Universidade Federal da Bahia - UFBA
Instituto de Matemática e Estatística - IME
Programa de Pós-Graduação em Matemática - PGMAT
Dissertação de Mestrado

Método de agrupamento multinível para dados mistos

Hellen Oliveira da Paz

Área de Concentração: Estatística

Salvador - Bahia
Junho de 2024

Método de agrupamento multinível para dados mistos

Hellen Oliveira da Paz

Dissertação de Mestrado apresentada ao Colegiado da Pós-Graduação em Matemática da Universidade Federal da Bahia (UFBA), como parte dos requisitos para obtenção do título de Mestre em Matemática. Área de concentração: Estatística.

Orientador: Prof. Dr. Anderson Luiz Ara Souza.
Coorientadora: Profª. Drª. Rosemeire Leovigildo Fiaccone.

Salvador - Bahia
Junho de 2024

Método de agrupamento multinível para dados mistos

Hellen Oliveira da Paz

Dissertação apresentada ao Colegiado do Curso de Pós-graduação em Matemática da Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Matemática.

Banca examinadora

gov.br Documento assinado digitalmente
ANDERSON LUIZ ARA SOUZA
Data: 12/07/2024 09:37:27-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Anderson Luiz Ara Souza (orientador - UFPR)

gov.br Documento assinado digitalmente
LILIA CAROLINA CARNEIRO DA COSTA
Data: 16/07/2024 14:22:58-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Lilia Carolina Carneiro da Costa (UFBA)

gov.br Documento assinado digitalmente
MARCELO RODRIGO PORTELA FERREIRA
Data: 15/07/2024 14:14:00-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcelo Rodrigo Portela Ferreira (UFPB)

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

P348 Paz, Hellen Oliveira da

Método de agrupamento multinível para dados mistos / Hellen Oliveira da Paz. – Salvador, 2024.

100 f.

Orientador: Prof. Dr. Anderson Luiz Ara Souza
Coorientadora: Profª. Drª. Rosemeire Leovigildo Fiaccone

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática e Estatística (IME), 2024.

1. Análise de agrupamento. 2. Agrupamento multinível. 3. *K-means*. 4. Dados multiníveis. 5. Dados mistos. I. Souza, Anderson Luiz Ara. II. Fiaccone, Rosemeire Leovigildo. III. Universidade Federal da Bahia. IV. Título.

CDU:510

Eu dedico este trabalho a todos que estiveram ao meu lado e sempre acreditaram que eu era capaz, mesmo quando nem eu acreditava. Em especial, aos meus pais.

Agradecimentos

À Deus, por Sua presença constante em minha vida, por guiar o meu caminho, por me fortalecer nos momentos que pareciam mais difíceis, por mais esta benção alcançada.

Aos meus pais, Lucineia e Admilson, por serem a minha base. Pai, obrigada por ser parceiro, pela tranquilidade e por me ensinar que dias melhores sempre virão. Mãe, obrigada pelo amor sem limites, pelo zelo e por me impulsionar na direção dos meus sonhos. Sou eternamente grata por todo amor, esforço, dedicação, conselho, incentivo e por sempre acreditarem em mim. Sem vocês eu não teria conseguido chegar até aqui!

Aos meus amados e queridos irmãos, pela amizade, apoio, incentivo e parceria de vida. Harllon, sempre muito sensato, obrigada por ouvir cada desabafo, por sempre me fazer ver uma luz no fim do túnel e me passar Esperança. Hallison, sempre muito perspicaz, obrigada por não se limitar, por sempre me fazer enxergar além (inclusive, enxergar que o mestrado poderia, sim, ser algo para mim) e me passar Confiança. Vocês me enchem de orgulho e são grandes inspirações! Um agradecimento especial à Kizzy, que me fez recordar o poder da Fé e que sempre tem uma dose de aprendizado de diversas esferas para compartilhar. Você é luz!

À Fernando, meu companheiro de todas as horas, pelo apoio, compreensão, amor, generosidade e paciência nos momentos de desalento. Você me fez ressignificar muitas coisas e me fortaleceu para que pudesse alcançar mais este objetivo. Eu realmente tenho muita sorte de ter uma pessoa tão especial na minha vida e, com certeza, este trabalho não seria o mesmo sem você!

Aos amigos, de perto ou de longe, agradeço a amizade, a compreensão pelas faltas em alguns (muitos) momentos, a força e estímulo que sempre me foram dados. Vocês sabem que estarão para sempre no meu coração.

Aos meus colegas de curso, pelas boas conversas, momentos de descontração e contribuições nos estudos, o que fizeram essa caminhada muito mais agradável. Obrigada Sandro e Nayguel pela partilha de angústias, apoio, incentivo, paciência e por sempre estarem dispostos a dividir suas experiências e conhecimentos. Um agradecimento especial ao José Guilherme, pela amizade consolidada, vídeochamadas, horas de estudos, conversas, desabafos e por sempre ter uma palavra de conforto e incentivo nessa jornada, por

vezes, árdua e solitária, principalmente em meio a pandemia da COVID-19.

Ao meu orientador, Anderson, pela paciência, disponibilidade e prontidão para aconselhar, orientar e debater ideias, sempre com muito entusiasmo e espírito de colaboração. À minha coorientadora, Rosemeire, por aceitar participar desse trabalho, por todo apoio e pelas sugestões que muito contribuíram. Também, gostaria de deixar registrado o meu reconhecimento e admiração por esses dois exemplos de profissionais com o qual tive a oportunidade de conviver nesse período e que sempre me lembrarei com carinho. Me deram a oportunidade de realizar este trabalho que muito me desafiou nas mais diversas esferas da vida e que me fez transpor muitas questões pessoais e profissionais.

Aos demais professores do PGMAT, em especial do DEST-UFBA, pelo comprometimento, dedicação e partilha de conhecimento bem como por se desdobrarem para manter a qualidade e excelência do ensino ao longo da pandemia da COVID-19 no modelo remoto.

À banca examinadora, composta pelos professores Anderson, Lilia e Marcelo, por disponibilizarem seu tempo para avaliar este trabalho e pelas valiosas contribuições.

À todos os funcionários da UFBA pela disposição e suporte.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

E, por fim, à todos que direta ou indiretamente contribuíram para a conclusão de mais essa etapa.

A sua coragem depende da sua habilidade de ser pouco ou muito vulnerável.
Brené Brown

Resumo

A Análise de agrupamento é uma área com vasto desenvolvimento metodológico nas diversas áreas do conhecimento. Esta dissertação propõe um novo método de agrupamento para dados mistos, levando em consideração a estrutura multinível das observações. A identificação de quão similares ou próximas as unidades de análise se encontram pode ser quantificada por meio de medidas de proximidade, que, juntamente com os algoritmos utilizados, são essenciais na metodologia de análise de agrupamento. Dados mistos são caracterizados pela presença conjunta de variáveis quantitativas e qualitativas. O termo "Agrupamento Multinível" é utilizado em diversas áreas do conhecimento, referindo-se a diferentes conceitos. Nossa proposta de agrupamento multinível adapta o algoritmo k-médias para dados multiníveis, incorporando a estrutura hierárquica dos dados no cálculo das distâncias entre as observações através de uma abordagem de ponderação da distância de Hellinger. Os resultados obtidos a partir de estudos de simulação e aplicações práticas são satisfatórios, apresentando melhores agrupamentos quando se tem mais de uma variável quantitativa. No entanto, ainda são necessários mais estudos em diversos cenários para aumentar a robustez da metodologia proposta.

Palavras-chave: Análise de agrupamento, agrupamento multinível, *k-means*, dados multiníveis, dados mistos.

Abstract

Cluster Analysis is an area with vast methodological development in different areas of knowledge. This dissertation proposes a new clustering method for mixed data, taking into account the multilevel structure of observations. The identification of how similar or close the units of analysis are can be quantified through proximity measures, which, together with the algorithms used, are essential in the cluster analysis methodology. Mixed data is characterized by the joint presence of quantitative and qualitative variables. The term "Multilevel Clustering" is used in different areas of knowledge, referring to different concepts. Our multilevel clustering proposal adapts the k-means algorithm to multilevel data, incorporating the hierarchical structure of the data in calculating the distances between observations through a Hellinger distance weighting approach. The results obtained from simulation studies and practical applications are satisfactory, presenting better groupings when there is more than one quantitative variable. However, more studies are still needed in different scenarios to increase the robustness of the proposed methodology.

Key words: Cluster analysis, multilevel clustering, k-means, multilevel data, mixed data.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AA Análise de Agrupamento

CES Censo da Educação Superior

DU Índice de avaliação de Dunn

IES Instituições de Ensino Superior

Inep Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

WSS *Whithin-cluster Sum of Square*

Capítulo 1

Introdução

A Análise de Agrupamento, também conhecida como Análise de *Cluster*, é uma metodologia da Estatística Multivariada, campo da Estatística que lida com a análise simultânea de diversas variáveis mensuradas em cada unidade de análise (ou observação). De acordo com Mingoti (2007), essa metodologia consiste em dividir as unidades de análise da amostra, ou população, em grupos de tal forma que, simultaneamente, unidades de análise pertencentes a um mesmo grupo sejam homogêneas entre si e heterogêneas entre grupos distintos.

Diversas áreas utilizam a análise de agrupamento nos mais diversos contextos. Na biologia, zoologia, medicina, psiquiatria, sociologia, geologia, geografia (SCOLTOCK, 1982), no *marketing* (PUNJ; STEWART, 1983), na psicologia (SPEECE; MCKINNEY; APPELBAUM, 1985), na segmentação de imagens de documentos (TAXT; FLYNN; JAIN, 1989), na recuperação de informação (SALTON, 1991), na ecologia (MCGARIGAL; CUSHMAN; STAFFORD, 2000) etc.

A Análise de Agrupamento também é conhecida por fazer parte do Aprendizado de Máquina Não Supervisionado. O aprendizado de máquina pode ser dividido em aprendizado supervisionado e não supervisionado, em que o supervisionado tem interesse na predição da variável resposta a partir de covariáveis e o não supervisionado retratar associações a partir das variáveis. Em razão disso, alguns autores consideram a análise de agrupamento como um método de classificação não supervisionado, uma vez que a atribuição dos rótulos (*labels*) é realizada através de abordagem baseada nos dados observados (*data driven*).

A identificação de quão similares ou próximas as unidades de análise se encontram pode ser quantificada através de medidas de dissimilaridade ou similaridade, ambas conhecidas como medidas de proximidade. A forma como o agrupamento deve ser realizado é feita pelos algoritmos, que determinam como as observações serão associadas com base nas medidas calculadas. Existem diversos algoritmos disponíveis, cada um com suas próprias

características. Assim sendo, as medidas de proximidade e os algoritmos utilizados são, essencialmente, os pontos-chaves da metodologia de análise de agrupamento.

A escolha da medida de proximidade e do algoritmo a serem utilizados tem relação direta com os tipos das variáveis envolvidas no estudo, podendo essas serem divididas em quantitativas (numéricas) ou qualitativas (categóricas). De acordo com Bussab e Morettin (2017), as variáveis quantitativas são resultados de uma contagem ou mensuração e podem ser divididas em: (I) discretas, em que os possíveis valores formam um conjunto finito ou enumerável; (II) contínuas, em que possíveis valores pertencem a um intervalo de números reais. Já as variáveis qualitativas, representam uma qualidade ou atributo do elemento em investigação, podendo ser: (I) nominal, em que não existe uma ordenação das possíveis categorias; (II) ordinal, em que existe uma ordem natural das possíveis categorias.

