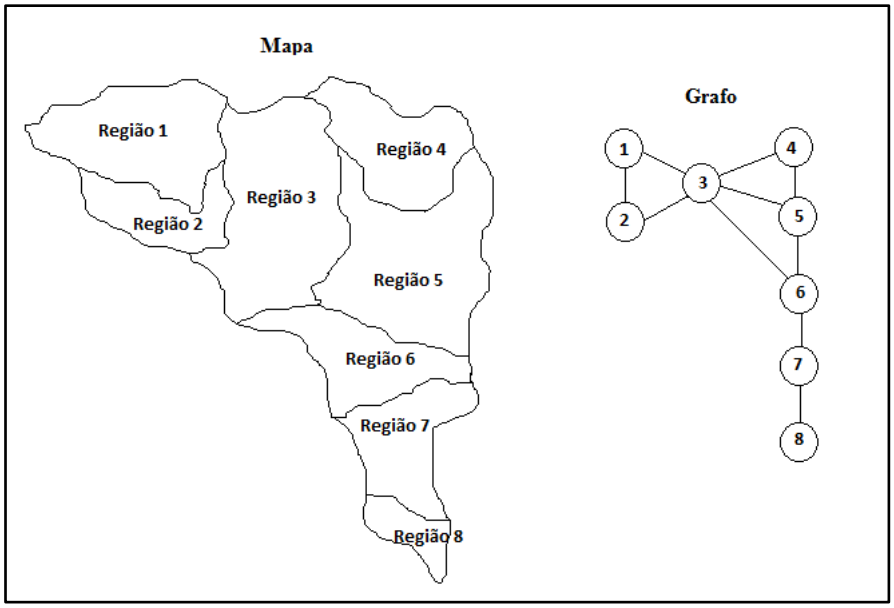


Figura 3.4 – Geração do grafo representante do mapa

3.3.3.2 – Calcular a árvore geradora mínima

Em seguida, o algoritmo calcula a árvore geradora mínima do grafo gerado no passo anterior. Existem diversos métodos que calculam a árvore geradora mínima de um grafo, sendo os mais utilizados, o de Prim e de Kruskal. Na implementação, foi utilizado o algoritmo de Kruskal.

A árvore geradora mínima é o subgrafo conexo contendo todos os vértices que possuem o menor custo total. Isso significa que, a soma dos custos de todas as arestas é o menor possível para um subgrafo conexo que possui todas os vértices do grafo original. Portanto, como o objetivo é encontrar os elementos com menores valores de dissimilaridade entre si, os pesos das arestas foram preenchidos com a medida da dissimilaridade entre os elementos.

A seguir, será mostrada a sequência do exemplo iniciado no passo anterior, para ilustrar o como é gerada a árvore geradora mínima. Os pesos das arestas foram adicionados aleatoriamente, apenas para mostrar o funcionamento do algoritmo.

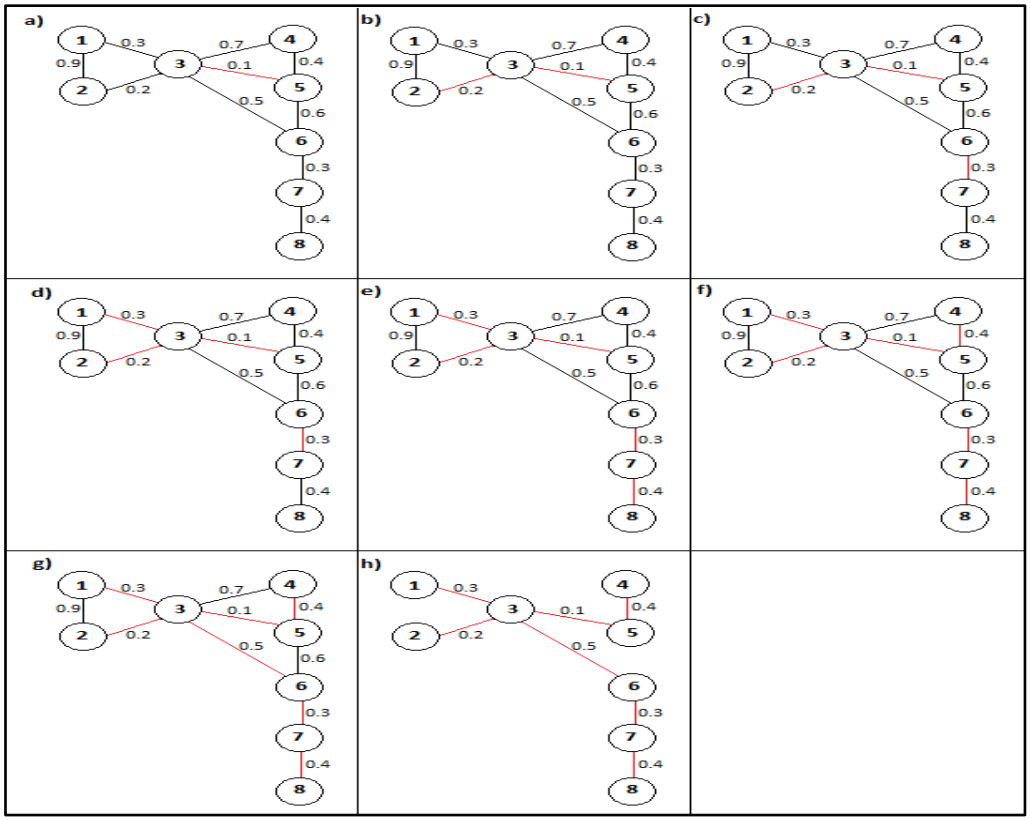


Figura 3.5 – Execução do algoritmo de Kruskal

3.3.3.3 – Fazer cortes na árvore até que haja K componentes conexas

Finalmente, o algoritmo deve gerar a quantidade K de agrupamentos espaciais contíguos que sejam o mais parecidos possível.

Por definição, em teoria de grafos, uma árvore com n vértices possui exatamente n-1 arestas de forma que, ao remover uma aresta, obrigatoriamente restarão duas componentes conexas. Analogamente, a cada nova aresta removida, uma nova componente conexa é gerada. Deste modo, para se obter K componentes conexas, K-1 arestas devem ser removidas.

O problema agora se resume em decidir quais arestas devem ser cortadas para que se obtenha os melhores agrupamentos possíveis. Existem diversas formas de se realizar esta etapa como, por exemplo, eliminar as K-1 arestas com maior valor, ou seja, menor similaridade

O metódo implementado leva em consideração a qualidade dos dois agrupamentos gerados ao remover cada aresta. Este método segue o seguinte algoritmo:

1. *Seja* G *o grafo que representa a árvore geradora mínima e* K *a quantidade de grupos que se deseja obter no final;*
2. *Enquanto a quantidade de componentes conexas de* G *for menor do que* K*, faça:*
    1. *Para cada aresta* A *existente em* G*, faça:*
        1. *Seja* $Q_t(A)$ *o valor da qualidade obtida ao remover* A *de* G*;*
    2. *Remova do grafo* G *a aresta* A *com maior valor de* $Q_t(A)$;
3. *Faça com que cada componente conexa do grafo* G *seja um grupo;*

A função $Q_t(A)$ é que vai definir quais arestas serão removidas e como serão formados os grupos. Portanto, quanto melhor for esta avaliação, melhor a qualidade do resultado. Nesta implementação, a função utilizada foi a seguinte:

$$Q_t(A) = SD_t(A) - \big(SD_a(A) + SD_b(A)\big),$$

onde:

$SD_t(A)$ é a soma das diferenças entre todos os elementos da componente conexa C_t onde se encontra a aresta A, que é dada por: $SD_t(A) = \sum_{i,j \subseteq C_t} d_{ij}$;

$SD_a(A)$ e $SD_b(A)$ são as somas das diferenças entre todos os elementos das duas componentes conexas C_a e C_b geradas pela remoção da aresta A do grafo G, como mostrado pela Figura 3.6. Ambas são calculadas da mesma forma que $SD_t(A)$, porém iterando nos seus respectivos domínios.

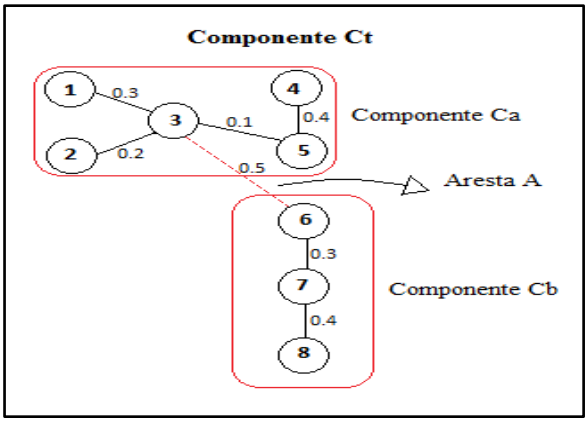


Figura 3.6 – Geração de duas componentes conexas pela remoçao de uma aresta

3.4 - Implementação

A implementação deste método foi feita na linguagem de programação JAVA, em forma de uma API, que é um conjunto de rotinas ajustadas para executar determinadas tarefas de forma que este possa ser reutilizado por outros softwares, sem que seja necessário ter conhecimento sobre sua implementação. O objetivo disso foi fazer com que o método implementado pudesse ser utilizado facilmente como ferramenta para análise de agrupamentos em qualquer problema.

Para utilizar essa API, os dados de entrada tanto espaciais como não-espaciais devem estar armazenados em um banco de dados do tipo PostgreSQL com extensão PostGIS, e um arquivo no formato XML deve ser gerado contendo as informações necessárias para que seja possível acessá-lo. Dessa forma, é possível executar as rotinas que serão responsáveis por recuperar os dados, executar o algoritmo de agrupamento e salvar a resposta no banco de dados.

Para facilitar o entendimento e a geração dos códigos, a organização foi feita dividindo-os em quatro diretórios principais: cd (classes de domínio), cs (classes de serviços), cc (classes de controle) e compartilhado. É possível ver a forma como essas classes estão ligadas entre si a partir do diagrama de classes apresentado na Figura 3.7.

Na seção seguinte, serão descritas todas as classes implementadas na API.

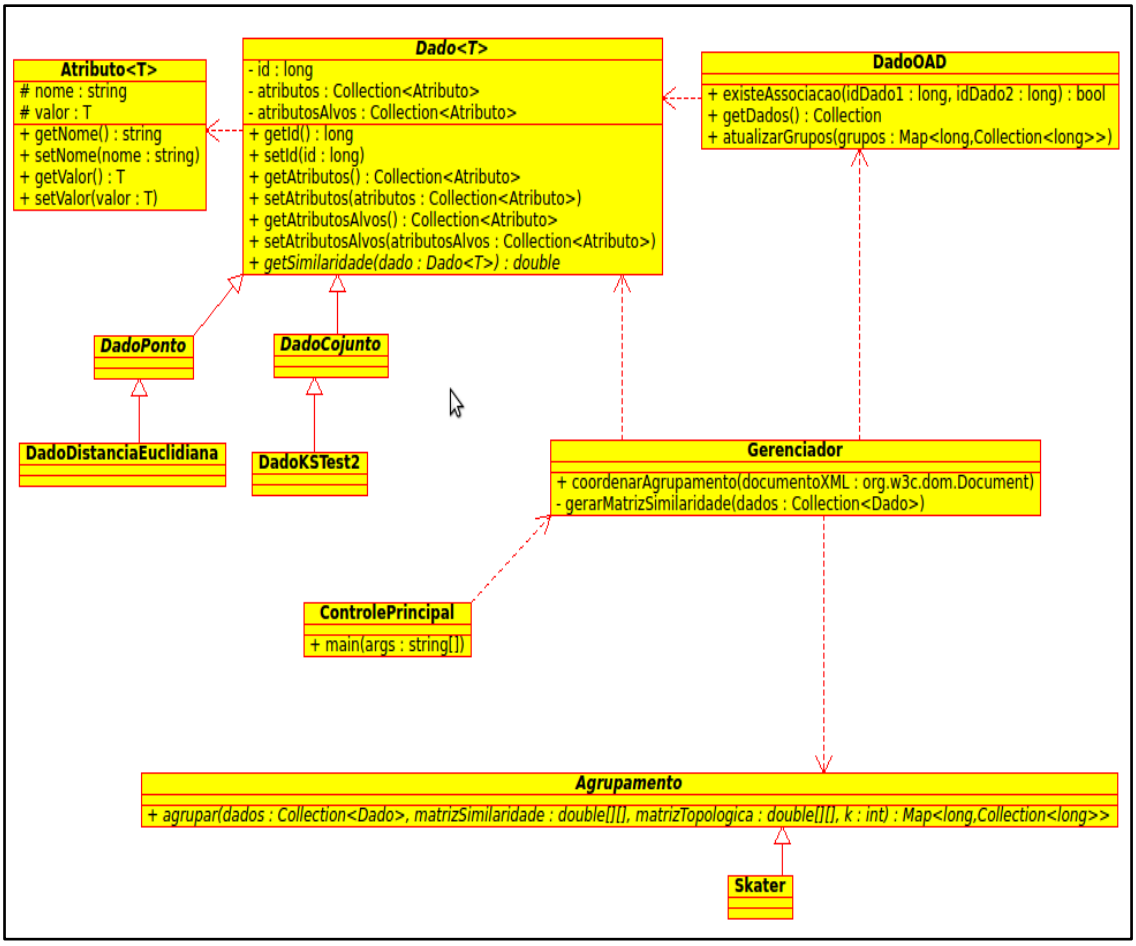


Figura 3.7 – Diagrama de Classes da API

3.4.1 – Classes de Domínio

Neste diretório, são armazenadas as entidades, que são as classes que representam todos os dados manipulados e a classe de rotinas responsáveis por fazer a comunicação entre o banco de dados e a aplicação. As entidades são:

Atributo:

Esta é uma classe concreta que representa cada uma das informações não-espaciais dos elementos que serão agrupados. Ela possui duas variáveis que são: o nome e o valor do atributo.

Dado:

Dado é uma classe abstrata que representa todos os elementos que serão agrupados pela aplicação. Esta classe possui o identificador, que é um número inteiro único para cada elemento; um conjunto de atributos que guarda todas as informações não-espaciais do elemento e um conjunto de atributos-alvo, que são os atributos que serão utilizados pelo algoritmo de agrupamento. Além disso, ela possui também o método abstrato, responsável por calcular o valor da similaridade entre o próprio elemento e algum outro, que deve ser implementado pelas classes concretas que a estenderem.

DadoPonto:

DadoPonto é uma classe abstrata que estende Dado, cujo objetivo é garantir que todas as classes que a estenderem possuam todos os atributos com apenas um valor. Este tipo de Dado deve ser utilizado para executar métodos que levem em conta apenas o valor médio dos atributos não-espaciais.

DadoConjunto:

Esta também é uma classe abstrata que estende Dado, porém as classes que a estenderem podem ter atributos com mais de um valor, ou seja, um conjunto de valores. Este é o tipo de Dado que foi utilizado para executar o método proposto, já que este leva em conta todos os valores dos atributos não-espaciais.

DadoDistanciaEuclidiana:

Esta é uma classe concreta que estende DadoPonto diretamente e Dado indiretamente, logo, ela possui as características das duas. Assim, nesta classe deve ser implementado o método que calcula a similaridade. O método foi implementado calculando a distância euclidiana entre os atributos não-espaciais dos elementos.

DadoKSTest2:

Esta é uma classe concreta que estende DadoConjunto diretamente e Dado indiretamente. Portanto, assim como no caso anterior, esta classe deve implementar o método que calcula a similaridade. O método foi implementado calculando o teste de Kolmogorov-Smirnov entre as amostras dos atributos não-espaciais dos elementos.