Para o caso de variáveis quantitativas, existe uma vasta discussão na literatura de análise de agrupamento, sendo o algoritmo k-médias (k-*means*) proposto por MACQUEEN et al. em 1967 um dos mais conhecidos. Jain (2010) apresenta uma visão geral do algoritmo abordando alguns pontos principais, dentre os quais: o histórico, as limitações e contribuições, melhorias e perspectivas futuras ao longo de 50 anos. De acordo com PITA (2019), a maioria dos algoritmos para dados categóricos se concentram em melhorar parte específica do algoritmo k-médias original, tornando-se extensões. Dentre os quais citam-se o k-modas (HUANG, 1998), que utiliza a moda para o processo de atualização de centroides no caso de variáveis qualitativas, e o k-protótipo (HUANG, 1997), que unifica o k-médias e o k-modas para o caso de dados quantitativos e qualitativos conjuntamente.

Na era atual de uma vasta quantidade de dados disponíveis, tanto em termos de observações quanto de variáveis, ressalta-se a ocorrência de mistura de tipos de variáveis. Dados mistos são caracterizados pela presença conjunta tanto de variáveis com escala de mensuração quantitativa quanto qualitativa. Esse tipo de dado ocorre com frequência nas diversas áreas do conhecimento, no entanto calcular a similaridade de dois pontos de dados se torna mais difícil configurando, então, um desafio para agrupar esse tipo de estrutura de dados.

Outro ponto a salientar é a possibilidade de diferente fontes de heterogeneidade entre as observações de variáveis com escala quantitativa e qualitativa, refletindo o desenho do estudo ou a estrutura dos dados. Por exemplo, dados com estrutura multinível são muito comuns em estudos que envolvem indivíduos dentro de comunidades, escolas, organizações, regiões geográficas etc. Nestes casos os indivíduos estão agrupados em unidades de nível mais baixo, que, por sua vez, se relacionam às unidades de um nível mais alto. Segundo Laros e Marciano (2008), os níveis podem ser considerados como agregados que exercem efeito sobre os comportamentos dos seus membros. Como exemplo, estudantes podem constituir o nível inferior, enquanto a escola em que estão inseridos definiria

o nível superior. Outra nomenclatura para as unidades nos níveis 1 e 2 são unidades a nível micro e unidades a nível macro ou conglomerado, respectivamente. Vale ressaltar que a estrutura multinível pode ser decorrente do delineamento da pesquisa ou mesmo pela visualização da referida estrutura de forma a utilizar procedimentos de análise mais sofisticados como, por exemplo, modelos que incorporem essa diferente fonte de heterogeneidade. Diversos desafios têm surgido na literatura para análise de dados com estrutura multinível (PEUGH, 2010).

Quando existe uma estrutura hierárquica na população de interesse, a análise multinível é a opção metodologicamente correta para estabelecer as relações entre as variáveis (LAROS; MARCIANO, 2008). A importância da estrutura multinível está na possibilidade de analisar a inter-relação existente entre os diferentes níveis, além de levar em consideração a variabilidade associada a cada nível de agrupamento. Por exemplo, na área de Ciências Sociais pode ser de interesse explorar o impacto do contexto (uma determinada condição) na resposta de interesse no nível 1. A noção de contexto é bastante geral e pode incluir contextos espaciais (estado, comunidade), temporais (história), organizacionais (escolas, hospitais) e socioeconômicos (classe social, grupo de etnia). Entretanto, as interações entre as unidades de análise e os conglomerados, aos quais os indivíduos pertencem, consistem não apenas nos impactos que os conglomerados têm sobre os indivíduos, como também as influências que os indivíduos causam nos grupos.

A terminologia Agrupamento Multinível, ou *Multilevel clustering*, é utilizada nas mais diversas áreas do conhecimento, porém referindo-se a conceitos diversos. Para a Estatística, o termo multinível é comumente empregado com a modelagem multinível (ou modelagem de efeito misto), que é uma extensão do modelo de regressão tradicional e que visa incorporar a variabilidade dos diferentes níveis, desde o nível micro até o macro. No entanto, para o contexto de análise de agrupamento, sobretudo para dados com estrutura multinível, percebe-se uma carência de literatura, principalmente no que tange dados mistos.

1.1 Objetivos

O presente trabalho tem como objetivo geral propor um novo método de agrupamento de dados mistos levando em consideração a estrutura multinível das observações[1]. Como objetivos específicos tem-se:

- Descrever algumas medidas de proximidade para dados quantitativos, qualitativos e mistos, bem como a composição delas;

[1] Acredita-se que o método seja inédito, pois não foram encontradas informações a respeito de tal conjuntura em pesquisas na literatura.

- Descrever alguns métodos de agrupamento para dados quantitativos, qualitativos e mistos;
- Descrever sobre dados com estrutura multinível;
- Definir uma proposta de agrupamento multinível para dados mistos;
- Avaliar o desempenho da metodologia proposta;
- Implementar a metodologia proposta em dados reais.

1.2 Organização do trabalho

Este trabalho divide-se em seis capítulos, entre os quais inclui-se este capítulo inicial no qual é apresentado, sumariamente, o problema de pesquisa. No Capítulo 2 é realizada uma revisão da literatura sobre os conceitos fundamentais em Análise de Agrupamento. Nesse capítulo, é apresentada a noção de proximidade e as principais medidas de dissimilaridade e similaridade encontradas na literatura para as situações envolvendo dados quantitativos, qualitativos e mistos. Discorre-se, por exemplo, sobre a distância Euclidiana, o coeficiente de Jaccard e o coeficiente de Gower, bem como exemplifica-se o cálculo dessas e outras medidas. Também, são sumarizados alguns métodos de agrupamento usualmente utilizados, tendo maior enfoque para o algoritmo k-médias (k-*means*). A Análise de Agrupamento Multinível é abordada no Capítulo 3, onde é apresentada uma visão geral do termo utilizado, a caracterização de estrutura multinível das observações e a definição da abordagem proposta. Já no Capítulo 4 são apresentados os resultados da avaliação de desempenho da metodologia apresentada por meio do estudo de simulação. No Capítulo 5 são exibidos os resultados da aplicação da metodologia proposta em dois cenários de aplicação. Por fim, no Capítulo 6, são apresentadas algumas considerações finais e perspectivas de trabalhos futuros.

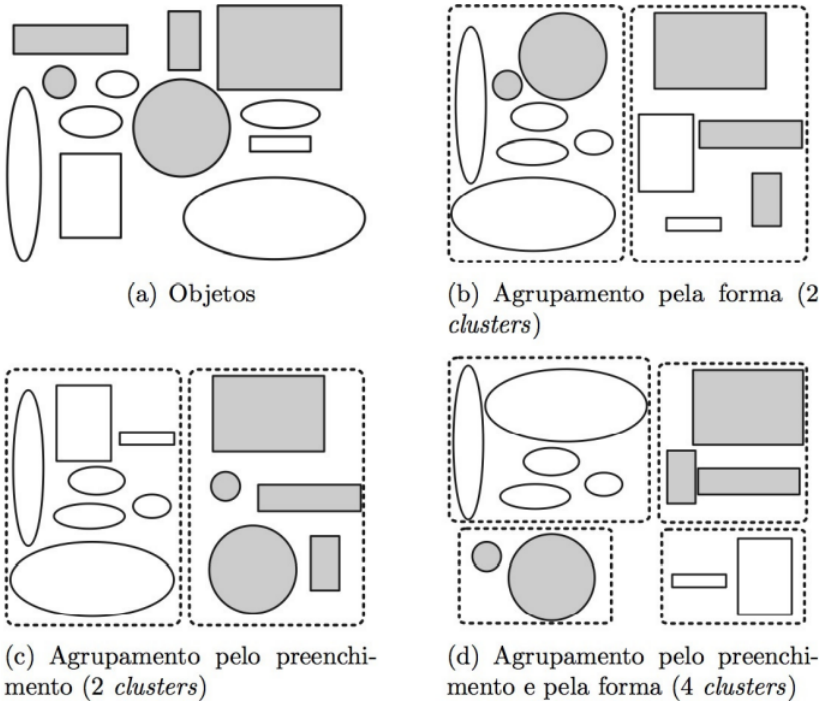
Capítulo 2

Análise de Agrupamento

2.1 Introdução

A Análise de Agrupamento tem como essência encontrar grupos nos dados a partir de padrões de comportamento similares. Dessa forma, tem-se homogeneidade entre as unidades de análises de cada grupo e heterogeneidade entre as unidades de análises dos diferentes grupos. Para ilustração, considere a Figura 2.1 em que inicialmente tem-se a coleção de objetos que se deseja agrupar, nela é possível observar três formas de agrupamento: (I) pelo formato; (II) pelo tipo de preenchimento; e (III) pela combinação do formato e do tipo de preenchimento. O critério utilizado para determinar até que ponto as unidades de análise podem ser consideradas semelhantes ou não depende da perspectiva adotada.

Figura 2.1: Exemplificação da ideia de agrupamento de objetos.



(a) Objetos

(b) Agrupamento pela forma (2 *clusters*)

(c) Agrupamento pelo preenchimento (2 *clusters*)

(d) Agrupamento pelo preenchimento e pela forma (4 *clusters*)

Fonte: Faceli (2011)

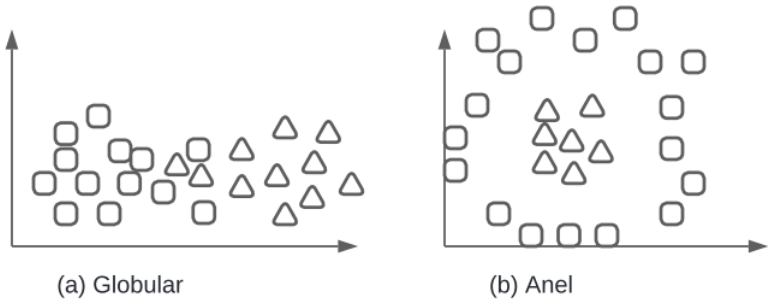
Os critérios de agrupamento são comumente divididos em:

- Compactação ou Homogeneidade: fundamenta-se em manter variação pequena dentro do agrupamento tendo efetividade em encontrar agrupamentos esféricos e/ou bem separados, não contemplando estruturas mais complexas. Como exemplo de algoritmo desse tipo de critério tem-se o *k-means*.

- Encadeamento/Ligação: fundamenta-se no fato de que vizinhos devem compartilhar o mesmo grupo, mas não contempla estruturas em que há pouca separação espacial entre os grupos. Como exemplo de algoritmo tem-se o agrupamento hierárquico.

- Separação espacial: fundamenta-se nas distâncias entre os grupos, mas proporciona pouca informação e pode resultar em soluções simples. Por isso, normalmente é utilizado em associação com outros critérios. Como exemplo tem-se o algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*).

Figura 2.2: Dados agrupados de acordo com diferentes critérios de agrupamento.



Fonte: Adaptado de Faceli (2011)

A Figura 2.2 apresenta dois conjuntos de dados em que no conjunto (a) tem-se dois grupos esféricos, compactos, mas não completamente separados e onde espera-se que algoritmos de compactação se comportem melhor e o conjunto (b) com grupos não compactos, em formato de anel e completamente separados e onde espera-se que algoritmos de encadeamento se comportem melhor.

O conceito de proximidade refere-se a generalização tanto da dissimilaridade quanto da similaridade e cada algoritmo de agrupamento é baseado em um critério de agrupamento o qual usa uma medida de proximidade e um método de busca para encontrar uma estrutura ótima ou subótima que descreva os dados, de acordo com o critério de agrupamento adotado (JIANG; TANG; ZHANG, 2004).

Existem diferentes níveis de proximidade que podem ser considerados em agrupamentos: a proximidade entre as unidades de análise, a proximidade entre a unidade de análise e um grupo de unidades de análise e a proximidade entre dois grupos de unidades de análises, que podem ser quantificadas por medidas de dissimilaridade ou similaridade. Todos os algoritmos de agrupamento consideram a similaridade/dissimilaridade entre objetos, e um mesmo algoritmo pode ser implementado considerando medidas diferentes (FACELI, 2011).

De acordo com Fávero et al. (2009), a análise de agrupamentos é composta pelas seguintes etapas:

- Análise das variáveis e das observações a serem agrupadas (estrutura dos dados);
- Seleção da medida de proximidade, entre pares de observações;
- Seleção do algoritmo de agrupamento;
- Escolha da quantidade de agrupamentos a serem formados; e
- Interpretação e validação dos agrupamentos.

2.2 Estrutura dos dados

Considere um conjunto de dados que pode ser organizado em uma matriz $\mathbf{X}$ com dimensão $n \times p$, em que n e p correspondem as quantidades de observações e de variáveis, respectivamente. Essa matriz pode ser representada da seguinte forma

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

em que x_{ij} é o valor da i-ésima unidade de análise na j-ésima variável $\forall\ i = 1, \ldots, n$ e $j = 1, \ldots, p$. Ainda, para a i-ésima unidade de análise tem-se um vetor multidimensional que pode ser expresso como

$$\mathbf{x}_{i.} = (x_{i1}\ x_{i2}\ \ldots\ x_{ip})$$

em que $\mathbf{x}_{i.}$ é o vetor contendo as p variáveis.

A proximidade entre o i-ésimo e o m-ésimo elemento pode ser armazenada em uma matriz simétrica de dimensão $n \times n$ denominada por Matriz de Proximidade ($S_{n\times n}$) ou Matriz de Similaridade/Dissimilaridade ($D_{n\times n}$). Tal matriz apresenta a similaridade/dissimilaridade entre pares de observações, podendo ser obtida pela similaridade, $s(\mathbf{x_i}., \mathbf{x_m}.) = s_{im}$, ou pela distância, $d(\mathbf{x_i}., \mathbf{x_m}.) = d_{im}$, conforme ilustrado abaixo

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1j} & \cdots & d_{1p} \\ d_{21} & 0 & \cdots & d_{2j} & \cdots & d_{2p} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & \cdots & 0 & \cdots & d_{ip} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nj} & \cdots & 0 \end{bmatrix}$$

Existem diversas medidas de proximidade e, consequentemente, cada uma pode produzir um agrupamento diferente dos dados. As medidas para variáveis quantitativas são de dissimilaridade e são as mais conhecidas, como a distância Euclidiana. Para o caso de variáveis qualitativas Mingoti (2007) diz que no geral tem-se duas opções: (I) transformar as variáveis em quantitativas a partir da atribuição de valores arbitrários para cada categoria e, em seguida, utilizar alguma medida de dissimilaridade para dados quanti-

tativos; (II) utilizar coeficientes de similaridade com as variáveis em sua forma natural. Esses coeficientes basicamente fazem comparações entre os vetores das duas observações em investigação, uma vez que espera-se que quanto mais similares duas observações sejam, mais termos em comum elas apresentem.

A análise de agrupamento é sensível a presença de *outliers*. No entanto, cabe esclarecer que é comum que indivíduos atípicos formem grupos isolados, o que, por vezes, é de interesse do próprio pesquisador esta constatação e, portanto, não necessariamente tais observações devem ser eliminadas da amostra (FÁVERO et al., 2009). Ainda, variáveis com medidas ou escalas diferentes podem distorcer o agrupamento, pois variáveis com maior dispersão acabam dominando as demais, sendo uma forma de contornar isso através do escalonamento das variáveis, afim de que as variáveis tenham escalas semelhantes. Amorim, Cavalcanti e Cruz (2023) fazem um estudo sobre os impactos da escolha da forma de escalonamento no desempenho de algoritmos de classificação. Eles consideram padronização e normalização como métodos de escalonamento de variável, dependendo de um termo translacional e/ou fator de escala para tornar os dados mais concentrados ou espalhados.

A forma mais utilizada para escalonamento é a padronização dos dados, que consiste em transformar cada variável em escore padrão (Z *scores*), permitindo que seja eliminado o viés decorrente das diferenças de escalas (FÁVERO et al., 2009). Esse método é menos sensível a valores discrepantes e faz com que a variável tenha média zero e desvio padrão 1 sendo obtida como:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{s_{.j}}, i = 1, ..., n \; ; j = 1, ..., p$$

em que $\bar{x}_{.j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ e $s_{.j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{.j})^2}$.

Em relação a normalização, diversos métodos podem ser encontrados na literatura, sendo os principais:

- *Range* [-1, 1]: a variável transformada fica no intervalo entre -1 e 1 e é obtida por $\frac{x_{ij}}{\max(x_{.j}) - \min(x_{.j})}$;
- Min-Max: a variável transformada tem variação entre 0 e 1 e é obtida por $\frac{x_{ij} - \min(x_{.j})}{\max(x_{.j}) - \min(x_{.j})}$. Para Amorim, Cavalcanti e Cruz (2023) essa transformação facilita um melhor desempenho em algoritmos baseados em distância;
- Máxima amplitude: a variável transformada tem valor máximo de 1 e é obtida por $\frac{x_{ij}}{\max(x_{.j})}$.

É ressaltado por Mingoti (2007) que qualquer medida de distância usada para

variáveis quantitativas pode ser transformada em um coeficiente de similaridade, podendo ser feito da seguinte forma:

$$s_{im} = 1 - d^*_{im},$$

em que s_{im} é o coeficiente de similaridade, $d^*_{im} = \frac{d(x_i,x_m)-\min(D)}{\max(D)-\min(D)}$ e $\min(D)$ e $\max(D)$ representam o menor e o maior valor de distância na matriz de distância D, respectivamente, sem considerar a diagonal principal.

Nas áreas aplicadas é extremamente comum estudos envolvendo tanto variáveis quantitativas quanto qualitativas conjuntamente, e ao longo da literatura e em concordância com Mingoti (2007) nota-se quatro abordagens usuais nesse caso:

1. Transformar as variáveis qualitativas em quantitativas atribuindo valor numérico a cada categoria e utilizar alguma medida de dissimilaridade para dados quantitativos;

2. Categorizar as variáveis quantitativas e utilizar algum coeficiente de similaridade. Entretanto essa abordagem é menos comum uma vez que há perda de informação das variáveis quantitativas;

3. Agrupar cada tipo de variável de forma distinta e combinar os resultados;

4. Construir medidas de semelhança mistas ou utilizar medidas mistas, sendo o coeficiente de Gower (1971) comumente utilizado nessa abordagem.

2.3 Medidas de dissimilaridade e similaridade

As medidas de dissimilaridade ou similaridade satisfazem algumas propriedades. De acordo com Deza e Deza (2009), as medidas de distância (ou dissimilaridade) devem satisfazer as propriedades 1, 2 e 3 $\forall\ \mathbf{x}_{i.}, \mathbf{x}_{m.}, \mathbf{x}_{n.} \in \mathbf{X}$. Se, além dessas, satisfizerem as propriedades 4 e 5, tem-se uma métrica.

1. Não negatividade: $d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) \geq 0$;

2. Simetria: $d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = d(\mathbf{x}_{m.}, \mathbf{x}_{i.})$;

3. Reflexividade: $d(\mathbf{x}_{i.}, \mathbf{x}_{i.}) = 0$;

4. $d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_m$;

5. Desigualdade triangular: $d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) \leq d(\mathbf{x}_{i.}, \mathbf{x}_{n.}) + d(\mathbf{x}_{n.}, \mathbf{x}_{m.})$.

Por outro lado, para as medidas de similaridade tem-se as seguintes propriedades $\forall\ \mathbf{x}_{i.}, \mathbf{x}_{m.}, \mathbf{x}_{n.} \in \mathbf{X}$:

1. Não negatividade: $s(\mathbf{x}_{i.}, \mathbf{x}_{m.}) \geq 0$;

2. Simetria: $s(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = s(\mathbf{x}_{m.}, \mathbf{x}_{i.})$;

3. $s(\mathbf{x}_{i.}, \mathbf{x}_{m.}) \leq s(\mathbf{x}_{i.}, \mathbf{x}_{i.})$, com igualdade apenas quando $\mathbf{x}_i = \mathbf{x}_{m.}$.

A escolha da medida mais adequada está relacionada à área de aplicação e os tipos das variáveis envolvidas no estudo.

2.3.1 Medidas para dados quantitativos

As medidas de distância são consideradas medidas de dissimilaridade, pois quanto maiores os valores, mais dissimilares são as observações, ou seja, mais afastadas as observações estão. Abaixo são apresentadas as principais medidas de distância para dados quantitativos.

Distância Euclidiana

A distância Euclidiana, também chamada de *Norma* L_2, entre duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ é definida como:

$$d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = [(\mathbf{x}_i - \mathbf{x}_m)'(\mathbf{x}_i - \mathbf{x}_m)]^{1/2} = \sqrt{\sum_{j=1}^{p} (\mathbf{x}_{ij} - \mathbf{x}_{mj})^2}.$$

Essa medida de distância é a mais frequentemente utilizada, sendo uma generalização do teorema de Pitágoras. Além disso, possui variações como a distância euclidiana média, a padronizada e a padronizada média. Crispim, Fernandes e Albuquerque (2019) utilizam a distância euclidiana visando agrupar indicadores sociais, econômicos, habitacionais e de saneamento ambiental e Filho et al. (2008) comparam métodos de agrupamento com dissimilaridade euclidiana média padronizada e generalizada de mahalanobis considerando cultivares de feijão.

Distância de Manhattan

A distância de Manhattan, também chamada de *Norma* L_1 ou *City-Block* ou distância de táxi ou absoluta, entre duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ é dada por:

$$d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \sum_{j=1}^{p} |\mathbf{x}_{ij} - \mathbf{x}_{mj}|.$$

Logo, é dada pela soma das diferenças absolutas entre valores de cada variável. Barbosa (2022) propõe o uso das distâncias de Manhattan e Chebyshev na avaliação da acurácia

posicional de produtos cartográficos. Apesar de ter observado comportamentos diferentes com o uso de tais distância, estatisticamente não foi encontrada diferença em relação ao que é padrão na área, ou seja, ao uso da distância Euclidiana. Nascimento et al. (2019) estudam casos de Leishmaniose Visceral no Cariri do estado do Ceará onde utilizando agrupamento e a distância de Manhattan dividem a cidade em três grupos e verificam a existência de locais com níveis semelhantes da doença.