DadoOAD:

Esta é uma classe de serviço conhecida como Objeto de Acesso a Dados (OAD), que é um tipo de classe responsável por fazer a comunicação com o banco de dados. Como foi dito anteriormente, para executar a aplicação, é necessário um arquivo XML contendo as informações necessárias para realizar uma conexão ao banco de dados onde os dados estão armazenados. Com essas informações é possível abrir uma conexão com o banco a partir do JDBC, que é uma API em Java especializada em realizar consultas em diversos tipos de banco de dados, e realizar a comunicação. Esta classe possui três métodos implementados. O primeiro método recupera todos os elementos do banco de dados com seus dados não-espaciais e os retorna em uma coleção de Dado's. O segundo verifica se dois elementos estão ligados ou não fisicamente a partir dos dados espaciais, que também se encontram no banco. Por último, o método que salva no banco de dados o número do grupo a que cada elemento pertence.

3.4.2 – Classes de Serviços

No diretório das classes de serviços, estão as classes que possuem as rotinas que são responsáveis pela execução algoritmo. Neste diretório existem três classes, são elas:

Gerenciador:

Esta é a principal classe de serviço, que é responsável por fazer a comunicação com os outros módulos e coordenar a aplicação. A primeira rotina é a que coordena todo o método, desde buscar os dados do arquivo XML até salvar o resultado do algoritmo no banco. Dentro dela é que são chamados os métodos da classe DadoOAD, o de

criação da matriz de similaridade e o responsável pela execução do algoritmo de agrupamento. A outra rotina, que é chamada pela primeira, gera a matriz de similaridade a partir dos dados recuperados do banco de dados.

Agrupamento:

Esta é uma classe abstrata responsável por definir o padrão para as classes que implementam o algoritmo de agrupamento. Em outras palavras, qualquer outra classe que estenda esta e implemente um algoritmo de agrupamento baseado em matriz de similaridade poderá ser utilizada pelo método.

Skater:

Esta é a classe de serviço que implementa o algoritmo de agrupamento SKATER e é a única que estende a classe Agrupamento. A classe Skater que é chamada pela rotina principal da classe Gerenciador, executa o algoritmo e retorna o resultado para que ela termine o gerenciamento, salvando o resultado no banco de dados.

3.4.3 – Classes de Controle

Esta é o diretório que contém uma única classe, que é a primeira a ser executada pelo aplicativo. O objetivo desta classe é fazer a inicialização e os preparativos para que o método possa ser executado.

ControlePrincipal:

Esta classe é responsável por fazer a leitura do parâmetro de entrada, que é o diretório do arquivo XML com as configurações de acesso ao banco de dados. Ela busca este arquivo e o carrega no aplicativo de um forma que ele possa ser lido mais rapidamente, utilizando a API em Java DOM, que é especializada em acessar documentos eletrônicos. Em seguida, ela chama a rotina principal, que coordena todo o restante do método.

3.4.4 – Classes Compartilhadas

As classes localizadas neste diretório são as utilitárias, que podem ser acessadas por qualquer outro diretório (ou módulo). Elas oferecem auxílio básico, mas que pode ser utilizado muitas vezes e em classes diferentes.

Constantes:

A primeira classe, como o próprio nome já diz, nada mais é do que uma lista das constantes que são utilizadas em todo o aplicativo. Neste arquivo, estão escritos os nomes padrões que devem ser utilizados no arquivo XML, para que possa ser possível a recuperação das informações contidas nele.

Util:

Esta é uma classe de serviço que contém métodos básicos que podem ser chamados em diversos lugares do código. Como exemplo, há o método de calcular a média e outro para o cálculo do desvio-padrão, além das rotinas de transformar uma variável de um determinado tipo em outro.

3.5 – Considerações finais

Neste capítulo, foi mostrado que ao calcular alguma estatística dos dados ao invés de utilizar todas as observações disponíveis pode causa uma queda na qualidade da análise. Com isso, houve uma motivação em buscar algum meio de comparar conjuntos ou amostras de dados, com a finalidade de utilizar este método de comparação em um algoritmo de agrupamento. Em seguida, o teste de hipóteses de Kolmogorov-Smirnov para duas amostras foi apresentado como uma opção para comparar amostras de dados, baseado em suas distribuições empíricas. Por fim, um algoritmo de agrupamento baseado no p-valor de um teste de Kolmogorov-Smirnov foi implementado para que seja possível avaliar o seu desempenho e dizer se é uma boa alternativa para os métodos que são utilizados atualmente.

Capítulo 4 – Avaliação do método proposto

4.1 – Introdução

Neste capítulo, foi feita uma avaliação dos resultados dos experimentos com base no método proposto nesta dissertação. Estes experimentos tiveram por objetivo ilustrar a aplicação do método proposto em um banco de dados controlado.

Os experimentos consistem na aplicação da abordagem proposta para agrupar regiões espaciais a partir de seus atributos não-espaciais e suas relações topológicas. Além disso, uma abordagem diferente, que representa os conjuntos de observações pela estatística da média, também foi implementada e executada utilizando os mesmos dados com objetivo de comparar os resultados Assim, conduziu-se uma análise com objetivo de verificar e investigar os resultados gerados para comprovar a eficácia do método, bem como o cenário de aplicação.

4.2 – Ambiente de experimentação

A realização dos experimentos é parte fundamental da avaliação da proposta desta dissertação. Para isso, um protótipo foi desenvolvido na linguagem de programação JAVA, onde o método proposto e o agrupamento baseado na estatística foram implementados. O banco de dados PostgreSQL com sua extensão espacial PostGIS, o servidor de mapas MapServer, o MATLAB e a ferramenta de planilhas Microsoft Excel também serviram de apoio à condução dos experimentos.

O protótipo em JAVA está integrado ao banco de dados PostgreSQL e sua extensão PostGIS. Assim, as informações espaciais e não-espaciais podem ser armazenadas e recuperadas facilmente, para ser processadas pelo programa. O resultado obtido é, então armazenado no mesmo banco de dados, de forma a ser exportado para plataformas onde seja possível a análise.

Uma aplicação foi desenvolvida, utilizando o framework MapServer (MAPSERVER, 2014), para apresentar o mapa dos grupos. Desta forma, esperou-se verificar visualmente a distribuição espacial dos grupos de regiões gerados pelos

métodos aplicados. Na Figura 4.1, o mapa gerado representa um esquema de agrupamento, onde cada cor corresponde a um grupo.

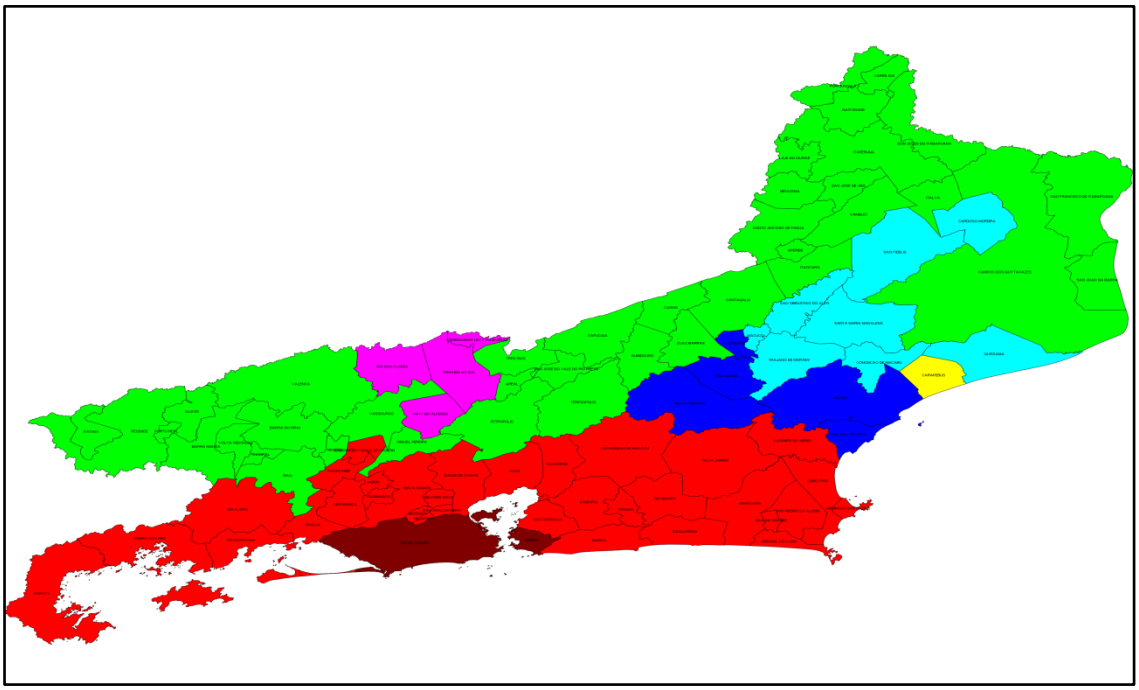


Figura 4.1 – Visualização do resultado pela aplicação no MapServer

Com o MATLAB e o Excel, foi possível fazer análises estatísticas dos resultados gerando gráficos, calculando medidas estatísticas e manipulando variáveis.

4.3 – Estudo de caso

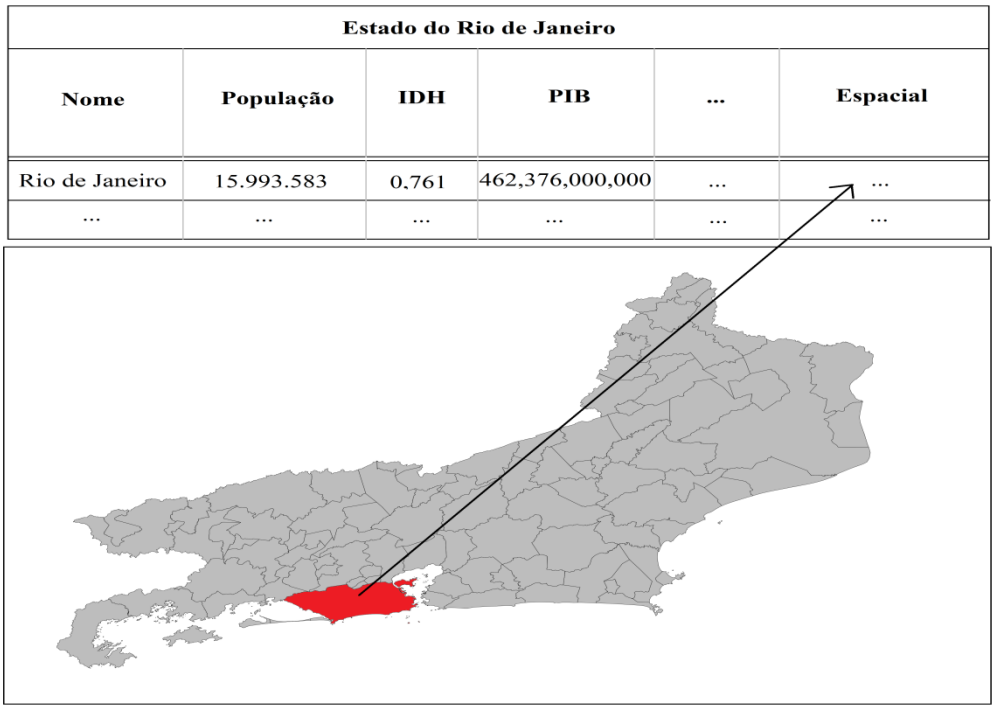
Nesta seção, serão descritos o estudo e a metodologia adotada para realização dos experimentos. Primeiramente, o problema será apresentado no contexto do agrupamento espacial e, em seguida, os passos para executar o experimento de forma correta.

4.3.1 – Problema

O Exame Nacional do Ensino Médio (ENEM) vem se tornado cada vez mais importante no Brasil, o que o torna uma rica fonte de informações sobre indicadores de educação e de sua distribuição entre cidades, estados ou até do país. Desta forma, a

descoberta de grupos de regiões cujos estudantes tiveram desempenho semelhante, pode auxiliar a tomada de decisões, tanto no contexto da administração pública, como mesmo para estudos de desenvolvimento educacional.

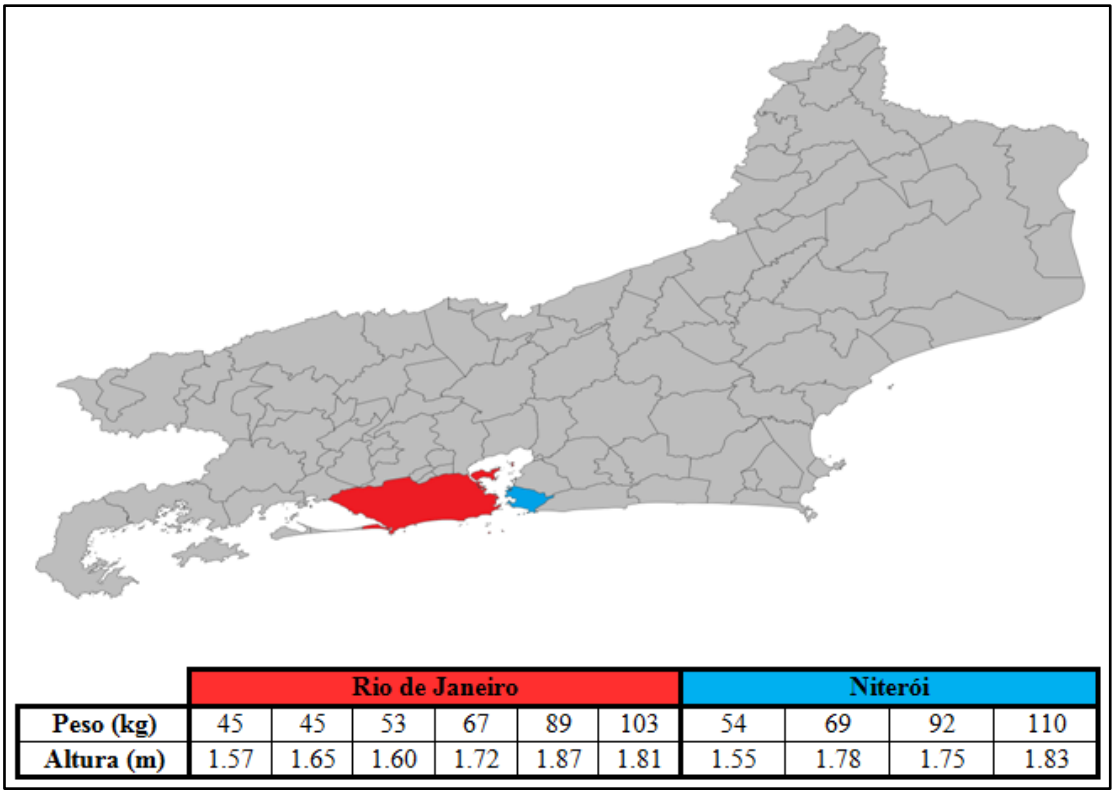
As técnicas clássicas de agrupamento espacial existentes na literatura tratam objetos (regiões) espaciais como instâncias mapeadas no espaço de entrada de seus atributos. Por exemplo, o IDH, a população e o PIB de cada região.



Estado do Rio de Janeiro					
Nome	**População**	**IDH**	**PIB**	**...**	**Espacial**
Rio de Janeiro	15.993.583	0,761	462,376,000,000	...	...
...	...	...	...	...	...

Figura 4.2 – Tabela espacial com dados das cidades do estado do Rio de Janeiro

Contudo, há casos em que para um objeto espacial (região) várias instâncias estão associadas. Por exemplo, peso e altura dos habitantes de um município.



	Rio de Janeiro						Niterói			
Peso (kg)	45	45	53	67	89	103	54	69	92	110
Altura (m)	1.57	1.65	1.60	1.72	1.87	1.81	1.55	1.78	1.75	1.83

Figura 4.3 – Dados com várias instâncias dos municípios do Rio de Janeiro e Niterói

Para contornar este tipo de problema, pode-se calcular, por exemplo, as médias dos atributos para cada região e posteriormente aplica-se um algoritmo de agrupamento. Desta forma, regiões que tenham vetores de médias parecidas tendem a ser alocadas no mesmo grupo.