Distância de Chebyshev

A distância de Chebyshev, também chamada de *supremum* ou *dominance metric*, entre duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ é definida como:

$$d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \max_{1 \leq j \leq p} |\mathbf{x}_{ij} - \mathbf{x}_{mj}|.$$

Essa distância representa a diferença absoluta máxima entre as p variáveis. Gomes et al. (2019) fazem análise de agrupamento com as distâncias Euclidiana, Chebyshev e Mahalanobis afim de identificar zonas com velocidades dos ventos distintas no estado do Ceará, encontrando 5 zonas. Gomes et al. (2020) fizeram a caracterização de genótipos de mandioca através da formação de dois grupos de genótipos por meio da análise de agrupamento com a distância de Chebyshev.

Distância de Mahalanobis

A distância de Mahalanobis entre duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ é definida como:

$$D^2 = (\mathbf{x}_{i.} - \mathbf{x}_{m.})' \Sigma^{-1} (\mathbf{x}_{i.} - \mathbf{x}_{m.}),$$

Sendo Σ^{-1} a inversa da matriz de variância-covariância (Σ) considerando a variabilidade. De acordo com Xu e Wunsch (2008) e Fávero et al. (2009), o cálculo da inversa de Σ pode ter alta carga computacional em dados de grande escala e a utilização da distância de Mahalanobis colabora para atenuar efeito de multicolinearidade. Benin et al. (2003) comparam medidas de dissimilaridade (Euclidiana e Mahalanobis) no contexto agrícola e Pereira (2009) utiliza Mahalanobis em um estudo comparativo das funções disciminantes de Fisher e redes neurais artificais.

Métrica de Minkowski

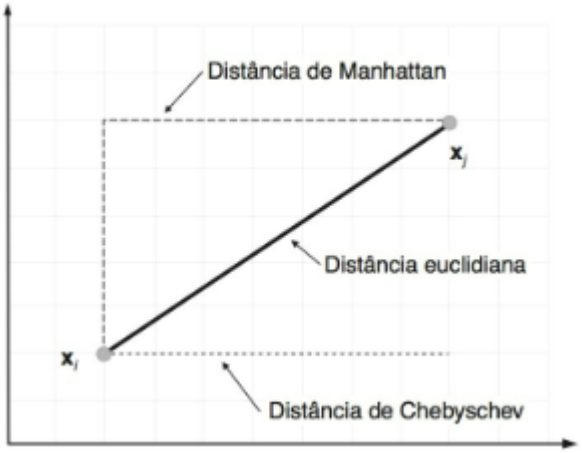
A métrica de Minkowski, ou distância L_p, entre duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ é definida como:

$$d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \sqrt[a]{\sum_{j=1}^{p} |x_{ij} - x_{mj}|^a},$$

em que $1 \leq a < \infty$, ou seja, de acordo com diferentes valores de a obtem-se variações da métrica. As medidas mais utilizadas para dados contínuos tem como base a métrica de Minkowski, como a distância Euclidiana, a distância de Manhattan e a distância *supremum*. As variações mais frequentes da métrica de Minkowski são apresentadas a seguir e suas representações no espaço R^2 são apresentadas na Figura 2.3.

- $a = 1$: Distância de Manhattan;
- $a = 2$: Distância Euclidiana;
- $a \to \infty$: Distância de Chebyshev.

Figura 2.3: Interpretação das medidas com base na Métrica de Minkowski das observações $\mathbf{x}_i$ e $\mathbf{x}_j$.



Fonte: Faceli (2011).

A métrica de Minkowsky é menos afetada por valores discrepantes (*outliers*) para menores valores de a (MINGOTI, 2007; FACELI, 2011).

Distância Generalizada

A distância generalizada, também chamada de ponderada, entre duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ é definida por:

$$d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = [(\mathbf{x}_{.i} - \mathbf{x}_{m.})' A (\mathbf{x}_{i.} - \mathbf{x}_{m.})]^{1/2}$$

em que A é uma matriz de ponderação (positiva definida) de dimensão $p \times p$. Ainda, se A for:

- Matriz identidade: a distância generalizada se equivale a distância Euclidiana;
- Σ^{-1}: a distância generalizada se equivale a distância de Mahalanobis (1936);
- $diag\left(\frac{1}{p}\right)$: a distância generalizada se equivale a distância euclidiana média.

A escolha da matriz $A_{p\times p}$ reflete o tipo de informação que o pesquisador deseja utilizar na ponderação das diferenças das coordenadas dos vetores que estão sendo comparados.

2.3.2 Medidas para dados qualitativos

Os dados com escala de mensuração qualitativa são classificados em nominais e ordinais. Ainda, pode-se dividir em dicotômicos (ou binário), em que se tem duas categorias, e politômicos, em que se tem três ou mais categorias.

2.3.2.1 Medidas para dados dicotômicos

Dados dicotômicos, ou binários, referem-se a presença ou ausência de uma determinada característica. Frequentemente são representados pela variável indicadora (*dummy)* assumindo valor 1 para a presença e 0 para a ausência. Para dados binários é comum a utilização de medidas de similaridade. Essas podem ser divididas em invariantes (ou simétricas) e não invariantes (ou assimétricas). As similaridades invariantes consideram os pares de respostas simultâneas $(0,0)$ e $(1,1)$ com igual importância, enquanto que as não invariantes ignoram os pares $(0,0)$ pelo fato de não ser informativo, pois pode ser devido ao tipo de categorização utilizada. Considerando os vetores multidimensionais de duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ pode-se construir uma tabela de contingência conforme a Tabela 2.1

Tabela 2.1: Tabela de contingência para dados binários.

		i-ésimo indivíduo		
		1	0	Total
m-ésimo indivíduo	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$p=a+b+c+d$

em que a quantifica presença em ambos os indivíduos, b a presença no m-ésimo indivíduo e ausência no i-ésimo, c a ausência no m-ésimo e presença no i-ésimo e d a ausência em ambos.

Para as similaridades invariantes podemos considerar a seguinte forma:

$$s(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \frac{a+d}{a+d+w(b+c)},$$

em que para $w = 1$ tem-se o coeficiente de correspondência simples, $w = 2$ o coeficiente de Rogers e Tanimoto (1960) e $w = 1/2$ o coeficiente de Sokal e Sneath (1963). Para o coeficiente de correspondência simples, a medida de dissimilaridade correspondente de $d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = 1 - s(\mathbf{x}_{i.}, \mathbf{x}_{m.})$ é conhecida como a distância de Hamming (XU; WUNSCH, 2008). Neto e Negreiros (2017) avaliam a distância euclidiana, do cosseno, de hamming e os coeficientes de Jaccard estendido e correlação de Pearson no contexto de agrupamento de objetos textuais.

Já para as similaridades não invariantes temos:

$$s(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \frac{a}{a+w(b+c)},$$

em que para $w = 1$ tem-se o coeficiente de Jaccard (1908), $w = 2$ o coeficiente de Sokal e Sneath (1963) e $w = 1/2$ o coeficiente Dice (1945). Emygdio et al. (2003) utilizam os coeficientes de Jaccard, Dice, Rogers e Tanimoto para avaliar eficiência em genótipos de feijão e Rezende, Marcacini e Moura (2011) descrevem técnicas e algoritmos não supervisionados para dados textuais com as similaridades do cosseno e Jaccard. A Tabela 2.2 sumariza as principais similaridades para dados binários.

Tabela 2.2: Principais medidas de similaridade para dados binários.

Coeficiente	**Tipo**	**Fórmula**
Correspondência simples	Invariante	$\frac{a+d}{a+d+(b+c)}$
Rogers e Tanimoto (1960)	Invariante	$\frac{a+d}{a+d+2(b+c)}$
Sokal e Sneath (1963)	Invariante	$\frac{a+d}{a+d+(1/2)(b+c)}$
Jaccard (1908)	Não invariante	$\frac{a}{a+(b+c)}$
Sokal e Sneath (1963)	Não invariante	$\frac{a}{a+2(b+c)}$
Dice (1945)	Não invariante	$\frac{a}{a+(1/2)(b+c)}$

Meyer (2002) e Albuquerque et al. (2016) estudaram diversas medidas de similaridade observando que as medidas invariantes tiveram comportamentos similares entre si e as não invariantes também. Diversas medidas são encontradas na literatura, mas o uso específico de cada relaciona-se com a área do conhecimento em estudo.

Distância de Hamming

Uma medida muito utilizada para dados qualitativos é a distância de Hamming que contabiliza o número de atributos categóricos com valores diferentes no par em inves-

tigação. Esta medida pode variar entre $[0, p]$ em que 0 indica a maior similaridade entre duas observações e é definida como:

$$d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \sum_{j=1}^{p} a(\mathbf{x}_{i.}, \mathbf{x}_{m.}),$$

em que:

$$a(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \begin{cases} 1, & \text{se } \mathbf{x}_{ij} \neq \mathbf{x}_{mj} \\ 0, & \text{se } c.c. \end{cases}$$

2.3.2.2 Medidas para dados politômicos

Medidas para dados nominais

Considerando dados com escala de mensuração nominal, uma possibilidade é a transformação de cada categoria em uma variável binária (*one-hot-encoding*), no entanto essa abordagem pode resultar em um maior número de variáveis. Alternativamente, pode-se empregar a seguinte fórmula para o cálculo da dissimilaridade entre duas observações

$$d(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \frac{V - v}{V},$$

em que V representa o número total de variáveis nominais e v o número de variáveis em que $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ são iguais.

Medidas para dados ordinais

Para dados qualitativos ordinais, primeiramente deve-se fazer um *ranking* com base na ordem natural das categorias da j-ésima variável e, seguidamente, calcular b_{ij} que é dado por:

$$b_{ij} = \frac{l_{ij} - 1}{C_j - 1},$$

em que l_{ij} é o valor no *ranking* e C_j o número de categorias possíveis da j-ésima variável. Em seguida, para o cálculo da distância, basta utilizar alguma medida para dados quantitativos.

2.3.3 Medidas para dados mistos

Medida de Gower

Para este cenário, uma medida frequentemente utilizada é o coeficiente geral de similaridade que, de acordo com Xu e Wunsch (2008), consiste de um método poderoso que foi proposto por Gower em 1971 e posteriormente estendido por Kaufman e Rousseeuw em 1990, e que define a similaridade entre duas observações $\mathbf{x}_{i.}$ e $\mathbf{x}_{m.}$ como:

$$s(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = \frac{\sum_{j=1}^{p} \delta_{imj} T_{imj}}{\sum_{j=1}^{p} \delta_{imj}}$$

em que T_{imj} refere-se a contribuição da j-ésima variável no intervalo $[0, 1]$ e δ_{imj} assume valores 0 ou 1, pois depende da comparação dos elementos da j-ésima variável. Para variáveis quantitativas tem-se que $T_{imj} = 1 - \frac{|\mathbf{x}_{ij} - \mathbf{x}_{mj}|}{\max(\mathbf{x}_{.j}) - \min(\mathbf{x}_{.j})}$ e qualitativas tem-se $T_{imj} = 1$ se $\mathbf{x}_{i.} = \mathbf{x}_{m.}$ e 0 caso contrário.

Medida mista

O coeficiente da medida mista entre a i-ésima e a m-ésima observação é dado por:

$$c(\mathbf{x}_{i.}, \mathbf{x}_{m.}) = w_f c_f(i, m) + w_q c_q(i, m)$$

em que f refere-se as variáveis quantitativas, q as variáveis qualitativas, c_f, c_q são as respectivas medidas (dis)similaridade entre i e m, w_p, w_q são pesos de ponderação. Segundo Mingoti (2007), é necessário que os coeficientes c_f, c_q tenham mesma direção e mesmo intervalo de variação. A dificuldade dessa abordagem é a definição dos pesos, sendo uma alternativa atribuí-los como função do número de variáveis quantitativas e qualitativas como, por exemplo, $w_f = \frac{f}{f+q}$ e $w_q = \frac{q}{f+q}$.

2.3.4 Exemplos de aplicação

Considere os dados apresentados na Tabela 2.3, em que idade (Ida.) e altura (Alt.) são variáveis quantitativas, Casado e Operado variáveis binárias, tipo de emprego (Empr.) e tipo sanguíneo (T.S.) qualitativas nominais e categoria de altura (C. alt.) e grau de escolaridade (Esc.) qualitativas ordinais. Esses dados são fictícios e foram gerados a fim de exemplificar o cálculo das medidas de dissimilaridade e similaridade apresentadas anteriormente. Para simplificação, são usadas as seguintes correspondências: $d(\mathbf{x}_A, \mathbf{x}_B) = d_{AB}$ e $s(\mathbf{x}_A, \mathbf{x}_B) = s_{AB}$.

Tabela 2.3: Base sintética para exemplificação.

Ind.	**Ida.**	**Alt.**	**C. alt.**	**Casado**	**Operado**	**T. S.**	**Empr.**	**Esc.**
A	31	1,73	Médio	0	1	B	Público	Superior
B	59	1,64	Médio	1	0	O	Privado	Fundamental
C	38	1,56	Baixo	1	1	B	Privado	Médio
D	56	1,98	Alto	1	0	A	Autônomo	Médio
E	40	1,89	Alto	1	0	B	Autônomo	Fundamental

Dados quantitativos

Utilizando as variáveis quantitativas **Idade** (discreta) e **Altura** (contínua), tem-se:

1. Distância Euclidiana

   A distância Euclidiana entre os elementos A e B é dada por:

   $$d_{AB} = \sqrt{(31 - 59)^2 + (1,73 - 1,64)^2} \simeq 28,00.$$

   Analogamente, para os elementos A e C obtem-se:

   $$d_{AC} = \sqrt{(31 - 38)^2 + (1,73 - 1,56)^2} \simeq 7,00.$$

   Dessa maneira, a matriz de dissimilaridade é dada por:

   Tabela 2.4: Matriz de dissimilaridade (Euclidiana).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	28,00	7,00	25,00	9,00
B	28,00	0	21,00	3,02	19,00
C	7,00	21,00	0	18,00	2,03
D	25,00	3,02	18,00	0	16,00
E	9,00	19,00	2,03	16,00	0

   Pela distância Euclidiana as observações C e E são mais similares considerando as variáveis correspondentes.

2. Distância de Manhattan

   A distância de Manhattan entre os elementos A e B pode ser obtida como:

   $$d_{AB} = |31 - 59| + |1,73 - 1,64| = 28,09.$$

   De forma similar, para os elementos A e C tem-se:

   $$d_{AC} = |31 - 38| + |1,73 - 1,56| = 7,17.$$

Desse modo, a matriz de dissimilaridade é dada por:

Tabela 2.5: Matriz de dissimilaridade (Manhattan).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	28,09	7,17	25,25	9,16
B	28,09	0	21,08	3,34	19,25
C	7,17	21,08	0	18,42	2,33
D	25,25	3,34	18,42	0	16,09
E	9,16	19,25	2,33	16,09	0

Pela distância de Manhattan as observações C e E são mais similares considerando as variáveis correspondentes.

3. Distância de Chebyshev

   A distância de Chebyshev entre os elementos A e B é dada por:

$$d_{AB} = max(|31 - 59|, |1,73 - 1,64|) = 28,00.$$

   Similarmente, para os elementos A e C tem-se:

$$d_{AC} = max(|31 - 38|, |1,73 - 1,56|) = 7,00.$$

   Dessa maneira, a matriz de dissimilaridade é dada por:

Tabela 2.6: Matriz de dissimilaridade (Chebyshev).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	28,00	7,00	25,00	9,00
B	28,00	0	21,00	3,00	19,00
C	7,00	21,00	0	18,00	2,00
D	25,00	3,00	18,00	0	16,00
E	9,00	19,00	2,00	16,00	0

Pela distância Chebyshev as observações C e E são mais similares considerando as variáveis correspondentes.

4. Distância de Minkowski

   A distância de Minkowski entre os elementos A e B é dada por:

$$d_{AB} = \sqrt[2]{|31 - 59|^2 + |1,73 - 1,64|^2} \simeq 28,00.$$

   Analogamente, para os elementos A e C obtem-se:

$$d_{AC} = \sqrt[2]{|31-38|^2 + |1,73-1,56|^2} \simeq 7,00.$$

Dessa maneira, a matriz de dissimilaridade é dada por:

Tabela 2.7: Matriz de dissimilaridade (Minkowski).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	28,00	7,00	25,00	9,00
B	28,00	0	21,00	3,02	19,00
C	7,00	21,00	0	18,00	2,03
D	25,00	3,02	18,00	0	16,00
E	9,00	19,00	2,03	16,00	0

Pela distância Minkowski as observações C e E são mais similares considerando as variáveis correspondentes.

5. Distância de Mahalanobis

A distância de Mahalanobis entre os elementos A e B é obtida como:

$$d^2_{AB} = \begin{bmatrix} 31-59 & 1,73-1,64 \end{bmatrix}^\intercal S^{-1} \begin{bmatrix} 31-59 \\ 1,73-1,64 \end{bmatrix} \simeq 6,49.$$

Similarmente, para os elementos A e C tem-se:

$$d^2_{AC} = \begin{bmatrix} 31-38 & 1,73-1,56 \end{bmatrix}^\intercal S^{-1} \begin{bmatrix} 31-38 \\ 1,73-1,56 \end{bmatrix} \simeq 1,77,$$

em que $S^{-1} = \begin{bmatrix} 0,007 & -0,114 \\ -0,114 & 34,970 \end{bmatrix}$.

Desse modo, a matriz de dissimilaridade é dada por:

Tabela 2.8: Matriz de dissimilaridade (Mahalanobis).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	6,49	1,63	5,25	1,15
B	6,49	0	3,01	4,34	5,86
C	1,63	3,01	0	6,78	3,69
D	5,25	4,34	6,78	0	1,80
E	1,15	5,86	3,69	1,80	0

Pela distância Mahalanobis as observações A e E são mais similares considerando as variáveis correspondentes.

6. Distância Generalizada

A distância Generalizada entre os elementos A e B é obtida como:

$$d_{AB} = \sqrt{\begin{bmatrix} 31-59 & 1,73-1,64 \end{bmatrix}^{\intercal} A \begin{bmatrix} 31-59 \\ 1,73-1,64 \end{bmatrix}} \simeq 28,00.$$

Similarmente, para os elementos A e C tem-se:

$$d_{AC} = \sqrt{\begin{bmatrix} 31-38 & 1,73-1,56 \end{bmatrix}^{\intercal} A \begin{bmatrix} 31-38 \\ 1,73-1,56 \end{bmatrix}} \simeq 7,00,$$

em que $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Desse modo, a matriz de dissimilaridade é dada por:

Tabela 2.9: Matriz de dissimilaridade (Generalizada).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	28,00	7,00	25,00	9,00
B	28,00	0	21,00	3,02	19,00
C	7,00	21,00	0	18,00	2,03
D	25,00	3,02	18,00	0	16,00
E	9,00	19,00	2,03	16,00	0

Pela distância Generalizada as observações C e E são mais similares considerando as variáveis correspondentes.

Dados dicotômicos

Utilizando as variáveis qualitativas binárias **Operado** (invariante/simétrico) e **Casado** (não invariante/assimétrico) obtem-se as Tabelas de contingência 2.10 e 2.11 referentes aos elementos A e B, e A e C, assim:

Tabela 2.10: Tabela de contingência referente aos elementos A e B.

		Indivíduo A	
		1	**0**
Indivíduo B	**1**	0	1
	0	1	0

Tabela 2.11: Tabela de contingência referente aos elementos A e C.

		Indivíduo A	
		1	**0**
Indivíduo C	**1**	1	0
	0	1	0

1. Coeficiente de Correspondência Simples

   O coeficiente de correspondência simples entre os elementos A e B é dado por:

$$s_{AB} = \frac{0+0}{0+0+1+1} = 0.$$

   Analogamente, para os elementos A e C obtem-se:

$$s_{AC} = \frac{1+0}{1+0+0+1} = 0,5.$$

   Dessa maneira, a matriz de dissimilaridade utilizando a correspondência $d = 1 - s$ é dada por:

Tabela 2.12: Matriz de dissimilaridade.

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,50	1,00	1,00
B	1,00	0	0,50	0,00	0,00
C	0,50	0,50	0	0,50	0,50
D	1,00	0,00	0,50	0	0,00
E	1,00	0,00	0,50	0,00	0

   Pelo coeficiente de correspondência simples as observações B e D, B e E, D e E são mais similares considerando as variáveis correspondentes.

2. Coeficiente de Rogers e Tanimoto (1960)

   O coeficiente de Rogers e Tanimoto (1960) entre os elementos A e B pode ser obtido como:

$$s_{AB} = \frac{0+0}{0+0+2(1+1)} = 0.$$

   De forma similar, para os elementos A e C tem-se:

$$s_{AC} = \frac{1+0}{1+0+2(0+1)} \simeq 0,33.$$

Desse modo, a matriz de dissimilaridade utilizando a correspondência $d = 1 - s$ é dada por:

Tabela 2.13: Matriz de dissimilaridade (Rogers e Tanimoto).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,67	1,00	1,00
B	1,00	0	0,67	0,00	0,00
C	0,67	0,67	0	0,67	0,67
D	1,00	0,00	0,67	0	0,00
E	1,00	0,00	0,67	0,00	0

Pelo coeficiente de Rogers e Tanimoto as observações B e D, B e E, D e E são mais similares considerando as variáveis correspondentes.