No entanto, essa técnica pode levar regiões com médias semelhantes, mas com distribuições distintas, pertencerem ao mesmo grupo.

4.3.2 – Metodologia

Nesta seção será apresentada a metodologia aplicada para realizar os experimentos necessários para validar o método proposto. Os passos da metodologia são ilustrados no fluxograma abaixo.

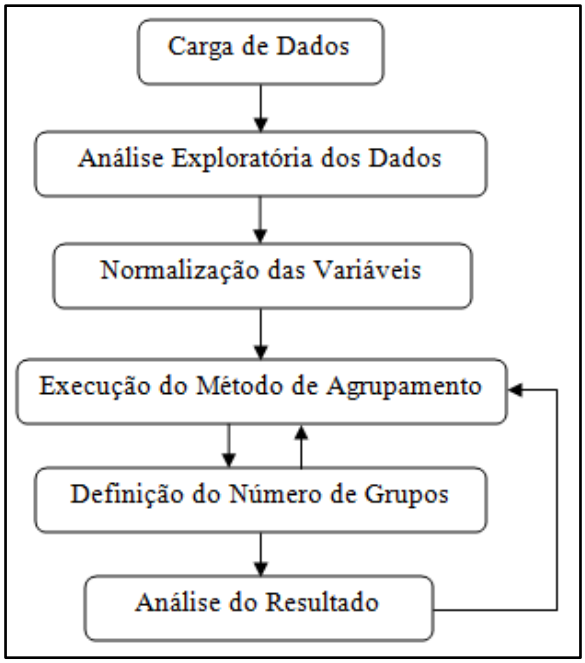


Figura 4.4 – Fluxograma da metodologia

Nas próximas seções, cada um dos itens listados acima será descrito de forma a explicar como foi realizado o experimento.

4.3.2.1 – Carga dos dados

Os dados utilizados são relativos ao Enem do ano de 2011, obtidos atráves do portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2014). Além disso, a base espacial referente à geometria dos municípios do estado do Rio de Janeiro foi adquirida através do portal do Instituto Brasileiro de Geografia e Estatística (IBGE, 2014). Como os dados foram obtidos por fontes externas, foi necessário armazená-los no ambiente de experimentação para que fosse possível utiliza-los quando necessário nos experimentos. O banco de dados PostgreSQL com sua extensão para dados espaciais, PostGIS, foi utilizado para esse propósito

Os dados espaciais e não-espaciais extraídos foram armazenados em uma tabela com os atributos que podem ser visto na Figura 4.5.

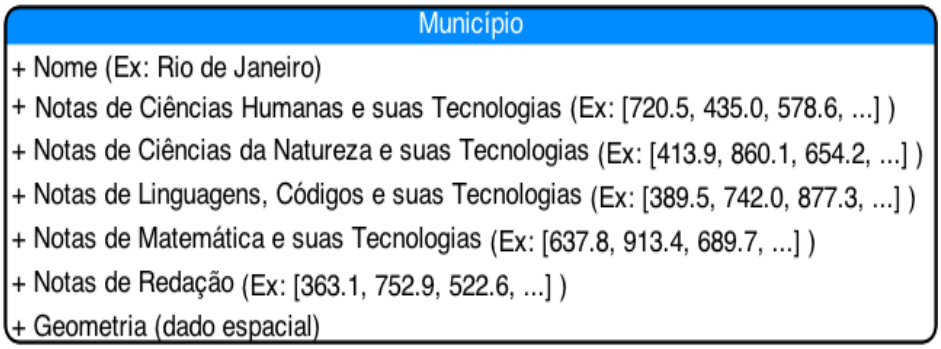


Figura 4.5 – Diagrama da entidade que representa a tabela dos munícipios

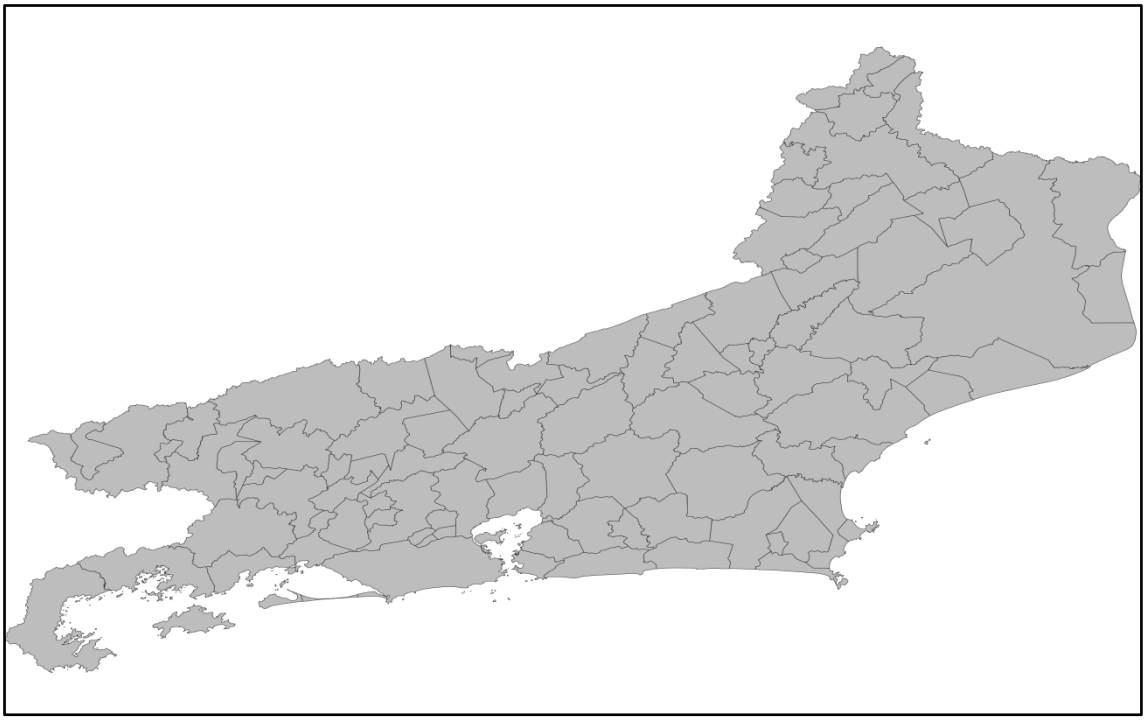


Figura 4.6 – Mapa com os munícipios do estado do Rio de Janeiro gerado pela base do IBGE

O mapa da Figura 4.6, que foi gerado a partir dos dados espaciais dos municípios, apresenta a divisão dos 92 municípios do estado do Rio de Janeiro. O banco de dados associou cada um desses municípios às notas de todos os participantes que fizeram a prova do Enem no estado. Desta forma, a tabela gerada possui 92 linhas indicando cada município e as suas respectivas notas.

4.3.2.2 – Análise exploratória dos dados

Nesta seção apresentamos a análise exploratória dos dados, com objetivo de

extrair o máximo de informações a partir de técnicas estatísticas, tanto gráficas como quantitativas, de forma a construir uma visão crítica sobre a natureza dos dados utilizados nos experimentos.
O MATLAB e a planilha eletrônica Excel foram utilizados nesta avaliação, visando o suporte da análise realizada.

Nesta etapa, utilizou-se as seguintes técnicas:

- Gráficos de dispersão:
  - histograma;
  - diagrama de caixa (boxplot);
- Medidas resumo:
  - medidas de tendências e variabilidade;

Como o experimento a ser desenvolvido baseia-se nas notas de matemática do Enem-2011 no estado do Rio de Janeiro, onde cada município possui um subconjunto de notas associado aos alunos em suas respectivas localidade. Este subconjunto será considerado uma amostra da variável aleatória das notas de matemática, cuja distribuição esta associada a localidade, no caso município fluminense.

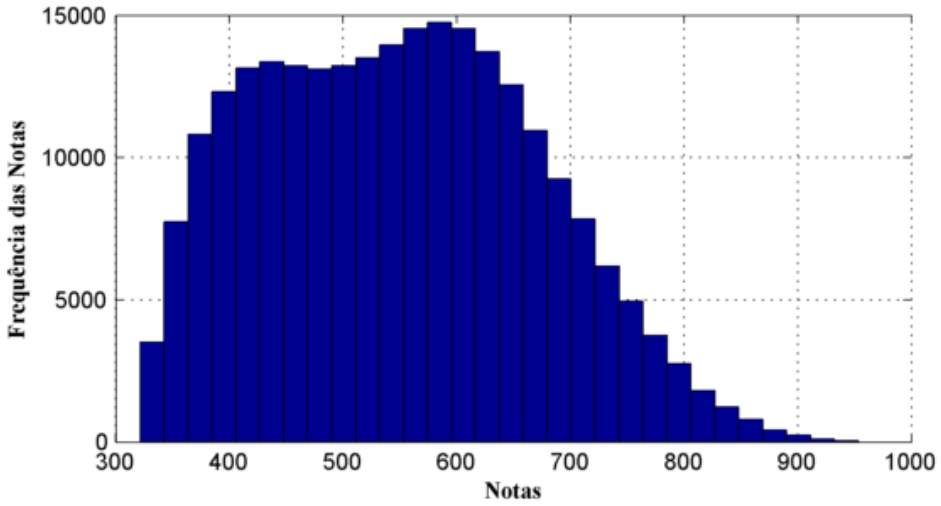


Figura 4.7 – Histograma das notas de todos os municípios

A Figura 4.7 mostra o histograma das notas de todos os municípios do Rio de Janeiro. De acordo com este, é possível verificar a moda das notas próxima de 600 e

encontra-se a “segunda moda”, um pouco menor, próxima de 450. Além disso, a maior parte das notas concentram-se entre 400 e 650. Pode-se destacar que, a partir de 650, o histograma apresenta queda acentuada.

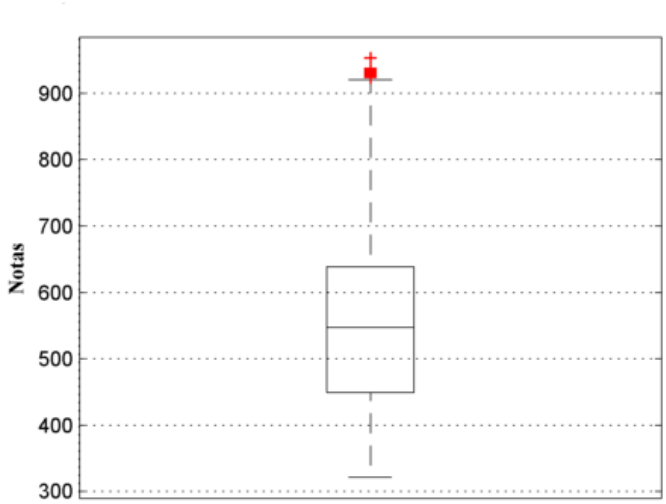


Figura 4.8 – Boxplot das notas de todos os municípios

A Figura 4.8 ilustra o boxplot das notas dos municípios do estado do Rio de Janeiro. A partir deste, nota-se que a mediana encontra-se próxima de 550, ou seja, metade dos estudantes obtiveram nota abaixo deste valor. O boxplot portanto confirma a grande concentração de pessoas que alcançaram notas entre 450 e 650. É possível notas ainda algumas notas consideradas outliers no gráfico, devido à distância entre estas e o intervalo interquartis.

Tabela 4.1 – Medidas estatísticas de todas as notas do estado do Rio de Janeiro

Medidas estatísticas	
Mínimo	321.6000
1º Quartil	449.8000
Mediana	547.7000
Média	550.2949
3º Quartil	638.8000
Máximo	953.0000
Desvio Padrão	121.9680
Quantidade	248409

A Tabela 4.1 confronta o que foi analisado pelos gráficos. É possível observar

que as notas estão distribuídas entre 321.6 e 953.0, com média em 550.3, pouco destoada da mediana. No total, 248409 alunos realizaram a prova dentro do estado.

Analisando as notas de cada município indivídualmente, destacam-se as médias de Niterói, com aproximadamente 601, por ser a maior e a de Carapebus, com pouco mais do que 470, por ser a menor. Além disso, Carapebus também se destaca por ter a menor quantidade de candidatos fazendo a prova, com apenas 72 pessoas. Em contrapartida, o município do Rio de Janeiro foi o que obeteve a maior quantidade de participantes, com 104808 candidatos. Este valor é mais de 90000 a mais que São Gonçalo, que é o segundo com mais candidatos. Estes três municípios (Carapebus, Niterói e Rio de Janeiro) serão analisados mais profundamente a seguir, objetivando ilustrar as características da distribuição de notas de cada um.

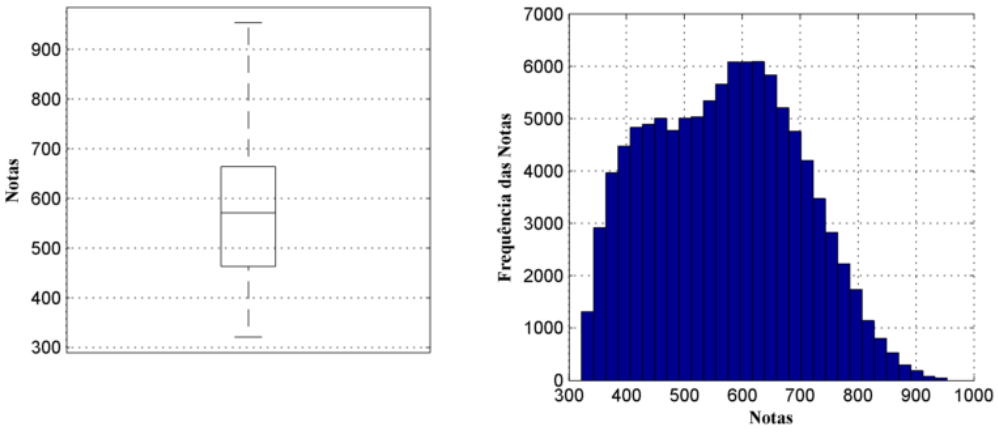


Figura 4.9 – Boxplot e Histograma das notas do município do Rio de Janeiro

Na Figura 4.9, o boxplot e o histograma das notas do município do Rio de Janeiro estão apresentados. O primeiro ponto que chama atenção é a semelhança entre esses gráficos e os gerados utilizando todas as notas de todos os municípios. Isto já era de se esperar, visto que as notas deste município correspondem a mais de 40% do total. Comparando os dois boxplots, se pode notar que a mediana deste está entre 550 e 600, diferenciando do outro que está abaixo de 550. Além disso, verifica-se a moda um pouco deslocada, para um valor entre 600 e 650, enquanto no gráfico para todo estado esta se encontra bem próxima de 600.

A Tabela 4.2 descreve a distribuição por quartis das notas do município do Rio de Janeiro. Pelo quadro é possível ver que a média e a mediana estão um pouco maiores quando comparadas com as notas de todos os municípios. Adicionalmente, quase

metade dos estudantes obtiveram nota entre 463 e 663, confirmando a semelhança com as estatísticas de quartis de todos os municípios.

Tabela 4.2 – Medidas estatísticas das notas do município do Rio de Janeiro

Medidas estatísticas	
Mínimo	321.6000
1º Quartil	463.7000
Mediana	571.4000
Média	569.1287
3º Quartil	663.8500
Máximo	953.0000
Desvio Padrão	127.7577
Quantidade	104808

Na Figura 4.10, os dados são relativos ao município de Carapebus. Nota-se que, pelo fato da amostra de notas não ser muito grande, o histograma não é muito preciso. Por conta disso, fica difícil tentar prever qual seria a distribuição original dela. Entretanto, é possível tirar algumas conclusões básicas, que auxiliam na análise dos dados, como média, desvio padrão etc. Os gráficos mostram que as maiores notas estão entre 700 e 750. Além disso, é possível ver que a grande maioria das pessoas tiraram nota abaixo de 500. O boxplot mostra a diferença entre essas notas e a distribuição do estado, por exemplo.

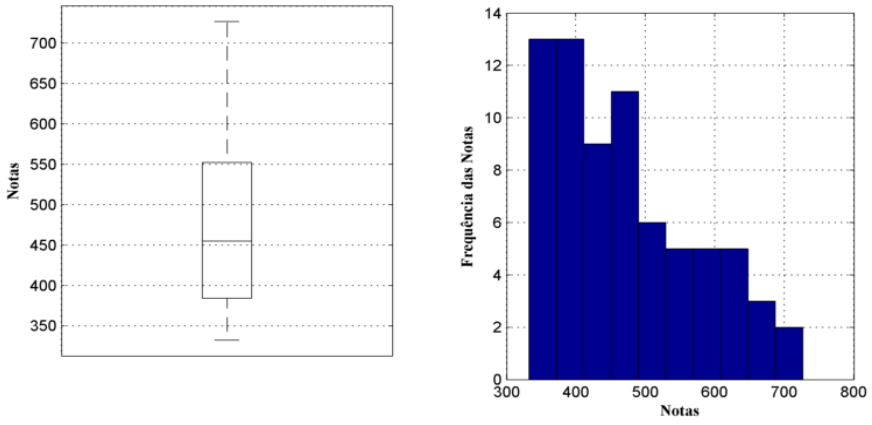


Figura 4.10 – Boxplot e Histograma das notas do município de Carapebus

A Tabela 4.3 confirma o que foi concluido pela Figura 4.10, sendo possível

perceber que metade dos candidatos obtiveram notas entre 332.5 e 445.05 e que 75% tiraram abaixo de 552.15. Nenhum estudante tirou nota maior do que 726.1. Além disso, a média foi de 470.1, a menor entre todos os municípios.

Tabela 4.3 – Medidas estatísticas das notas do município de Carapebus

Medidas estatísticas	
Mínimo	332.5000
1º Quartil	384.1500
Mediana	455.0500
Média	470.0986
3º Quartil	552.1500
Máximo	726.1000
Desvio Padrão	102.2904
Quantidade	72

Foi possível concluir que o município de Carapebus possui dados bem diferenciados dos demais, pois sua distribuição e a sua média são bem diferentes das dos outros municípios. Portanto, é razoável especular que este se encontre em um grupo diferente do grupo do Rio de Janeiro e Niterói, por exemplo.

A Figura 4.11 mostra o boxplot e o histograma das notas do município de Niterói, que é o município com a maior média de notas. De acordo com esses gráficos, é possível notar que a amostra tende a uma distribuição normal com média próxima de 600. Também é possível constatar que existe uma grande concentração de candidatos entre as notas 500 e 700, com a moda nas proximidades de 650.

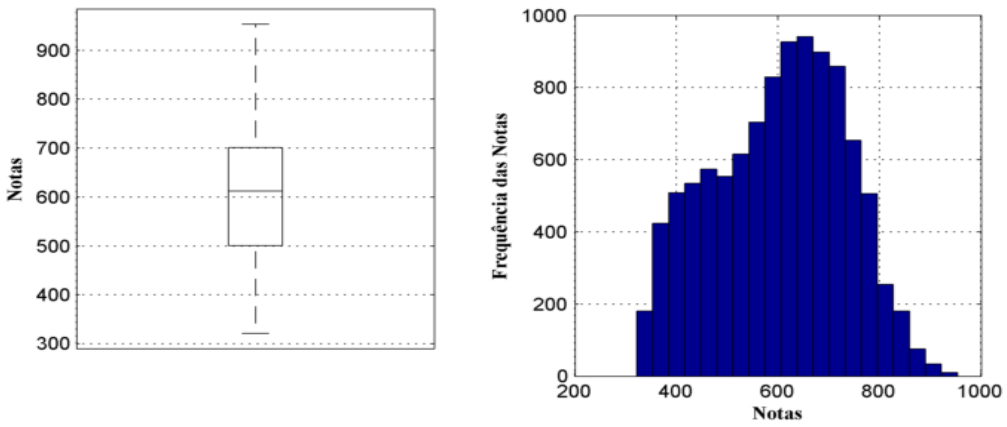


Figura 4.11 – Boxplot e Histograma das notas do município de Niterói

Como podemos observar na Tabela 4.4, o município de Niterói possui a maior média entre todos os outros, de aproximadamente 601.0. Adicionalmente, nota-se que apenas 25% dos candidatos obtiveram nota inferior a 500 e que mais da metade pontuou acima de 612. Essas informações sugerem que o município possui um ótimo índice de educação dentro do estado do Rio de Janeiro. Niterói possui média de notas de matemática bem acima das outras cidades, sendo o município com segunda maior média é Nova Friburgo, com 585.15. Portanto, ao aplicar o algoritmo de agrupamento espera-se que esta cidade acabe em um grupo isolado.

Tabela 4.4 – Medidas estatísticas das notas do município de Niterói

Medidas estatísticas	
Mínimo	321.6000
1º Quartil	500.7000
Mediana	612.5000
Média	601.0078
3º Quartil	700.8000
Máximo	953.0000
Desvio Padrão	130.5162
Quantidade	10261

Existem ainda situações que são interessantes de se analisar. Por exemplo, há casos de dois municípios que são vizinhos e possuem médias próximas, mas as distribuições de suas notas são relativamente diferentes.

As Figura 4.12 e Figura 4.13 apresentam a comparação entre as notas dos municípios de São Gonçalo e Maricá. Pelo histograma (Figura 4.12) é possível ver que as notas de São Gonçalo tem moda um pouco abaixo de 600 e logo após isso, ocorre uma queda bem acentuada. Já no município de Maricá, a moda se encontra um pouco antes, por volta de 500, e depois as notas caem de forma mais suave até próximo de 700, onde começam a cair acentuadamente como São Gonçalo.

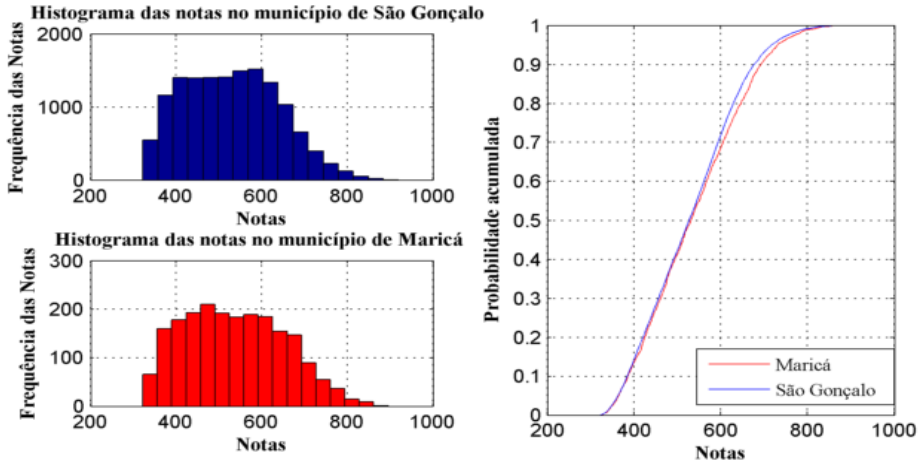


Figura 4.12 – Histograma e as fda's empíricas de São Gonçalo e Maricá

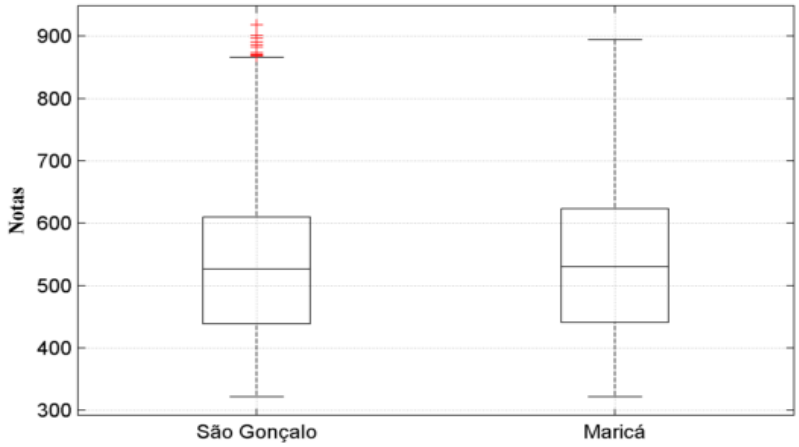


Figura 4.13 – Boxplot de São Gonçalo e Maricá

A Figura 4.12 apresenta também as funções de distribuições acumuladas (fda), construídas empiricamente a partir das notas destes dois municípios. Note que pouco depois do 400, pode-se observar que a fda de São Gonçalo está um pouco acima da outra. Isto é confirmado pelo histograma, onde é possível observar que as notas de São Gonçalo sobem um pouco mais acentuamente do que as de Maricá. Na sequência elas se igualam a medida que atingem a moda. Quando as notas estão perto de 600, próximas da moda das notas de São Gonçalo, as duas funções voltam a se distanciar. Após isto, as funções se reaproximam devido a queda das notas de Maricá ser mais lenta.

A Figura 4.13 apresenta o boxplot das notas destes dois municípios. A primeira coisa que se observa é a presença dos *outliers* no boxplot de São Gonçalo. Estes indicam que a maior nota possui valor acima de 900, o que não ocorre em Maricá, que não tem nenhuma nota acima de 900. Apesar disto, esse gráfico mostra que as notas de São Gonçalo são concentradas um pouco abaixo das notas de Maricá, o que faz com que

as notas muito afastadas sejam consideradas outliers.

A Tabela 4.5 confirma nossa constantação sobre a maior concentração das notas de São Gonçalo estarem um pouco abaixo das de Maricá. Além disso, a maior nota de São Gonçalo é 918.4 e a de Maricá é 894.3. A média dos dois municípios são relativamente próximas, o que justificaria agrupá-los em um mesmo grupo ao aplicar um algoritmo de agrupamento baseado nas médias das notas de cada um deles. Entretanto, por conta do baixo p-valor entre eles, mostrando que existe uma considerável diferença entre as distribuições das notas dos dois municípios, o método proposto por esta dissertação poderia alocá-los em grupos diferentes.

Tabela 4.5 – Medidas estatísticas das notas dos municípios de São Gonçalo e Maricá

Medidas estatísticas		
	São Gonçalo	**Maricá**
Mínimo	321.6000	321.6000
1º Quartil	438.7000	441.3000
Mediana	526.6000	530.4500
Média	529.4996	536.9130
3º Quartil	610.1000	623.6000
Máximo	918.4000	894.3000
Desvio Padrão	111.1331	115.9286
Quantidade	14208	2068
P-Valor	0.00150852	

Por fim, analisamos as Figura 4.14, Figura 4.15 e Figura 4.16, que ilustram os municípios Teresópolis e Rio das Ostras. Estes são separados por alguns municípios, porém possuem uma grande semelhança com relação a distribuição das notas de matemática. Assim, devido a restrição de contiguidade eles podem acabar em grupos diferentes. Note que os histogramas apresentam que os dois possuem praticamente a mesma moda e tendem a uma distribuição normal. O gráfico da função de probabilidade acumulada empírica e o boxplot confirmam que as duas distribuições são muito parecidas.

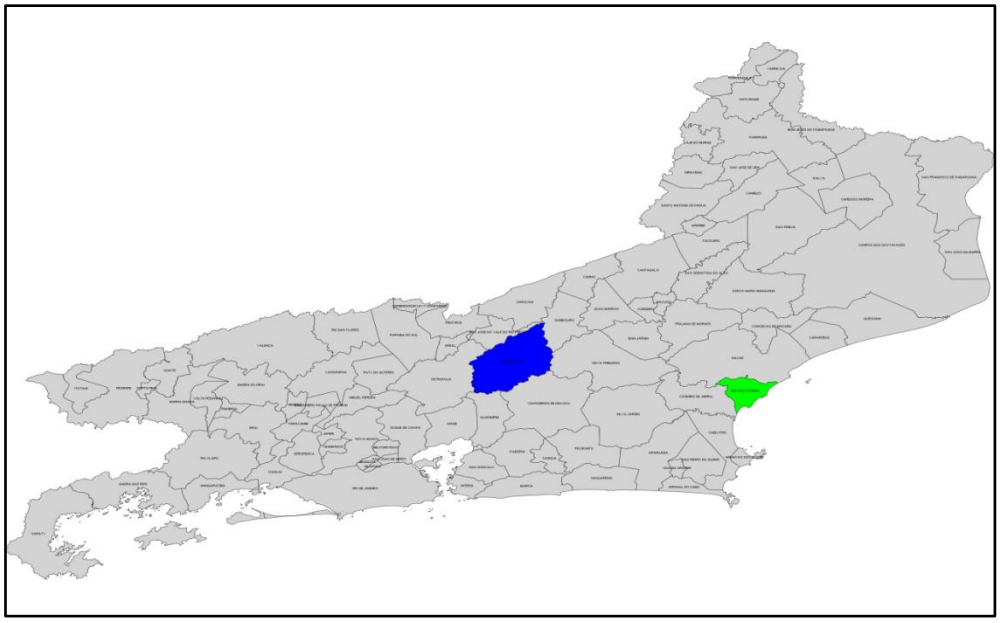


Figura 4.14 – Mapa mostrando a localização dos municípios de Teresópolis e Rio das Ostras

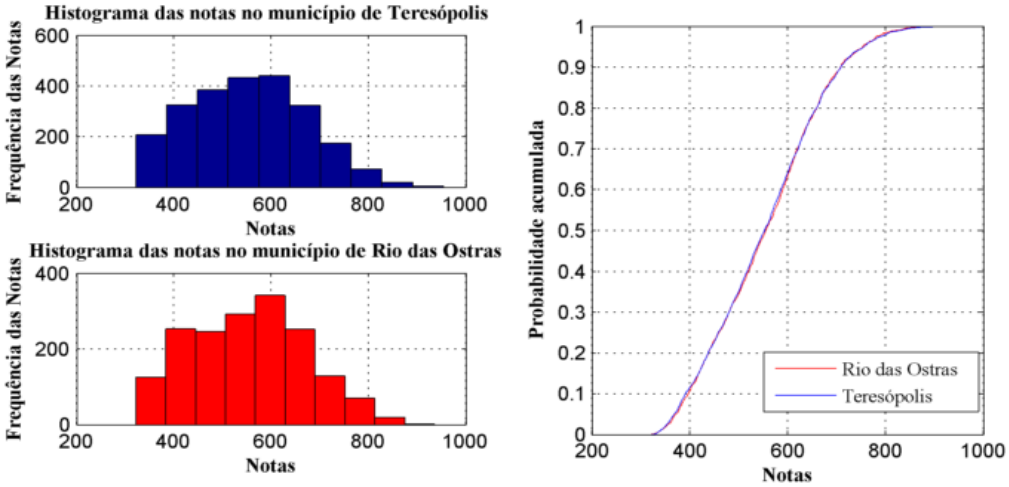


Figura 4.15– Histograma e fds's empíricas de Teresópolis e Rio das Ostras

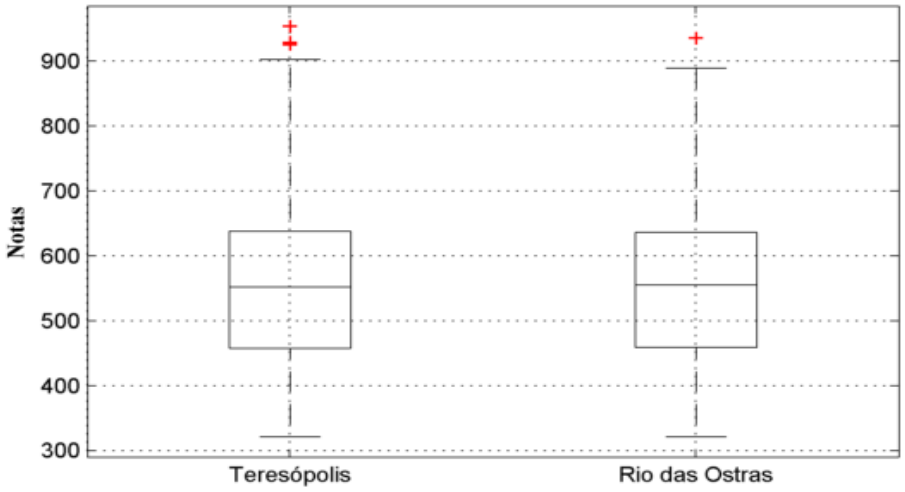


Figura 4.16 – Boxplot de Teresópolis e Rio das Ostras

Na Tabela 4.6 é possível notar a semelhança entre as medidas dos dois municípios, que confirmam nossa análise segundo os gráficos anteriores. O p-valor apresentado confirma a grande semelhança entre a distribuição das notas dos dois municípios, sendo ele bem perto de 1.