3. Coeficiente de Sokal e Sneath (1963)

   O coeficiente de Sokal e Sneath entre os elementos A e B pode ser obtido como:

$$s_{AB} = \frac{0+0}{0+0+(1/2)(1+1)} = 0.$$

   De forma similar, para os elementos A e C tem-se:

$$s_{AC} = \frac{1+0}{1+0+(1/2)(0+1)} \simeq 0,67.$$

   Desse modo, a matriz de dissimilaridade utilizando a correspondência $d = 1 - s$ é dada por:

Tabela 2.14: Matriz de dissimilaridade (Sokal e Sneath).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,33	1,00	1,00
B	1,00	0	0,33	0,00	0,00
C	0,33	0,33	0	0,33	0,33
D	1,00	0,00	0,33	0	0,00
E	1,00	0,00	0,33	0,00	0

   Pelo coeficiente de Sokal e Sneath as observações B e D, B e E, D e E são mais similares considerando as variáveis correspondentes.

4. Coeficiente de Jaccard

   O coeficiente de Jaccard entre os elementos A e B é obtido como:

$$s_{AB} = \frac{0}{0+1(1+1)} = 0.$$

Similarmente, para os elementos A e C tem-se:

$$s_{AC} = \frac{1}{1+1(0+1)} = 0,5.$$

Desse modo, a matriz de dissimilaridade utilizando a correspondência $d = 1 - s$ é dada por:

Tabela 2.15: Matriz de dissimilaridade (Jaccard).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,50	1,00	1,00
B	1,00	0	0,50	0,00	0,00
C	0,50	0,50	0	0,50	0,50
D	1,00	0,00	0,50	0	0,00
E	1,00	0,00	0,50	0,00	0

Pelo coeficiente de Jaccard as observações B e D, B e E, D e E são mais similares considerando as variáveis correspondentes.

A observação A é mais similar a observação C considerando as duas variáveis.

5. Coeficiente de Sokal e Sneath

O coeficiente de Sokal e Sneath entre os elementos A e B é obtido como:

$$s_{AB} = \frac{0}{0+2(1+1)} = 0.$$

Similarmente, para os elementos A e C tem-se:

$$s_{AC} = \frac{1}{1+2(0+1)} \simeq 0,33.$$

Desse modo, a matriz de dissimilaridade utilizando a correspondência $d = 1 - s$ é dada por:

Tabela 2.16: Matriz de dissimilaridade (Sokal e Sneath).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,67	1,00	1,00
B	1,00	0	0,67	0,00	0,00
C	0,67	0,67	0	0,67	0,67
D	1,00	0,00	0,67	0	0,00
E	1,00	0,00	0,67	0,00	0

Pelo coeficiente de Sokal e Sneath as observações B e D, B e E, D e E são mais similares considerando as variáveis correspondentes.

6. Coeficiente Dice (1945)

   O coeficiente Dice entre os elementos A e B é obtido como:

$$s_{AB} = \frac{0}{0 + (1/2)(1+1)} = 0.$$

   Similarmente, para os elementos A e C tem-se:

$$s_{AC} = \frac{1}{1 + (1/2)(0+1)} \simeq 0,67.$$

   Desse modo, a matriz de dissimilaridade utilizando a correspondência $d = 1 - s$ é dada por:

Tabela 2.17: Matriz de dissimilaridade (Dice).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,33	1,00	1,00
B	1,00	0	0,33	0,00	0,00
C	0,33	0,33	0	0,33	0,33
D	1,00	0,00	0,33	0	0,00
E	1,00	0,00	0,33	0,00	0

   Pelo coeficiente Dice as observações B e D, B e E, D e E são mais similares considerando as variáveis correspondentes.

7. Distância de Hamming

   A distância de Hamming entre os elementos A e B é obtido como:

$$d_{AB} = 1 + 1 = 2.$$

   Similarmente, para os elementos A e C tem-se:

$$d_{AC} = 1 + 0 = 1.$$

   Desse modo, a matriz de distância é dada por:

Tabela 2.18: Matriz de distância de Hamming.

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	2,00	1,00	2,00	2,00
B	2,00	0	1,00	0,00	0,00
C	1,00	1,00	0	1,00	1,00
D	2,00	0,00	1,00	0	0,00
E	2,00	0,00	1,00	0,00	0

Pelo a distância de Hamming as observações B e D, B e E, D e E são mais similares considerando as variáveis correspondentes.

Dados politômicos

Dados nominais

Para o caso das variáveis nominais **Tipo Sanguíneo** e **Emprego** tem-se que a distância entre os elementos A e B é dada por:

$$d_{AB} = \frac{2-0}{2} = 1.$$

De forma análoga, para os elementos A e C obtem-se:

$$d_{AC} = \frac{2-1}{2} = 0,5.$$

Dessa maneira, a matriz de distância é dada por:

Tabela 2.19: Matriz de distância nominal.

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,50	1,00	0,50
B	1,00	0	0,50	1,00	1,00
C	0,50	0,50	0	1,00	0,50
D	1,00	1,00	1,00	0	0,50
E	0,50	1,00	0,50	0,50	0

Pelo a distância nominal as observações A e C, A e E, B e C, C e E, D e E são mais similares considerando as variáveis correspondentes.

Dados ordinais

Considerando as variáveis ordinais **C. alt.** (categoria de altura) e **Escolaridade** (grau de escolaridade) inicialmente se estabelece o ranking das categorias de cada variável e calcula-se o valor b_{ij} equivalente, como pode ser visto na Tabela 2.20.

Tabela 2.20: Esquematização do cálculo para dados ordinais.

	Ranking		b_{ij}	
Indivíduo	Cl. alt.	Escolaridade	Cl. alt.	Escolaridade
A	2	3	0,50	1,00
B	2	1	0,50	0,00
C	1	2	0,00	0,50
D	3	2	1,00	0,50
E	3	1	1,00	0,00

Agora em posse de variáveis apenas quantitativas, basta utilizar alguma medida de distância para dados quantitativos vistas anteriormente. A Tabela 2.21 apresenta a matriz de dissimilaridade considerando a distância euclidiana para b_{ij}.

Tabela 2.21: Matriz de dissimilaridade (ordinal).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	1,00	0,71	0,71	1,12
B	1,00	0	0,71	0,71	0,50
C	0,71	0,71	0	1,00	1,12
D	0,71	0,71	1,00	0	0,50
E	1,12	0,50	1,12	0,50	0

Pela distância ordinal as observações B e E, D e E são mais similares considerando as variáveis correspondentes.

A observação A é mais similar a observação C e D considerando as duas variáveis.

Dados mistos

A distância de Gower entre os elementos A e B é dada por:

$$d_{AB} = \frac{1 \times 1 + 1 \times 0,21 + 1 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 0 + 1 \times 1}{8} \simeq 0,78,$$

pois:

Idade: $d_{ijp} = \frac{|31-59|}{59-31} = 1$

Altura: $d_{ijp} = \frac{|1,73-1,64|}{1,98-1,56} = 0,21$

Casado: $x_A \neq x_B => d_{ijp} = 1$

Operado: $x_A \neq x_B => d_{ijp} = 1$

Tipo Sanguíneo: $x_A \neq x_B => d_{ijp} = 1$

Emprego: $x_A \neq x_B => d_{ijp} = 1$

Altura 2: $d_{ijp} = \frac{|0,5-0,5|}{1-0} = 0$

Escolaridade: $d_{ijp} = \frac{|1-0|}{1-0} = 1$

De forma análoga, para os elementos A e C obtem-se:

$$d_{Ac} = \frac{1 \times 0,25 + 1 \times 0,40 + 1 \times 1 + 1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 0,5 + 1 \times 0,5}{8} \simeq 0,46,$$

pois:

Idade: $d_{ijp} = \frac{|31-38|}{59-31} = 25$

Altura: $d_{ijp} = \frac{|1,73-1,56|}{1,98-1,56} \approx 0,40$

Casado: $x_A \neq x_B => d_{ijp} = 1$
Operado: $x_A = x_B => d_{ijp} = 0$
Tipo Sanguíneo: $x_A \neq x_B => d_{ijp} = 1$
Emprego: $x_A \neq x_B => d_{ijp} = 1$
Altura 2: $d_{ijp} = \frac{|0,5-0|}{1-0} = 0.5$
Escolaridade: $d_{ijp} = \frac{|1-0.5|}{1-0} = 0.5$
Dessa maneira, a matriz de dissimilaridade é dada por:

Tabela 2.22: Matriz de dissimilaridade (Gower).

Indivíduo	**A**	**B**	**C**	**D**	**E**
A	0	0,78	0,46	0,81	0,65
B	0,78	0	0,49	0,49	0,47
C	0,46	0,49	0	0,71	0,54
D	0,81	0,49	0,71	0	0,29
E	0,65	0,47	0,54	0,29	0

Pela distância de Gower as observações D e E são mais similares considerando todas as variáveis.

Pela exemplificação fica tangível que o valor da proximidade varia conforme a medida utilizada. Ao analisar a dissimilaridade com as variáveis quantitativas nota-se que a distância Euclidiana, de Manhattan e Generalizada obtiveram o mesmo valor, conforme o esperado, mostrando que as observações C-E seriam as mais próximas. Para o caso binário, a proximidade foi observada entre as observação B-D, B-E e D-E. Ainda, observa-se que no caso ordinal a identificação das observações mais próximas foi melhor do que no nominal, ou seja, as observações na base de dados original tiveram padrões de respostas parecidas. Já considerando todas as variáveis, foi possível identificar as observações que realmente tiveram respostas mais semelhantes na base de dados inicial.

2.4 Algoritmos de agrupamentos

Os diversos métodos de agrupamento apesar de terem o mesmo objetivo comum, se diferenciam pelas diferentes estruturas de dados. De acordo com Faceli (2011), diferentes tipos de estruturas encontradas em algoritmo de agrupamento são, por exemplo, partições, hierarquias, partições *fuzzy*, em que cada algoritmo utiliza uma única estrutura. Algumas categorizações encontradas são: exclusivo *vs.* não exclusivo, hierárquico *vs.* particional, baseado em teoria dos grafos *vs.* álgebra matricial etc.

O agrupamento exclusivo basicamente fragmenta o conjunto de dados em grupos de tal forma que cada observação é relacionada somente a um grupo. Formalmente,

dado o conjunto de dados $\mathbf{X}$, uma partição em K *clusters* pode ser definida como: $\pi = G_1, G_2, ..., G_K$ com $K < n$, tal que (XU; WUNSCH, 2008):

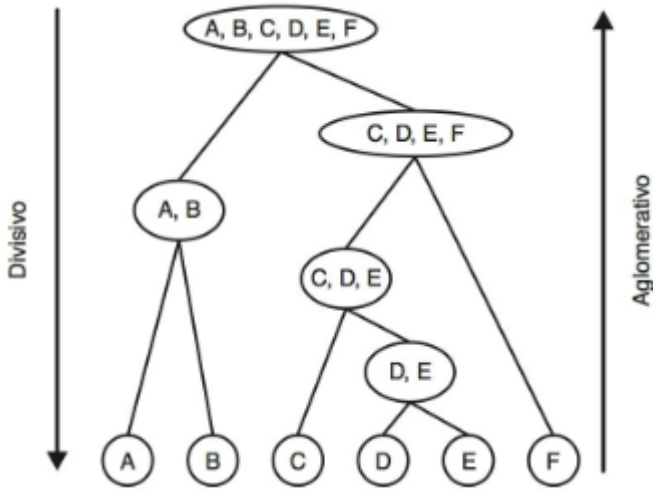
1. $G_k \neq \emptyset, k = 1, 2, ..., K$ (todos os grupos contêm pelo menos uma observação);
2. $\bigcup_{k=1}^{K} G_K = \mathbf{X}$ (todas as observações pertencem a algum grupo);
3. $G_{k1} \bigcap G_{k2} = \emptyset$, $k1, k2 = 1, ..., K$ e $k1 \neq k2$ (cada observação pertence exclusivamente a um único grupo).