Tabela 4.6 – Medidas estatísticas das notas dos município de Teresópolis e Rio das Ostras

Medidas estatísticas		
	Teresópolis	**Rio das Ostras**
Mínimo	321.6000	321.6000
1º Quartil	457.7250	458.8500
Mediana	552.3000	555.1000
Média	552.9754	554.0535
3º Quartil	637.4000	636.2250
Máximo	953.0000	935.0000
Desvio Padrão	118.6806	117.2753
Quantidade	2389	1733
P-Valor	0.929169	

4.3.2.3 – Normalização das variáveis

Após análise, os dados serão normalizados de maneira que seus atributos sejam alocados em um intervalo entre zero e um. O algoritmo utilizado para normalizar foi o MIN-MAX, que faz com que o valor mais alto dentro da amostra passe a ser 1 e o mais baixo passe a ser 0. Esta normalização tem como base a seguinte fórmula (JAIN *et al.*, 2005):

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

onde,

- X é o conjunto de valores que se deseja normalizar;
- z é o valor de $x \in X$ após a normalização.

Embora a normalização pudesse ser aplicada também aos dados utilizados no método proposto, optou-se por manter os dados na escala original para evitar qualquer destoação na distribuição destes.

4.3.2.4 - Execução do método de agrupamento

Nesta seção, serão analisados os passos que o programa executa para processar os métodos de agrupamento. Os dados armazenados no banco de dados são recuperados pelo programa em Java e, a partir destes, a matriz com a topológia e a matriz de similaridade são calculadas. Em seguida, os algoritmos de agrupamentos são executados para valores de k variando entre 2 e 92, que é a quantidade total de municípios.

A matriz com a topológia foi gerada com base na contiguidade dos municípios, que foi calculada a partir dos dados espaciais. Porém, os municípios Rio de Janeiro e Niterói, que possuem o maior Índice de Desenvolvimento Humano (IDH) dentro do estado, não são contíguos. O acesso entre eles é feito pela Ponte Presidente Costa e Silva (Ponte Rio-Niterói), uma das vias de transporte mais utilizadas no estado. Dada a importância desses dois municípios, foi estabelecida uma ligação entre eles na matriz, com o intuito de enriquecer a análise.

4.3.2.5 - Definição do número de grupos

A definição da quantidade de grupos para o algoritmo de agrupamento é feita de forma subjetiva. No entanto, diversas técnicas podem auxilliar a escolha deste valor.

Segundo ALDENDERFER & BLASHFIELD (1984), ao analisar agrupamentos, o que se busca é identificar grupos homogêneos, ou seja, minimizar a soma das distâncias de elementos dentro do mesmo grupo (intragrupos) e maximizar a soma das distâncias dos elementos em grupos distintos (intergrupos). A análise das distâncias intergrupo pode ser prejudicada no contexto de dados espaciais, já que o algoritmo pode separar duas regiões com características semelhantes que não respeitem o critério espacial.

Como critério de avaliação para a escolha da quantidade de grupos, utilizou-se a soma das variações intragrupo. A partir deste critério, fez-se uma análise interna dos grupos, verificando a sua homogeneidade. Quanto menor a soma das distâncias internas entre os elementos dos grupos, melhor a qualidade do agrupamento. Para isso, o método calcula a soma das distâncias dentro dos grupos, que é dada pela fórmula a seguir:

$$D_{total} = \sum_{k}^{K} \sum_{i,j}^{N_k} (1 - Sim[i][j])$$

onde:

- D_{total} é a soma das distâncias intragrupo para um determinado agrupamento;
- K é a quantidade de grupos;
- N_k é a quantidade de elementos dentro do grupo k;
- $Sim[i][j]$ é a similaridade entre os elementos i e j do grupo k;

Note que, na equação, utiliza-se $1 - Sim[i][j]$ para representar a distância (dissimilaridade) entre dois elementos. Desta forma, calcula-se a distância entre todos os elementos dentro do mesmo grupo. Em seguida, soma-se as distâncias obtidas para cada grupo resultando na medida D_{total} .

Na Figura 4.17 apresentamos os valores D_{total} para grupos de 2 a 20, executando-se o método de agrupamento. Observou-se não ser interessante para o estudo final criar muitos grupos.

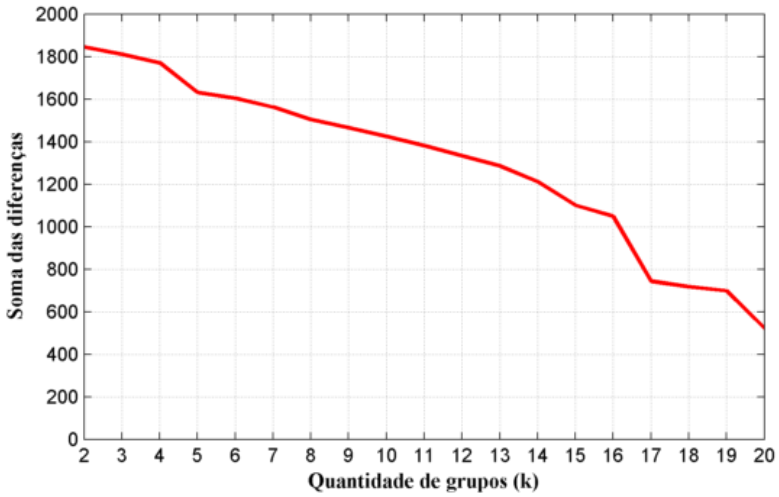


Figura 4.17 – Soma das diferenças intragrupo

Note que esta função decresce a medida que a quantidade de grupos aumenta. Grandes quedas de D_{total} significam bons ganhos ao se criar mais um grupo. No gráfico, é possível observar uma grande queda do 4 para o 5 e depois uma queda praticamente nula de 5 para 6. Portanto, um bom número para escolher seria o 5, que ofereceu um ganho relativo de homogeneidade. É possível notar outras quedas, destacam-se aquelas em 17 e 20. Neste caso, o número 5 seria o mais indicado, pois os outros dois são muito grandes e vão particionar demais a região, atrapalhando o estudo final.

4.3.2.6 - Análise do resultado

Depois de definida a quantidade de grupos, os agrupamentos resultantes são analisados. Nesta etapa, os resultados dos dois métodos serão estudados e comparados com o objetivo de avaliar o desempenho de nossa proposta.

As melhores escolhas de quantidade de grupos para o método proposto são 5, 17 e 20 grupos. Entretanto, como a quantidade de elementos a agrupar são 92 municípios, dividi-los em 17 ou 20 grupos resultaria em muitos grupos com poucos elementos, o que não é interessante para o estudo. Assim, a quantidade definida para as análises foi de 5 grupos.

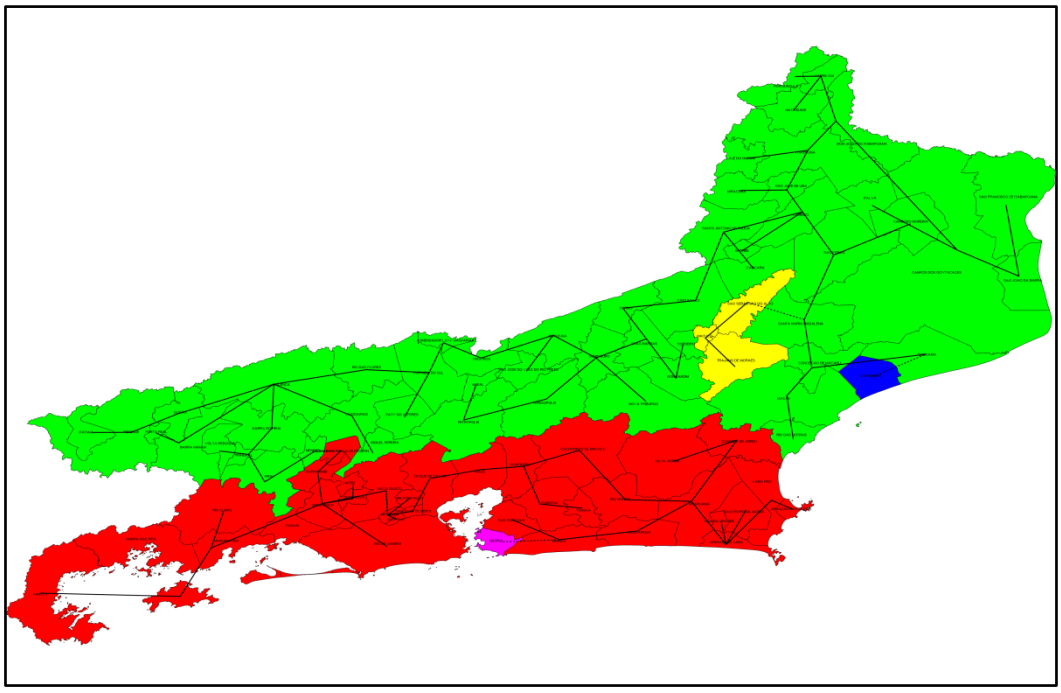


Figura 4.18 – Mapa do resultado do método proposto com 5 grupos

Na Figura 4.18 é apresentado o mapa do estado do Rio de Janeiro dividido em 5 grupos pelo método proposto. As linhas representam a árvore geradora mínima obtida a partir do método de agrupamento e as linhas tracejadas são as arestas que foram cortadas para formar os grupos. Cada grupo foi composto pelos seguintes municípios:

- **Grupo 1 (Vermelho) –** Paraty, Angra dos Reis, Rio Claro, Mangaratiba, Itaguaí, Engenheiro Paulo de Frontin, Paracambi, Seropédica, Rio de Janeiro, Japeri, Queimados, Nova Iguaçu, Belford Roxo, São João de Meriti, Mesquita, Nilópolis, Duque de Caxias, Magé, Guapmirim, São Gonçalo, Maricá, Cachoeiras de Macacu, Itaboraí, Tanguá, Rio Bonito, Saquarema, Silva Jardim,

Araruama, São Pedro da Aldeia, Iguaba Grande, Arraial do Cabo, Casimiro de Abreu, Cabo Frio e Armação dos Búzios;

- **Grupo 2 (Verde) –** Itatiaia, Resende, Porto Real, Quatis, Barra Mansa, Volta Redonda Valença, Barra do Piraí, Pinheral, Piraí, Mendes, Rio das Flores, Vassouras, Paty do Aferes, Miguel Pereira, Comendados Levy Gasparian, Paraíba do Sul, Petrópolis, Três Rios, Areal, Sapucaia, São José do Valo do rio Preto, Teresópolis, Carmo, Sumidouro, Duas Barras, Nova Friburgo, Cantagalo, Cordeiro, Bom Jardim, Rio das Ostras, Macaé, Conceição de Macabu, Santa Maria Madalena, Quissamã, Campo dos Goytacazes, São João da Barra, São Francisco de Itabapoana, Cardoso Moreira, Italva, São Fidelis, Itaocara, Aperibé, São Antônio de Pádua, Cambucí, Italva, Miracema, São José de Uba, Laje de Muriaé, Itaperuna, Bom Jesus do Itabapoana, Natividade, Porciúncula e Varre-sai;
- **Grupo 3 (Azul) –** Carapebus;
- **Grupo 4 (Amarelo) –** Macuco, Trajano de Moraes e São Sebastião do Alto;
- **Grupo 5 (Rosa) –** Niterói.

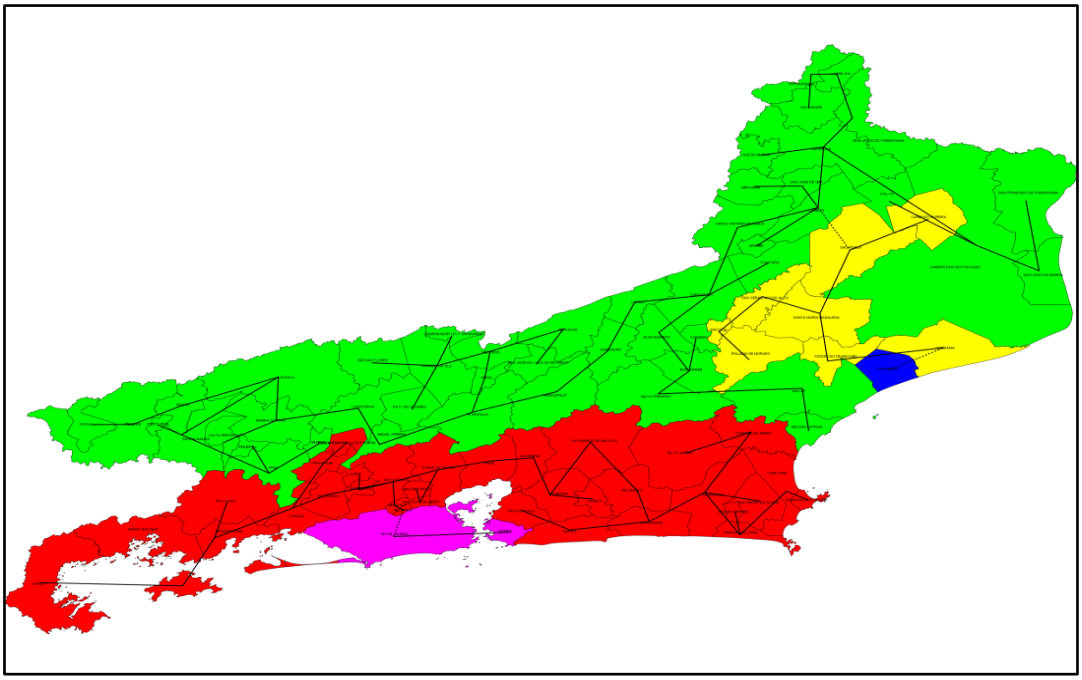


Figura 4.19 – Mapa do resultado do método que utiliza a média com 5 grupos

Na Figura 4.19, é possível observar o mapa do estado do Rio de Janeiro dividido pelo método de agrupamento aplicado nas médias das notas de cada município. Os

grupos são compostos pelos seguintes municípios:

- **Grupo 1 (Vermelho) –** Paraty, Angra dos Reis, Rio Claro, Mangaratiba, Itaguaí, Engenheiro Paulo de Frontin, Paracambi, Seropédica, Japeri, Queimados, Nova Iguaçu, Belford Roxo, São João de Meriti, Mesquita, Nilópolis, Duque de Caxias, Magé, Guapmirim, São Gonçalo, Maricá, Cachoeiras de Macacu, Itaboraí, Tanguá, Rio Bonito, Saquarema, Silva Jardim, Araruama, São Pedro da Aldeia, Iguaba Grande, Arraial do Cabo, Casimiro de Abreu, Cabo Frio e Armação dos Búzios;
- **Grupo 2 (Verde) –** Itatiaia, Resende, Porto Real, Quatis, Barra Mansa, Volta Redonda Valença, Barra do Piraí, Pinheral, Piraí, Mendes, Rio das Flores, Vassouras, Paty do Aferes, Miguel Pereira, Comendados Levy Gasparian, Paraíba do Sul, Petrópolis, Três Rios, Areal, Sapucaia, São José do Valo do rio Preto, Teresópolis, Carmo, Sumidouro, Duas Barras, Nova Friburgo, Cantagalo, Cordeiro, Bom Jardim, Rio das Ostras, Macaé, Campo dos Goytacazes, São João da Barra, São Francisco de Itabapoana, Italva, Itaocara, Aperibé, São Antônio de Pádua, Cambucí, Italva, Miracema, São José de Uba, Laje de Muriaé, Itaperuna, Bom Jesus do Itabapoana, Natividade, Porciúncula e Varre-sai;
- **Grupo 3 (Azul) –** Carapebus;
- **Grupo 4 (Amarelo) –** Macuco, Trajano de Moraes, São Sebastião do Alto, Conceição de Macabu, Santa Maria Madalena, Quissamã, São Fidelis e Cardoso Moreira;
- **Grupo 5 (Rosa) –** Niterói e Rio de Janeiro.

Na Figura 4.20, apresentamos as fda's estimadas a partir das notas de todos os municípios de cada grupo, gerado a partir da execução do método proposto. É possível notar que cada grupo possui uma distribuição bem diferente dos demais, com exceção dos grupos 1 e 2, que possuem distribuições mais próximas. Além disso, o gráfico confirma que os municípios de Niterói (Grupo 5) e Carapebus (Grupo 3) destoam bastante dos demais.

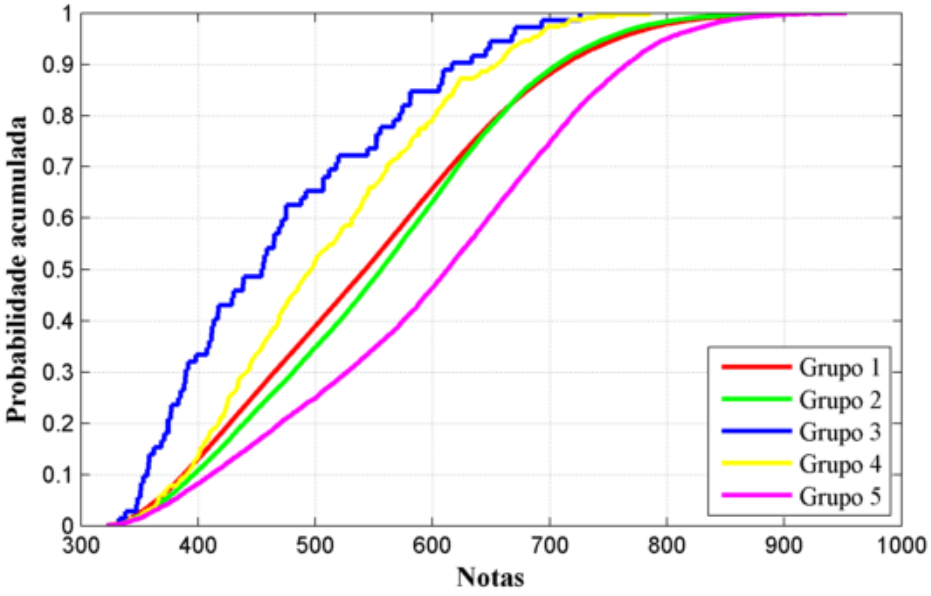


Figura 4.20 – Gráfico das fda's de todos os grupos para o método proposto

A Figura 4.21 apresenta as fda's de todos os grupos obtidos a partir do método que utiliza a média das notas. É possível observar, comparando com o gráfico da Figura 4.20, que as curvas se encontram mais próximas, mostrando que as distribuições dos grupos não estão tão bem definidas quanto as resultantes do método proposto.

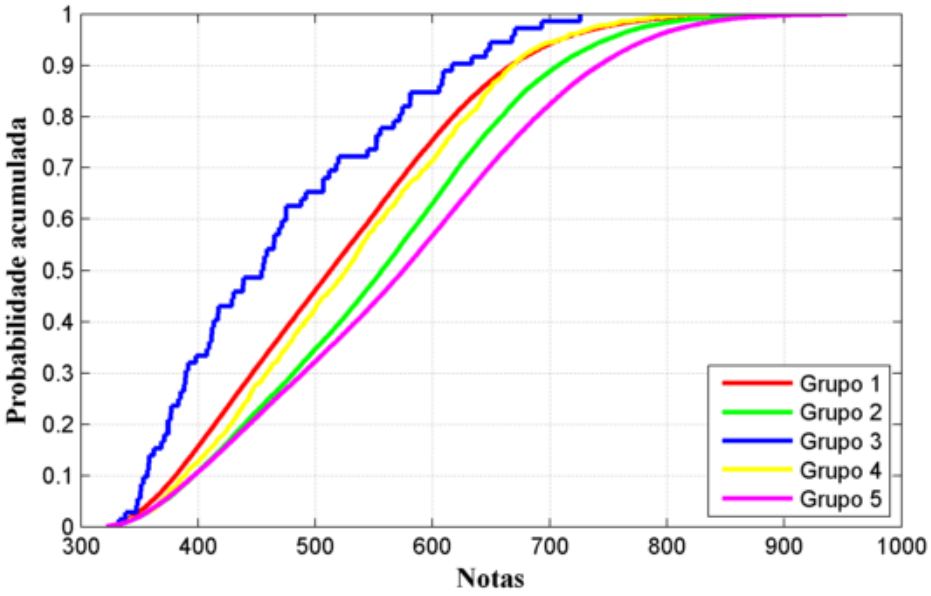


Figura 4.21 – Gráfico das fda's de todos os grupos para o método que utiliza a media

A seguir, será realizada uma análise individual de cada grupo focada no resultado obtido pelo método proposto e em seguida serão apontandas as principais

diferenças com a abordagem que utiliza as médias das notas.

Grupo 1:

O grupo 1 é composto por basicamente todo o sul fluminense, com exceção do município de Niterói que ficou em um grupo isolado. Na abordagem que utiliza a média das notas, a diferença é apenas a ausência do município do Rio de Janeiro. Este grupo possui 34 cidades, que são as que estão pintadas de vermelho no mapa da Figura 4.18. Esse grupo é o que mais possui candidatos que fizeram o exame, por conta principalmente do Rio de Janeiro, que é o município mais diferente dos demais, tanto pela média como pela distribuição, vide Figura 4.22.

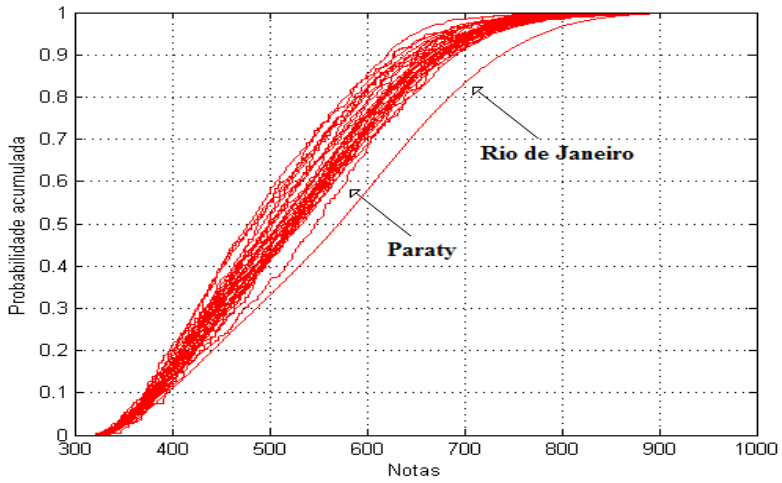


Figura 4.22 – Gráfico das fda's de todos os municípios do grupo 1

A Figura 4.22 mostra claramente a diferença entre as notas do Rio de Janeiro e as do restante do grupo. Paraty é o município cuja distribuição das notas mais se aproxima do Rio de Janeiro, mas ainda assim é possível ver a grande diferença entre as curvas apresentadas no gráfico, principalmente nas notas mais altas. Isto se deve à quantidade de grupos escolhida no passo anterior, pois quanto maior esta for, mais homogêneos os grupos se tornarão. Como o Rio de Janeiro é o município que mais se difere do restante de seu grupo, ele é um forte candidato a sair caso a quantidade de grupos aumentasse. Outro detalhe importante é que a quantidade de candidatos neste município é tão grande, que faz com que a distribuição de todo grupo (mostrada na Figura 4.20) seja muito afetada por ele. Desta forma, se separarmos o Rio de Janeiro do grupo 1, as fda's ficam como apresentado na Figura 4.23.

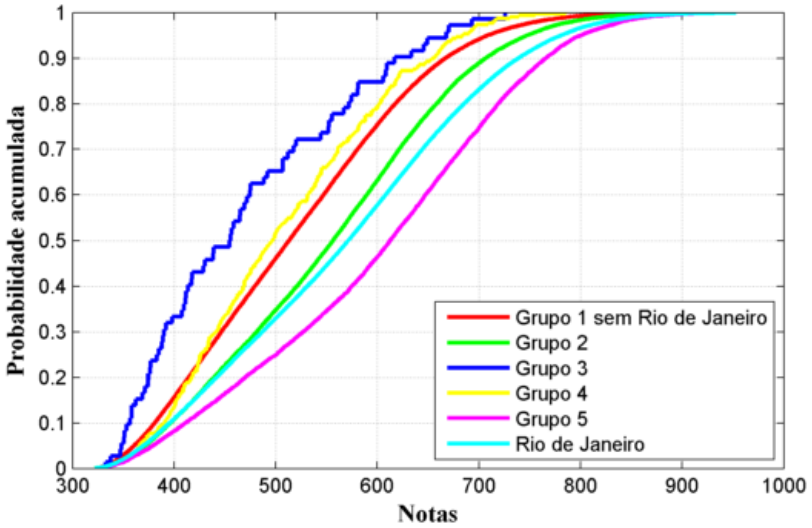


Figura 4.23 – Gráfico das fda's de todos os grupos isolando o Rio de Janeiro

Na Figura 4.23 temos as fdas do Rio de Janeiro e a do restante do grupo 1, além das dos outros grupos. Note que, a distribuição das notas do Rio de Janeiro se parece muito mais com as do grupo 2 do que as do grupo 1, entretanto a restrição espacial imposta pela matriz de vizinhança não permitiu que o município fizesse parte desse grupo. O município do Rio de Janeiro também poderia ter feito parte do grupo 5, porém a similaridade entre este e Niterói a partir do p-valor é muito baixa, fazendo com que ficassem em grupos diferentes.

Grupo 2:

Ao contrário do grupo 1, este grupo abrange praticamente todo o norte fluminense, com exceção apenas de alguns municípios da região serrana e Carapebus. Este é o grupo que mais possui cidades, com 53, que estão pintadas com a cor verde no mapa da Figura 4.18.

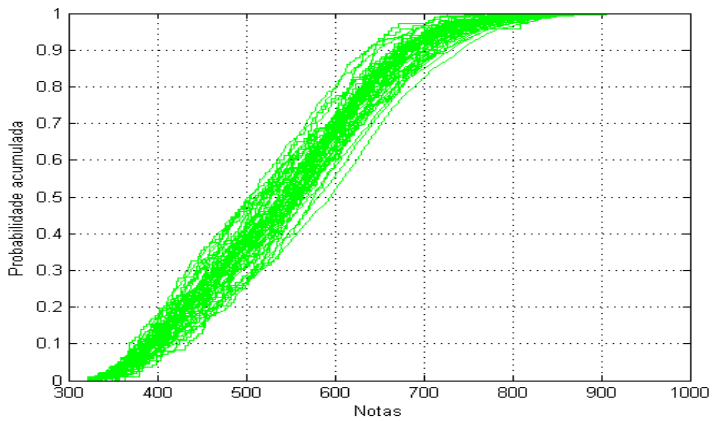


Figura 4.24 – Gráfico das fda's de todos os municípios do grupo 2

O gráfico da Figura 4.24 ilustra as fdas dos municípios do grupo 2. Este, por ser o mais abrangente, também é o menos homogêneo de todos. Escolhendo-se uma quantidade maior de grupos, a tendência é que permaneçam juntos apenas os municípios que estejam incluídos na faixa central. Mesmo sendo um grupo bastante abrangente, é possível ver que não existe nenhuma cidade que seja muito afastadas das demais, como ocorreu no grupo 1.

Novamente comparando com o grupo 1, já que os dois são os mais abrangentes, é possível notar que a fda do grupo 2 está mais deslocadas para a direita do que as do 1, com a exceção do Rio de Janeiro que é um outlier (Figura 4.22). Significa que o grupo 2 possui a maior concentração de notas num intervalo de valores maiores do que as do grupo 1, ou seja, o desempenho dos candidatos do grupo 2 foi superior aos do grupo 1.

Grupo 3:

O grupo 3 é representado por apenas um município, que é o de Carapebus. Ele aparece com a cor azul no mapa da Figura 4.18. Este município possui características bem diferentes de todos os outros, o que justifica deixá-lo em um grupo isolado. Também, é o município com a menor quantidade de candidatos e a menor média de notas. Observando a Figura 4.20, o grupo que mais se aproxima a este é o grupo 4, que além de não ser vizinho também possui distribuição diferente o bastante para mantê-lo isolado.

De acordo com o IBGE (2014), o município de Carapebus possui apenas uma escola que oferece turmas de ensino médio, sendo esta estadual. Isso pode justificar a baixa quantidade de candidatos deste município no ENEM e a qualidade das notas apresentadas por ele.

Grupo 4:

O grupo 4 é representado pelos municípios Macuco, São Sebastião do Alto e Trajano de Morais. Estas cidades são as que aparecem em amarelo no mapa da Figura 4.18. Este grupo é formado pelas três cidades que apresentam um desempenho muito baixo nas notas de matemática do ENEM para estar no grupo 2.

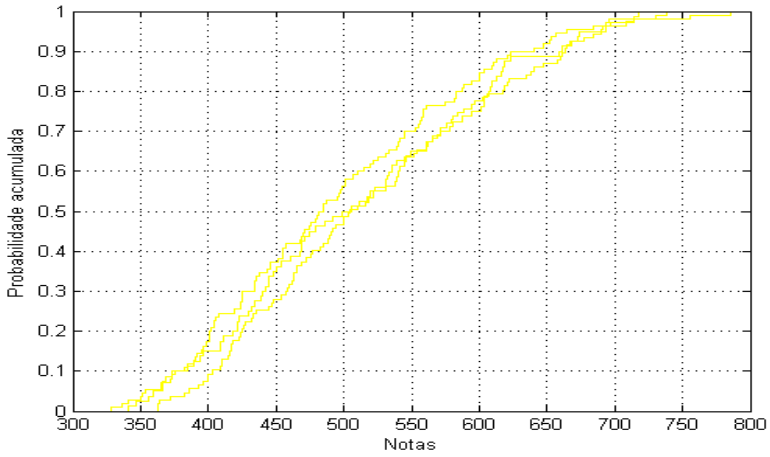


Figura 4.25 – Gráfico das fda's de todos os municípios do grupo 4

Na Figura 4.25, é possível ver que os três municípios possuem significativas diferenças entre si. Entretanto, as curvas não são suaves, o que significa que as amostras são pequenas. A quantidade de candidatos que realizaram o exame nesses municípios é dada na Tabela 4.7. de Macuco, São Sebastião de Alto e Trajano de Morais são respectivamente 80, 107 e 110.

Tabela 4.7 – Quantidade de Candidatos no grupo 4

	Quantidade de candidatos
Macuco	80
São Sebastião de Alto	107
Trajano de Morais	110

Desta forma, especula-se que as diferenças observadas no gráfico sejam decorrentes do tamanho das amostras. Para melhor avaliar esse grupo é preciso aplicar o teste de hipóteses de Kolmogorov-Smirnov para duas amostras neles, comparando-os dois à dois. Ao executar o método, obtem-se o seguinte resultado:

- Macuco e São Sebastião do Alto, p-valor = 0.815177 = 81.5177%;
- Macuco e Trajano de Morais, p-valor = 0.696102 = 69.6102%;
- São Sebastião do Alto e Trajano de Morais, p-valor = 0.221808 = 22,1808%.