Em contrapartida, o agrupamento não exclusivo pode relacionar cada observação a mais de um grupo, como ocorre nos agrupamentos do tipo *fuzzy* ou probabilístico.

A análise de agrupamento comumente é dividida em duas vertentes: a hierárquica e a não hierárquica. A principal diferença consiste no fato de que para o agrupamento não hierárquico é necessário pré definir o número de grupos desejados enquanto que para o hierárquico isso não é feito, sendo realizadas partições de forma aninhada.

O agrupamento hierárquico apresenta dois segmentos: o aglomerativo e o divisivo (Figura 2.4). O agrupamento hierárquico aglomerativo é mais frequente e inicia-se considerando que cada observação é o seu próprio grupo, realizando o agrupamento das observações mais próximas, um por vez, até que todas as observações façam parte de um único grupo. Em cada estágio do procedimento de agrupamento, os grupos são comparados através de alguma medida de similaridade (ou dissimilaridade) previamente definida (MINGOTI, 2007). Já o agrupamento hierárquico divisivo inicia-se considerando que todas as observações compõem um único grupo, se subdividindo em grupos menores até que cada observação seja o seu próprio grupo. No geral é representado graficamente pelo dendograma.

Figura 2.4: Exemplo de funcionamento dos algoritmos hierárquicos aglomerativos e divisivos.



Fonte: Faceli (2011).

As etapas do agrupamento hierárquico aglomerativo são:

1. Cada observação é um grupo de uma observação;
2. A cada passo, o par que possui maior similaridade é unido (somente um novo grupo é formado por vez);
3. Um novo grupo é obtido a partir da união de outros dois formados anteriormente. Dessa forma, ao se unir duas observações elas sempre estarão unidas nas etapas seguintes (hierarquia);
4. É possível a construção do dendograma, que exibe o histórico de agrupamento.

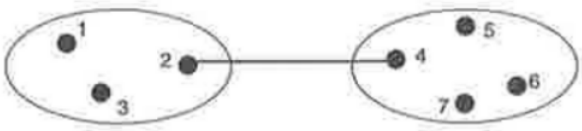
De acordo com Faceli (2011), o agrupamento hierárquico tem como pontos positivos a sua flexibilidade quanto ao nível de refinamento, fácil utilização de qualquer forma de (dis)similaridade e que se pode utilizar qualquer tipo de atributo, já como ponto negativo tem-se o critério de terminação subjetivo e que o algoritmo não melhora os grupos, pois uma vez construídos permanecem juntos até o final. Esse tipo de agrupamento resulta em *clusters* de formas convexas próprias e, em geral, possuem complexidade $O(n^2)$.

Sejam G_{k1} e G_{k2} dois grupos distintos, as principais métricas de integração do agrupamento hierárquico são:

- Método da Ligação Simples ou do Vizinho Mais Próximo (*Simple Linkage*): As observações são agrupadas de acordo com o menor valor de distância, ou seja, as duas observações mais parecidas.

$$d(G_{k1}, G_{k2}) = min\{d_{im} : i \in G_{k1}, m \in G_{k2}\}$$

Figura 2.5: Exemplificação de ligação simples.

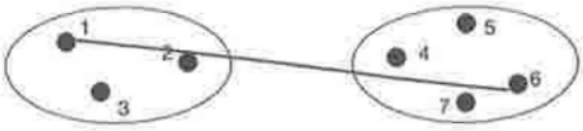


Fonte: Fávero et al. (2009)

- Método de Ligação Completa ou do Vizinho Mais Distante (*Complete Linkage*): As observações são agrupadas de acordo com o maior valor de distância, ou seja, as duas observações menos semelhantes.

$$d(G_{k1}, G_{k2}) = max\{d_{im} : i \in G_{k1}, m \in G_{k2}\}$$

Figura 2.6: Exemplificação de ligação completa.

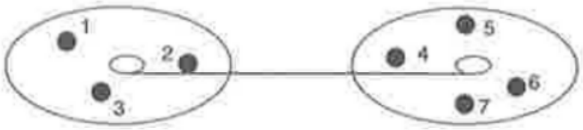


Fonte: Fávero et al. (2009)

- Método do Centroide (*Centroid Method*): Os grupos são unidos de acordo com os vetores de médias (centroides), ou seja, menor valor de distância entre os grupos.

$$d(G_{k1}, G_{k2}) = (\overline{x}_i - \overline{x}_m)'(\overline{x}_i - \overline{x}_m)$$

Figura 2.7: Exemplificação de centroide.



Fonte: Fávero et al. (2009)

Mingoti (2007) e Fávero et al. (2009) apresentam exemplos com cada etapa do processo de agrupamento hierárquico utilizando as diferentes formas de ligação.

O agrupamento não hierárquico necessita que seja definido previamente o número de grupos (K) a serem formados e funciona de forma iterativa para encontrar a melhor solução, segundo algum critério de qualidade de partição. Os métodos não hierárquicos não requerem o cálculo e armazenamento de uma nova matriz de distâncias a cada etapa do processo, o que reduz o tempo computacional e possibilita sua aplicação em grandes bases de dados Fávero et al. (2009).

Algoritmos de agrupamentos que tem como base o erro quadrático buscam encontrar a melhor separação por meio de iterações, em que cada observação pode ser alocada em outro grupo caso reflita em uma melhora do agrupamento. Minimizar o erro quadrático, ou a variação dentro de um *cluster*, é equivalente a maximizar a variação entre *cluster* (JAIN; DUBES, 1988). A partição que minimiza o critério da soma dos quadrados dos erros é considerada ótima e é chamada de partição de variância mínima (XU; WUNSCH, 2008), sendo o algoritmo k-médias o principal dessa classe.

2.4.1 K-médias (*K-means*)

O método K-médias (*K-means*) (MACQUEEN et al., 1967) é um dos métodos mais tradicionais e amplamente utilizado. Consiste nos seguintes passos:

1. Gera-se K centroides (sementes, protótipos);
2. Calcula-se as distâncias de cada observação a cada centroide, alocando-as ao grupo mais próximo. Normalmente a distância euclidiana é utilizada;
3. Atualizam-se os centroides pela média de cada atual grupo;
4. Repete-se as etapas 2 e 3 até que não haja mudança na formação dos grupos.

O K-médias objetiva particionar os dados em K grupos de forma a minimizar E, que é definido como:

$$E = \sum_{k=1}^{K} \sum_{\mathbf{x}_{i.} \in G_k} d(\mathbf{x}_{i.}, \bar{\mathbf{x}}^k)^2$$

em que $\mathbf{x}_{i.}$ é a i-ésima observação, $\bar{\mathbf{x}}^k$ é a média do k-ésimo grupo e $d(\mathbf{x}_{i.}, \bar{\mathbf{x}}^k)$ é a distância entre a observação e o centroide do k-ésimo grupo, tendo como base critério de mínimos quadrados. A complexidade do algoritmo K-médias é $O(n)$, uma vez que o número de iterações é tipicamente pequeno e $K << n$ (BARBARA, 2000).

Uma desvantagem do método é que a escolha do número K de grupos é subjetiva e que a escolha dos centroides iniciais influencia no resultado final. No entanto, existem métodos que focam em auxiliar essa escolha, dentre os quais:

- Técnica hierárquica aglomerativa: usada para obter K;
- Aleatória: escolhe dentro do conjunto de dados a ser analisado;
- Variável aleatória: escolhe a variável aleatória de maior variância e divide em K intervalos, a semente é o centroide de cada intervalo;
- Valores discrepantes: os K elementos discrepantes nas variáveis conjuntamente no conjunto de dados;
- Escolha prefixada: arbitrariamente pelo pesquisador;
- K primeiros valores: as K primeiras observações do conjunto de dados.

Ainda, K-médias é sensível ao modo de atualização dos centroides e em geral produz grupos desbalanceados. Também, conforme Chen et al. (2016), o K-médias e a maioria dos algoritmos de agrupamento são apresentados para agrupar dados numéricos e por isso tem sido desafiador agrupar dados que envolvem variáveis categóricas. Por esse motivo, várias versões para solucionar questões específicas do algoritmo original foram surgindo ao longo dos anos.

O K-medoide (*K-medoid)* foi proposto por Rdusseeun e Kaufman (1987), representa os centroides dos grupos a partir de pontos observados (medoids) de forma a minimizar a soma das distâncias em relação aos outros pontos do grupo e sendo não menos suscetível a valores atípicos do que K-médias. Já Huang (1998) idealizou o K-modas (*K-modes*) e o K-protótipo (*K-prototypes*). O primeiro utiliza-se apenas com dados categóricos, em que a medida de posição moda é empregada ao invés da média na atualização dos centroides. Já o segundo é utilizado no caso de dados mistos, onde considerando uma medida de dissimilaridade combinada é feita a agregação do K-medias e do K-modas.

2.4.2 Outros métodos

Muitos outros métodos de agrupamentos são encontrados na literatura objetivando melhorar o algoritmo base K-médias ou mesmo suprir deficiências desse, mas também existem outras abordagens devido as variadas áreas de aplicação e aos diversos critérios de agrupamento.

Ainda na vertente dos métodos hierárquicos, BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*), CURE (*Clustering Using Representatives*) e ROCK (*RObust Clustering using linKs*) são algumas abordagens visando melhorias como complexidade, por exemplo.

O algoritmo *fuzzy* c-*means* (BEZDEK; EHRLICH; FULL, 1984), assim como o K-médias, requer a pré-especificação do número de grupos desejados e se utiliza de processo iterativo, no entanto realiza a atribuição de grau de pertencimento de cada observação aos grupos. Algumas propostas de agrupamento para dados categóricos considerando a abordagem *fuzzy* são as de Huang (1998) e Kim, Lee e Lee (2004). Também em uma vertente próxima, existem os métodos baseados em densidade. Esses algoritmos assumem que os *clusters* são regiões de alta densidade de observações, separadas por regiões com baixa densidade Faceli (2011). Como exemplo tem-se o DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) (ESTER et al., 1996).

Uma outra abordagem para análise de agrupamentos é a utilização de redes neurais artificiais, em que as variáveis sob estudo são as variáveis de entrada e a de saída são os grupos. Alguns algoritmos desse tipo são o SOM (*Self-Organizing Map*) (KOHONEN, 2001) e o SOTA (*Self-Organizing Tree Algorithm*) (HERRERO; VALENCIA; DOPAZO,

2001).

Também, existem métodos de agrupamentos que tem como base a teoria dos grafos como o HSC (*Highly Connected Subgraph*) (HARTUV; SHAMIR, 2000) e o CLICK (*Cluster identification via Connectivity Kernels*) (SHARAN; SHAMIR, 2000). E, ainda, *ensemble* para combinação de agrupamentos utilizando somente um tipo de algoritmo ou vários conjuntamente.