O resultado mostra que os p-valores resultantes são relativamente altos, portanto,

nos três testes a hipótese nula não seria recusada, para os níveis de significância de 5% ou 1%. Ou seja, as distribuições dos três municípios podem ser consideradas parecidas, justificando que agrupá-los não seria um equívoco.

Grupo 5:

Assim como o grupo 3, o grupo 5 é representado por apenas um município, no caso Niterói. Este grupo é representado pela cor rosa no mapa da Figura 4.18. Niterói possui a maior média dentre todos os municípios do estado do Rio de Janeiro. A sua distribuição se diferencia bastante dos demais e isso foi definitivo para isolá-lo em um grupo. O gráfico das fda's apresentado na Figura 4.20 mostra que a curva deste grupo está bastante deslocada para a direita, o que mostra a grande concentração de notas mais altas.

Observando os mapas das Figura 4.18 e Figura 4.19, é possível observar que em ambos existe uma divisão idêntica entre os municípios do norte e do sul do estado, mostrando a grande semelhança entre os resultados. Contudo duas diferenças se destacam. A primeira é referente ao município do Rio de Janeiro e a outra é referente a alguns municípios do norte fluminense.

A Figura 4.26 mostra que no resultado da abordagem que utiliza as médias o município do Rio de Janeiro não fez parte do grande grupo 1, ficando com o de Niterói no grupo 5. No método proposto, Rio de Janeiro se manteve no grupo 1 e Niterói ficou sozinho no grupo 5.

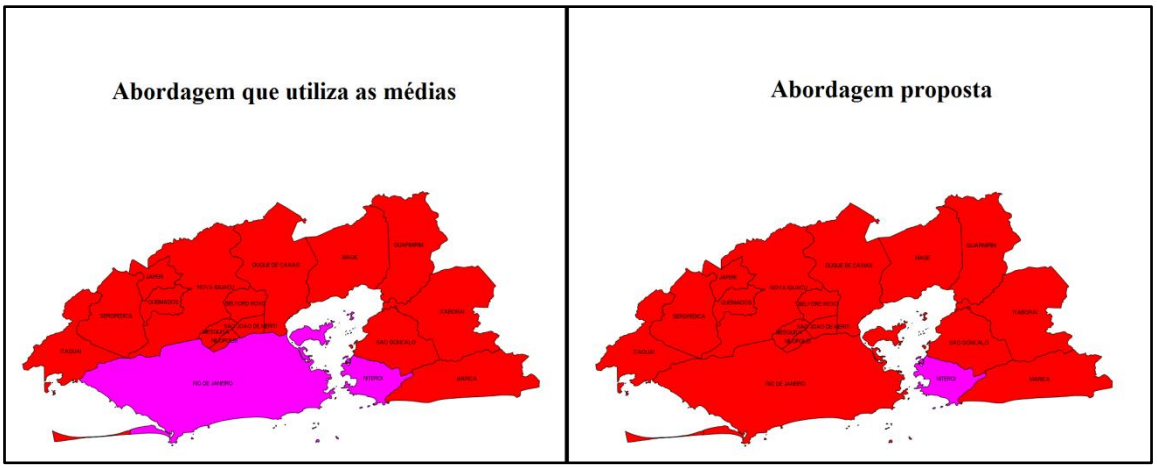


Figura 4.26 – Diferença do município do Rio de Janeiro entre as abordagens

Para identificar o que causou a diferença nos dois resultados, é necessário analisar os valores de similaridade obtidos entre Rio de Janeiro e cada um de seus adjacentes a partir de cada um dos métodos.

A Tabela 4.8 apresenta as comparações entre Rio de Janeiro e os municípios adjacentes a ele. A diferença entre as médias é utilizada na abordagem baseada em uma estatística da amostra, de forma que quanto menor a diferença maior o valor da similaridade. Assim, verifica-se que Niterói é o município que possui a média mais próxima do Rio de Janeiro, justificando o fato dos dois municípios pertencerem ao mesmo grupo na abordagem que utiliza a média. O p-valor é utilizado no método proposto como o próprio valor da similaridade, ou seja, quanto maior o p-valor mais parecidos eles são. Neste caso, é possível ver que Seropédica é o mais similar ao Rio de Janeiro. Além disso, Niterói apresenta uma das similaridades mais baixas entre os 8 vizinhos, fato que justifica a separação dos dois municípios no resultado do método proposto.

Tabela 4.8 – Comparação entre Rio de Janeiro e seus adjacentes

	Rio de Janeiro	
	Diferença entre as médias:	*P-valor:*
Duque de Caxias	$569.13 - 517.55 = 51.58$	$1.64257 \times e^{-317}$
Itaguaí	$569.13 - 523.92 = 45.21$	$2.09895 \times e^{-58}$
Mesquita	$569.13 - 520.40 = 48.73$	$3.28465 \times e^{-79}$
Nilópolis	$569.13 - 532.18 = 36.95$	$9.57618 \times e^{-48}$
Niterói	$601.01 - 569.13 = 31.88$	$5.57648 \times e^{-108}$
Nova Iguaçu	$569.13 - 519.24 = 49.89$	0.0
São João de Meriti	$569.13 - 517.46 = 51.67$	$6.53675 \times e^{-198}$
Seropédica	$569.13 - 521.02 = 48.11$	$8.0052 \times e^{-28}$

Analisando pelo ponto de vista educacional dos municípios envolvidos, observamos que Niterói apresenta um índice de educação muito superior à todos os municípios do estado, com 0.773. O município com segundo maior índice é Volta Redonda com 0.720, seguido por Rio de Janeiro com 0.719 e Nilópolis com 0.716. A Tabela 4.9 apresenta o índice de educação de todos os municípios que fazem fronteira com Rio de Janeiro, incluindo o próprio e Niterói, disponibilizados pelo IBGE (2014).

Tabela 4.9 - Índice de educação do Rio de Janeiro e seus adjacentes

	IDH-E
Niterói	0.773
Rio de Janeiro	0.719
Nilópolis	0.716
Mesquita	0.678
Seropédica	0.648
São João de Meriti	0.646
Nova Iguaçu	0.641
Itaguaí	0.638
Duque de Caxias	0.624

A Tabela 4.9 mostra que o índice de educação de Niterói é muito superior a todos listados, justificando o seu isolamento. Além disso, o índice do Rio de Janeiro está muito mais próximo do índice de Nílópolis e de Mesquita do que de Niterói.

A diferença entre os métodos no norte fluminense é apresentada pela Figura 4.27. Enquanto o método proposto gerou o grupo 4 com três municípios, o outro gerou o grupo 4 com oito municípios.

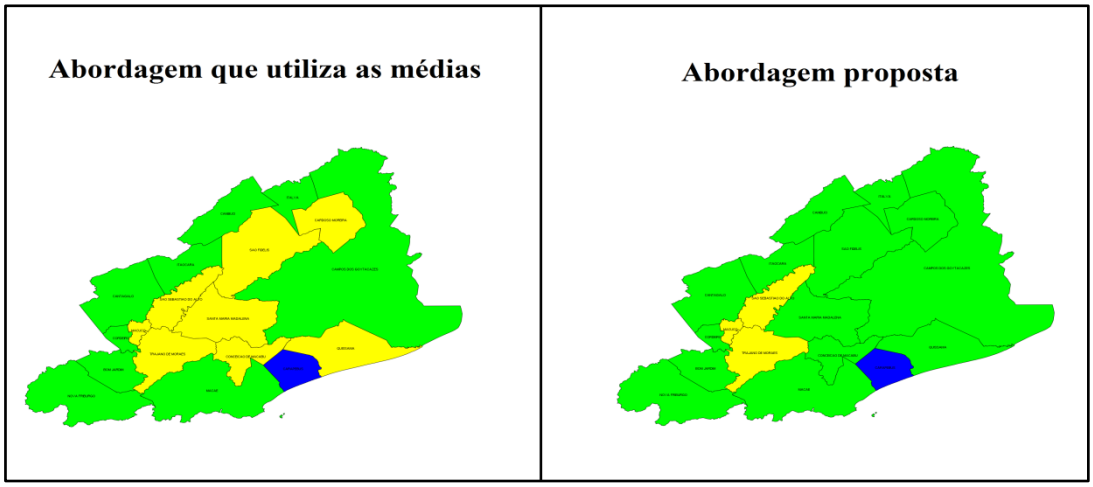


Figura 4.27 – Diferença nos municípios do norte fluminense entre os métodos

Essa grande diferença entre os dois resultados é compreendida a partir da análise dos municípios envolvidos. A seguir, será apresentada uma matriz contendo os p-valores referentes à todos estes municípios.

A Tabela 4.10 apresenta os p-valores dos municípios destacados pelo método baseado nas médias das notas, são eles Cardoso Moreira (CA), Conceição de Macabu (CM), Macuco (MA), Quissamã (QU), Santa Maria Madalena (SM), São Fidelis (SF), São Sebastião (SS), Trajano de Moraes (TR). Cada município foi pintado com a cor que aparece no resultado do método proposto. Além disso, para cada linha o maior p-valor foi destacado em vermelho.

Tabela 4.10 – P-valor entre os municípios destacados pelo método baseado nas médias das notas

	CA	CM	QU	SM	SF	MA	SS	TR
CA	1.0	0.0	0.0	0.0	0.2165	0.0	0.0	0.0
CM	0.0	1.0	0.7371	0.9972	0.0	0.0	0.0	0.0261
QU	0.0	0.7371	1.0	0.0	0.0	0.0	0.0	0.0
SM	0.0	0.9972	0.0	1.0	0.6931	0.0	0.6150	0.0678
SF	0.2165	0.0	0.0	0.6931	1.0	0.0	0.0972	0.0
MA	0.0	0.0	0.0	0.0	0.0	1.0	0.8152	0.6961
SS	0.0	0.0	0.0	0.6150	0.0972	0.8152	1.0	0.2218
TR	0.0	0.0261	0.0	0.0678	0.0	0.6961	0.2218	1.0

Analisando a Tabela 4.10, verifica-se que nenhum município apresentou o maior p-valor com um município de outro grupo, justificando o agrupamento obtido pelo método proposto.

Tabela 4.11 – Médias dos municípios em destaque e os seus adjacentes

	Média		Média
Bom Jardim	559.70	**Cardoso Moreira**	514.55
Cambuci	535.36	**Conceição de Macabu**	532.80
Campo dos Goytacazes	555.77	**Macuco**	510.67
Cantagalo	548.02	**Quissamã**	529.59
Cordeiro	569.32	**Santa Maria Madalena**	535,02
Italva	562.23	**São Fidelis**	539,44
Itaocara	562.82	**São Sebastião do Alto**	520,37
Macaé	561.34	**Trajano de Moraes**	495.58
Nova Friburgo	585.15		

A Tabela 4.11 contém as médias dos municípios do grupo 4 (cor azul) e seus adjacentes do grupo 2 (cor verde), separados pela abordagem que utiliza a média. A tabela evidencia a separação dos grupos a partir das médias, já que é facilmente notável

que os municípios do grupo 2 (verde) apresentam médias bem superiores as do grupo 4 (amarelo).

Analisando pelo ponto de vista dos índices de educação disponibilizados pelo IBGE (2014), podemos ver na Tabela 4.12 que há uma clara divisão entre os municípios com índice acima de 0.610 e os abaixo de 0.556, mostrando uma certa diferença entre os resultados obtidos por ambas abordagens e os dados estatísticos reais. Entretanto, alguns desses municípios não estão sendo bem representados devido a baixa quantidade de candidatos, fazendo com que, independente da abordagem utilizada, o resultado fuja um pouco dos dados reais. O principal exemplo deste problema é o município de Carapebus que apresentou a pior distribuição de notas, ficando muito abaixo de todos os outros, mas com índice de educação de 0.644, acima de todos os municípios apresentados na Tabela 4.12.

Tabela 4.12 - Índice de educação dos municípios do grupo 4 da abordagem que utiliza a média

	IDH-E
Cardoso Moreira	0.534
São Sebastião do Alto	0.536
Trajano de Moraes	0.543
Santa Maria Madalena	0.556
Quissamã	0.610
São Fidlis	0.611
Macuco	0.631
Conceição de Macabu	0.642

4.4 – Validação do método proposto

Nesta seção, será apresentado outro experimento realizado com intuito de verificar a validade do método proposto em uma base de dados gerada artificialmente. A ideia básica do método é calcular a distância entre as distribuições das amostras, mantendo as amostras mais próximas no mesmo grupo e as mais afastadas em grupos diferentes. Desta forma, se observarmos distribuições normais com o mesmo desvio padrão e apenas médias diferentes, a distância entre essas distribuições será maior quando a distância entre as médias também for maior. Neste caso, o método proposto ou um algoritmo de agrupamento baseado na distância das médias destas distribuições devem obter ao mesmo resultado.

Neste experimento, utilizamos novamente a base espacial dos municípios do estado do Rio de Janeiro. No entanto, as notas foram geradas artificialmente a partir de distribuições normais com o mesmo desvio padrão e médias diferentes. Novamente, as duas abordagens foram executadas e os resultados obtidos são apresentados nas Figura 4.28 e Figura 4.29.

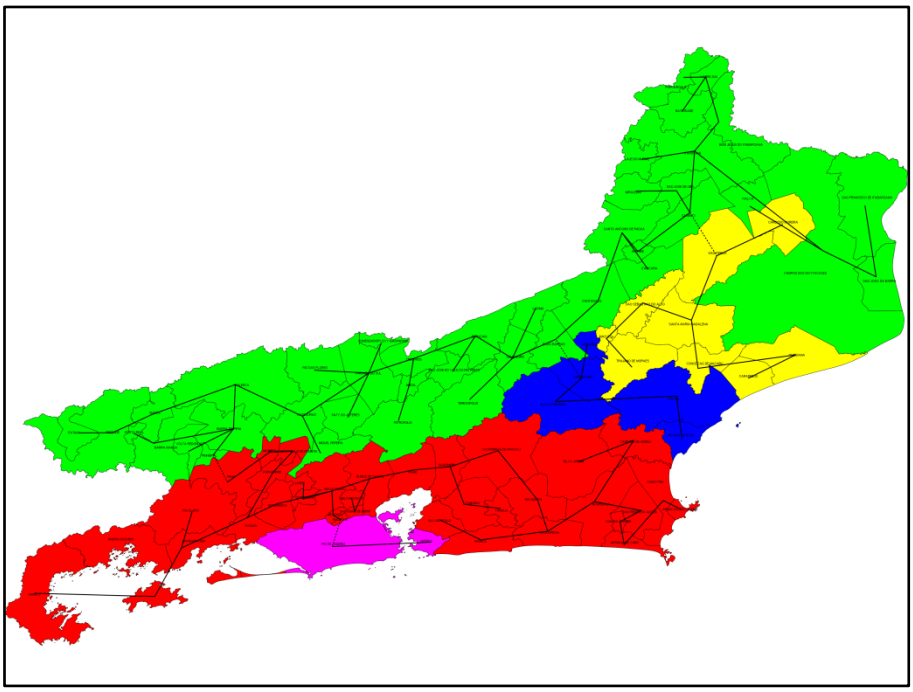


Figura 4.28 - Mapa do resultado do experimento de validação para o método proposto

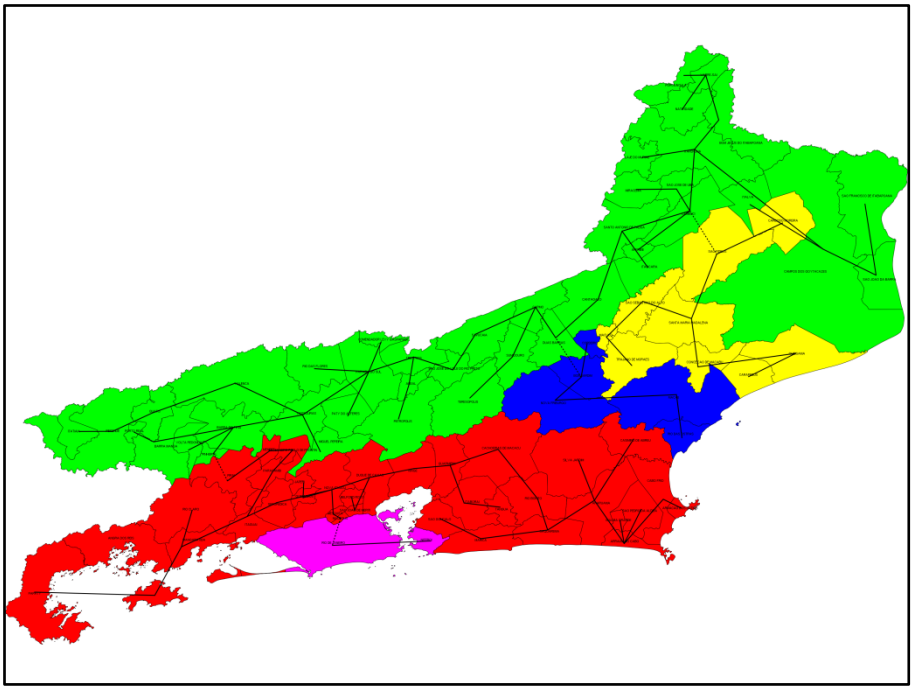


Figura 4.29 - Mapa do resultado do experimento de validação para a abordagem que utiliza a média

Como era esperado, as duas abordagens dividiram da mesma forma os municípios do estado. Isto mostra que o método proposto foi capaz de comparar as amostras de forma coerente, já que estas foram geradas a partir de distribuições normais que diferem apenas pela média. É possível ver que a árvore geradora mínima obtida pelos dois métodos são um pouco diferentes. Entretanto, isto ocorre por conta da imprecisão da amostra, que seria nula se as comparações fossem feitas a partir das populações.

4.5 – Considerações finais

Os experimentos descritos neste capítulo e as comparações realizadas na seção de análise de resultado, mostram que é viável utilizar o método proposto para o problema abordado.

O agrupamento obtido pela abordagem que utiliza uma estatística da amotra, mostrado na Figura 4.21, criou grupos de municípios mais parecidos do que os do método proposto, mostrado pela Figura 4.20, a partir das distâncias de suas funções. Além disso, gerou dois grupos com poucos municípios, mas que possuíam distribuições bem diferentes. Por outro lado, o método proposto permitiu o isolamento de municípios, como os de Niterói e Carapebus, com distribuições bem diferentes dos demais, criando assim grupos menos parecidos.

Os resultados mostraram que, ao utilizar a abordagem que calcula uma estatística da amostra, alguns municípios com características muito diferentes foram alocados no mesmo grupo, como Niterói e Rio de Janeiro. Este tipo de equívoco pode influenciar negativamente no estudo que se deseja fazer ao final do processo.

Capítulo 5 – Conclusão

Nesta dissertação, foi tratado o problema de aplicar um algoritmo de agrupamento em dados com atributos que possuam um conjunto de valores. Normalmente, calcula-se uma estatística desta amostra e passa-se a trabalhar com ela. Este procedimento pode não ser bom em situações em que não seja interessante comparar os dados a partir de uma estatística, mas sim por toda distribuição dos valores.

A abordagem sugerida nesta dissertação é a utilização do atributos com conjunto de valores e a aplicação do teste de Kolmogorov-Smirnov para fazer a comparação entre eles. Para isso, foi feita a revisão da literatura com o objetivo de buscar as principais metodologias aplicadas em análise de agrupamentos e definir a teoria fundamental para o desenvolvimento desta abordagem.

Um experimento foi realizado com o objetivo de comparar a qualidade dos agrupamentos formados pelas duas abordagens. Este experimento foi realizado utilizando a base de dados das notas de matemática do ENEM 2011 dos participantes do estado do Rio de Janeiro. Os resultados obtidos a partir deste experimento foram analisados e, por fim, concluiu-se que o método proposto por esta dissertação obtém agrupamentos de melhor qualidade com relação a distribuição dos valores dos atributos. Analisando a distribuição dos valores, é possível ver que este método gerou grupos com municípios mais parecidos entre si e diferentes, se comparados com os de outros grupos.

Por fim, os resultados obtidos no experimento serviram para mostrar que a abordagem proposta não é só viável, como também eficiente no tratamento objetos representados por um conjunto de instâncias.

5.1 – Trabalhos Futuros

Nesta seção, serão apresentados os principais trabalhos futuros possíveis. O trabalho futuro que merece mais destaque é a aplicação que se aborda nesta dissertação em dados com mais de um atributo. Em problemas de agrupamento, é normal que os dados possuam uma grande quantidade de atributos, portanto a aplicação da técnica em dados com mais de um atributo seria mais eficaz. Além disso, a utilização dos dados de todo o Brasil e eliminação da restrição de contiguidade. Outro trabalho que se pode

destacar é a utilização de outros métodos estatísticos, como o teste de duas amostras de Cramér–von Mises apresentado em ANDERSON (1962). Por fim, pode-se destacar a aplicação da proposta abordada, em problemas de agrupamento que não estejam relacionados com regionalização.

Referências

AGRAWAL, Rakesh; GEHRKE, J.; GUNOPULOS, D.; RAGHAVAN, P. **Automatic subspace clustering of high dimensional data for data mining applications**. ACM, 1998.

ALDENDENFER, Marks; BLASHFIELD, R. Cluster analysis. **Beverly Hills, CA: Sage**, 1984.

AMBROISE, Christophe; DANG, Mo. Spatial Data Clustering. **Data Analysis**, p. 289-318, 2009.

ANDERSON, T. W. et al. On the distribution of the two-sample Cramer-von Mises criterion. **The Annals of Mathematical Statistics**, v. 33, n. 3, p. 1148-1159, 1962.

ANKERST, Mihael; BREUNING, M. M.; KRIEGEL, H. P.; SANDER, J. Optics: Ordering points to identify the clustering structure. In: **ACM SIGMOD Record**. ACM, 1999. p. 49-60.

ARORA, Sanjeev; BARAK, Boaz. **Computational complexity: a modern approach**. Cambridge University Press, 2009.

ASSUNÇÃO, Renato M.; LAGE, Juliano P.; REIS, Edna. Análise de conglomerados espaciais via árvore geradora mínima. **Revista Brasileira de Estatística**, v. 63, n. 220, p. 7-24, 2002.

BERKHIN, Pavel. A survey of clustering data mining techniques. In:**Grouping multidimensional data**. Springer Berlin Heidelberg, 2006. p. 25-71.

CASELLA, George; BERGER, Roger L. **Statistical inference**. Belmont, CA: Duxbury Press, 1990.

CHANDRA, E.; ANURADHA, V. P. A Survey on Clustering Algorithms for Data in Spatial Database Management Systems. **International Journal of Computer Applications**, v. 24, 2011.

CHATURVEDI, Anil; GREEN, Paul E.; CAROLL, J. Douglas. K-modes clustering. **Journal of Classification**, v. 18, n. 1, p. 35-55, 2001.

DE MELLO, Carlos Eduardo Ribeiro. **Agrupamento de regiões: Uma abordagem utilizando acessibilidade**. 2008. Dissertação de Mestrado. UNIVERSIDADE FEDERAL DO RIO DE JANEIRO.

E.M. Mirkes, K-means and K-medoids applet. University of Leicester, 2011 . Disponível em: <http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html>. Acessado em: janeiro de 2014.

ESTER, Martin et al. Spatial data mining: database primitives, algorithms and efficient DBMS support. **Data Mining and Knowledge Discovery**, v. 4, n. 2-3, p. 193-216, 2000.

ESTER, Martin; KRIEGEL, H. P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **KDD**. 1996. p. 226-231.

ESTER, Martin; KRIEGEL, Hans-Peter; SANDER, Jörg. Algorithms and applications for spatial data mining. **Geographic Data Mining and Knowledge Discovery**, v. 5, n. 6, 2001.

EVERITT, B. S.; LANDAU, S.; LEESE, M. Cluster analysis. 2001. **Arnold, London**, 2001.

GOEL, Manuj; SHANKAR, Akshat. 2012. Hypothesis Testing. Disponível em: <http://simplifyingstats.com/data/HypothesisTesting.pdf>. Acessado em: Janeiro de 2014.

GOIL, Sanjay; NAGESH, Harsha; CHOUDHARY, Alok. MAFIA: Efficient and scalable subspace clustering for very large data sets. In: **Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. 1999. p. 443-452.

GUHA, Sudipto; RASTOGI, Rajeev; SHIM, Kyuseok. CURE: an efficient clustering algorithm for large databases. In: **ACM SIGMOD Record**. ACM, 1998. p. 73-84.

GUHA, Sudipto; RASTOGI, Rajeev; SHIM, Kyuseok. ROCK: A robust clustering algorithm for categorical attributes. In: **Data Engineering, 1999. Proceedings., 15th International Conference on**. IEEE, 1999. p. 512-521.

GÜTING, Ralf Hartmut. An introduction to spatial database systems. **The VLDB Journal—The International Journal on Very Large Data Bases**, v. 3, n. 4, p. 357-399, 1994.

HALKIDI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2-3, p. 107-145, 2001.

HAN, Jiawei; CAI, Yandong; CERCONE, Nick. Knowledge discovery in databases: An attribute-oriented approach. In: **VLDB**. 1992. p. 547-559.

HAN, Jiawei; KAMBER, Micheline. **Data mining: concepts and techniques**. Morgan kaufmann, 2006.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining: concepts and techniques**. Morgan kaufmann, 2011.

HINNEBURG, Alexander; KEIM, Daniel A. An efficient approach to clustering in large multimedia databases with noise. In: **KDD**. 1998. p. 58-65.

IBGE, 2014, "Instituto Brasileiro de Geografia e Estatística". Disponível em: <http://www.ibge.gov.br/home/>. Acessado em: Janeiro de 2014.

INEP, 2014, "Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira". Disponível em: <http://portal.inep.gov.br/>. Acessado em: Janeiro de 2014.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651-666, 2010.

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. **ACM computing surveys (CSUR)**, v. 31, n. 3, p. 264-323, 1999.

JAIN, Anil; NANDAKUMAR, Karthik; ROSS, Arun. Score normalization in multimodal biometric systems. **Pattern recognition**, v. 38, n. 12, p. 2270-2285, 2005.

JIAWEI, Han; KAMBER, Micheline. Data mining: concepts and techniques.**San Francisco, CA, itd: Morgan Kaufmann**, v. 5, 2001.

JOLLIFFE, Ian. **Principal component analysis**. John Wiley & Sons, Ltd, 2005.

KARYPIS, George; HAN, Eui-Hong; KUMAR, Vipin. Chameleon: Hierarchical clustering using dynamic modeling. **Computer**, v. 32, n. 8, p. 68-75, 1999.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data: an introduction to cluster analysis. **Wiley series in probability and mathematical statistics (Applied probability and statistics**, 2005.

LAWSON, Andrew B.; DENISON, David GT (Ed.). **Spatial cluster modelling**. CRC press, 2002.

LEHMANN, Erich L.; ROMANO, Joseph P. **Testing statistical hypotheses**. springer, 2006.

MACQUEEN, James. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. 1967. p. 14.

MAPSERVER, 2014, "MapServer v6.4.1". Disponível em: <http://mapserver.org/>. Acessado em: janeiro de 2014.

MONTGOMERY, Douglas C.; RUNGER, George C. **Applied Statistics and Probability for Engineers 6th edition**. Wiley, 2003.

NEVES, Marcos Corrêa et al. Procedimentos automáticos e semi-automáticos de regionalização por árvore geradora mínima. In: **Simpósio Brasileiro de Geoinformática, GeoInfo**. 2002. p. 2002.

NEVES, Marcos Corrêa. **Procedimentos Eficientes para Regionalização de Unidades Socioeconômicas em Bancos de Dados Geográficos.** 2003. Tese de Doutorado. INPE, São José dos Campos, SP, Brasil.

NG, Raymond T.. ; HAN, Jiawei. CLARANS: A method for clustering objects for spatial data mining. **Knowledge and Data Engineering, IEEE Transactions on**, v. 14, n. 5, p. 1003-1016, 2002.

NG, Raymond T.; HAN, Jiawei. Efficient and Effective Clustering Methods for Spatial Data Mining. 1994.

OPENSHAW, Stan; RAO, Liang. **Re-engineering 1991 census geography: serial and parallel algorithms for unconstrained zone design**. School of Geography, University of Leeds, 1994.

PELLEG, Dan; MOORE, Andrew W. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: **Proceedings of the Seventeenth International Conference on Machine Learning**. Morgan Kaufmann Publishers Inc., 2000. p. 727-734.

POSTGIS, 2014, "PostGIS v.2.1", Disponível em: <http://postgis.net/install>. Acessado em: Janeiro 2014.

POSTGRESQL, 2014, "PostgreSQL v.9.3". Disponível em: <http://www.postgresql.org/download/>. Acessado em: Janeiro de 2014.

RAZALI, Nornadiah Mohd; WAH, Yap Bee. Power comparisons of shapiro-wilk, Kolmogorov-Smirnov, lilliefors and anderson-darling tests. **Journal of Statistical Modeling and Analytics**, v. 2, n. 1, p. 21-33, 2011.

REYNOLDS, Alan P.; RICHARDS, Graeme; RAYWARD-SMITH, Vic J. The application of k-medoids and pam to the clustering of rules. In: **Intelligent Data Engineering and Automated Learning–IDEAL 2004**. Springer Berlin Heidelberg, 2004. p. 173-178.

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, p. 53-65, 1987.

SCHIKUTA, Erich; ERHART, Martin. The BANG-clustering system: Grid-based data analysis. In: **Advances in Intelligent Data Analysis Reasoning about Data**. Springer Berlin Heidelberg, 1997. p. 513-524.

SHALABI, Luai A.; SHAABAN, Zyad; KASASBEH, Basel. Data mining: A preprocessing engine. **Journal of Computer Science**, v. 2, n. 9, p. 735, 2006.

SHEIKHOLESLAMI, Gholamhosein; CHATTERJEE, Surojit; ZHANG, Aidong. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In: **VLDB**. 1998. p. 428-439.

SHEKHAR, Shashi et al. Spatial databases-accomplishments and research needs. **Knowledge and Data Engineering, IEEE Transactions on**, v. 11, n. 1, p. 45-55, 1999.

SKATER, 2014, "SKATER (Spatial 'K'luster Analysis by Tree Edge Removal)". Disponível em: <http://www.est.ufmg.br/leste/skater.htm>. Acessado em: Janeiro de 2014.

SONI, Neha; GANATRA, Amit. Categorization of Several Clustering Algorithms from Different Perspective: A Review. **International Journal**, v. 2, n. 8, 2012.

STATA. 2014. Disponível em: <http://www.stata.com/manuals13/rksmirnov.pdf>. Acessado em: janeiro 2014.

STEINBACH, Michael et al. A comparison of document clustering techniques. In: **KDD workshop on text mining**. 2000. p. 525-526.

WANG, Wei; YANG, Jiong; MUNTZ, Richard. STING: A statistical information grid approach to spatial data mining. In: **VLDB**. 1997. p. 186-195.

XU, Rui; DONALD WUNSCH, I. I. Survey of Clustering Algorithms. **IEEE TRANSACTIONS ON NEURAL NETWORKS**, v. 16, n. 3, p. 645, 2005.

ZHANG, Tian; RAMAKRISHNAN, Raghu; LIVNY, Miron. BIRCH: an efficient data clustering method for very large databases. In: **ACM SIGMOD Record**. ACM, 1996. p. 103-114.