2.5 Medidas de avaliação de agrupamentos

As medidas de avaliação de agrupamentos, também conhecidas como critérios ou índices, têm como objetivo avaliar a qualidade dos agrupamentos encontrados. Isso é especialmente importante uma vez que cada algoritmo e/ou parâmetros de entrada diferentes podem revelar estruturas de agrupamento distintas.

Ferrari et al. (2014) apresentam que as medidas de avaliação tem como base dois critérios:

- Compactação/Compacidade: reflete quão próximas as observações de cada grupo estão entre si (intragrupo);
- Separação/Separabilidade: reflete quão distante os grupos estão entre si (intergrupos).

Ainda, as medidas de avaliação são frequentemente divididas em:

- Medidas externas: avaliam a estrutura encontrada comparando-as com uma conhecida *a priori* (informação externa);
- Medidas internas: avaliam a estrutura encontrada, com relação a algum critério, unicamente com informações dos próprios dados (informação interna).

Embora todos esses índices sejam úteis em determinadas situações, eles não são de uso geral (SAITTA; RAPHAEL; SMITH, 2007), já que não se pode garantir que um bom desempenho em um conjunto de dados ocorra também em diferentes conjuntos. Conforme Xu e Wunsch (2008), não se deve depender de apenas uma regra para selecionar o número de grupos e, sim, resumir os resultados de várias técnicas.

2.5.1 Medidas internas de validação

2.5.1.1 *Within-cluster Sum of Squares* (WSS)

A medida *Within-cluster Sum of Squares* (WSS), ou também Soma de quadrados dentro do *cluster*, é uma das medidas mais conhecidas para determinar o número

ótimo de grupos. Ela tem como objetivo quantificar a homogeneidade interna dos grupos (intragrupos) e é definida como:

$$WSS = \sum_{k=1}^{K} WSS(G_k) = \sum_{k=1}^{K} \sum_{\mathbf{x}_{i.} \in G_k} d(\mathbf{x}_{i.}, \mathbf{c}_k)^2,$$

em que K representa o número total de grupos, G_k o k-ésimo grupo, $\mathbf{x}_{i.}$ a i-ésima observação e $\mathbf{c}_k$ o centroide do k-ésimo grupo. Logo, a WSS representa a soma dos quadrados das distâncias de cada elemento ao seu respectivo centroide.

Para a identificação do número ótimo de grupos, calcula-se a WSS considerando diferentes números de grupos (K) e plota-se os valores de WSS em relação ao número de grupos. Assim, o número ótimo é obtido quando se observa um "padrão cotovelo" no gráfico, por isso esse gráfico é popularmente conhecido como "Gráfico do cotovelo" ou "*Elbow plot*", ou seja, quando o gráfico apresenta uma redução na medida (não aumenta tanto a homogeneidade interna do grupo).

2.5.1.2 Dunn *Index* (DU)

De acordo com Saitta, Raphael e Smith (2007) o Dunn *index* (DU), ou também índice de Dunn, tem como objetivo maximizar a distância intergrupos enquanto minimiza a distância intragrupos. Isso é feito por meio da comparação da menor distância entre os grupos e o diâmetro do grupo com maior dispersão, sendo definido como:

$$DU = \min_{k1=1,\ldots,K} \left[\min_{k2=1,\ldots,K, k1 \neq k2} \left(\frac{d(G_{k1}, G_{k2})}{\max_{k3=1,\ldots,K} diam(G_{k3})} \right) \right],$$

em que $d(G_{k1}, G_{k2}) = min_{\mathbf{x_i}. \in G_{k1}, \mathbf{x_m}. \in G_{k2}} d(\mathbf{x_i}., \mathbf{x_m}.)$ e $diam(G_{k3}) = \max_{\mathbf{x_i}., \mathbf{x_m}. \in G_{k3}} d(\mathbf{x_i}., \mathbf{x_m}.)$.

Assim como para WSS, plota-se o gráfico do índice *versus* diferentes quantidades de grupos (K), no entanto não se espera observar nenhuma tendência. Conforme Faceli (2011) e Xu e Wunsch (2008), um grande valor de DU sugere a presença de grupos bem separados e compactos, e o valor de K correspondente pode ser uma indicação do número de grupos que melhor se ajusta aos dados. Vale ressaltar que quanto maior for o número de grupos maior é o custo computacional.

2.5.1.3 Davies-Bouldin *Index* (DB)

O Davies-Bouldin *index* (DB), ou índice de Davies-Bouldin, foi criado por Davied Davies e Donald Bouldin em 1979. Segundo Ferrari et al. (2014), se baseia nos trabalhos de Dunn, determinando a qualidade do agrupamento através das medidas intragrupos e intergrupos, sendo definido como:

$$DB = \frac{1}{k}\sum_{k1=1}^{K} \max_{k2=1,\ldots,K,k1\neq k2} \left[\frac{diam(G_{k1}) + diam(G_{k2})}{d(G_{k1}, G_{k2})}\right]$$

em que K é o número de grupos, G são os grupos, $diam(G) = \max_{i,m\in G} d(i,m)$ e $d(G_{k1}, G_{k2}) = \min_{i\in G_{k1}, m\in G_{k2}} d(i,m)$.

Esse índice assume valores no intervalo $[0, \infty)$ e, de acordo com Saitta, Raphael e Smith (2007) e Xu e Wunsch (2008), pequenos valores revelam o número potencial de grupos além de indicarem baixa dispersão intragrupo e grandes distâncias intergrupos.

2.5.1.4 *Silhouette index* (sil)

O *silhouette index* (sil), ou índice de silhueta, foi criado por Rousseeuw (1987), sendo um dos mais conhecidos para estimar o número de grupos. Para construir silhuetas, precisamos de duas coisas: a partição que obtivemos (pela aplicação de alguma técnica de agrupamento) e a coleta de todas as proximidades entre os objetos sendo definida como:

$$sil(i) = \frac{b(i) - a(i)}{\max\left[a(i), b(i)\right]}$$

em que $a(i)$ é a distância média da i-ésima observação a todas as demais do mesmo grupo e $b(i)$ é a distância mínima da i-ésima observação a todas as demais que não pertencem ao seu grupo.

A silhueta geral é dada pela média das silhuetas dos grupos que, por sua vez, equivalem a média de $sil(i)$ de todas as observações pertencentes ao correspondente grupo. Segundo Rousseeuw (1987) e Ferrari et al. (2014), o índice pode assumir valores no intervalo [-1,1] e uma maneira de escolher o número de grupos é selecionar o valor de K que apresenta maior índice.

Capítulo 3

Análise de Agrupamento Multinível

Neste capítulo são apresentados aspectos relacionados à metodologia proposta. Inicialmente apresentamos uma revisão de literatura sobre Análise de Agrupamento Multinível, na qual nota-se que não existe um consenso sobre tal abordagem. Em seguida, expomos as características de uma estrutura multinível, com motivações e exemplificações de possíveis áreas de aplicação. Posteriormente, apresentamos a distância de Hellinger e, por fim, a construção de uma medida mista multinível, bem como a nossa proposta de Análise de Agrupamento Multinível.

3.1 Visão geral sobre Agrupamento Multinível

O termo agrupamento multinível não surgiu de forma unificada. Percebe-se que tal termo é utilizado em diversas áreas do conhecimento nas mais diversas circunstâncias, sendo empregado desde áreas como a Engenharia e Ciência da Computação, até a Psicologia, Geofísica e Saúde Pública. No entanto, de forma geral, observa-se que este termo comumente é relacionado à computação distribuída, modelagem multinível (também conhecida como modelo linear hierárquico ou modelo linear misto) ou à análise de agrupamento.

Nas áreas de Ciência e Engenharia da Computação, Das, Chaki e Biswas (2013) e Feng Zhongyan; Feng Z.;Wan (2006) utilizam o termo *cluster* multinível referente à estrutura da rede de sensores e à arquitetura do sistema de armazenamento, respectivamente. Das, Chaki e Biswas (2013) apresentam um protocolo baseado em *cluster* multinível para melhorar a eficiência energética da rede de sensores sem fio (WSN). Já Feng Zhongyan; Feng Z.;Wan (2006) apresentam e implementam o MCFS (sistema de arquivos de *cluster multinível*; *Multilevel Cluster File System*).

A modelagem multinível considera a estrutura natural aninhada (níveis) dos dados e a variabilidade dentro e entre os níveis, permitindo compreender interações complexas.

Na área da Saúde, Merlo et al. (2005) tem como premissa que o contexto da vizinhança (nível 2) desempenha um papel importante na determinação da saúde e bem-estar dos indivíduos (nível 1). Dessa forma, estudam o impacto das características da vizinhança na saúde individual, como densidade populacional e a disponibilidade de serviços de saúde. Na área da Educação, Andrade e Laros (2007) estuda os fatores associados ao desempenho escolar de alunos da 3ª série do Ensino Médio nas disciplinas de Língua Portuguesa e Matemática com base em dados do Sistema de Avaliação da Educação Básica (SAEB). Verificaram que as variáveis como recursos culturais (a nível escolar) e atraso escolar (a nível aluno) apresentam grande influência.

No que diz respeito a Análise de agrupamento, na Geofísica, por exemplo, Inzoli, Giudici e Huisman (2015) utiliza análise de agrupamento e PCA (*Principal Component Analysis*) para inferir propriedades sedimentológicas de depósitos aluviais por meio da comparação da variabilidade elétrica e sedimentológica intracluster (homogeneidade interna). Como produto final estabeleceram ligação entre os parâmetros geológicos e geofísicos e os resultados podem ser usados para caracterizar sedimentos. Na Engenharia, Wilschut et al. (2017) propõem um novo algoritmo de agrupamento baseado no agrupamento de Markov. Na Ciência da Computação, por sua vez, Kang, Wu e Yang (2013) propõe uma nova abordagem não supervisionada para a construção automática de clusters de imagens a partir de imagens não ordenadas. Consiste em fazer a representação da imagem a partir do *bit*, formando uma sequência binária, para então descrever a imagem quantitativamente a partir de medidas de popularidade e dissimilaridade de imagens inspiradas na teoria da informação.

3.2 Estrutura multinível

Em muitas áreas, como nas Ciências Sociais e Humanas, os dados coletados são frequentemente de pessoas agrupadas em *clusters* (LAROS; MARCIANO, 2008), circunstância na qual surge a estrutura multinível dos dados. Dessa forma, uma estrutura multinível pode ser caracterizada pela presença de dados aninhados, que podem ser vistos como dados provenientes de estruturas hierárquicas naturais, também chamadas de níveis de agregação. Esse é o tipo de estrutura considerado nesse trabalho.

Dados com estrutura multinível possibilitam que as unidades de análise em cada nível de agregação sejam correlacionadas, ou seja, possibilita considerar as variações existentes em cada nível, desde o nível micro até o macro, e entre os níveis. De acordo com Peugh (2010) e Laros e Marciano (2008), esse tipo de dado aparece com frequência em pesquisas educacionais onde, por exemplo, pode-se considerar para o 1º nível os alunos, para o 2º nível os professores ou as turmas e para o 3º nível as escolas.

Para ilustrar um exemplo, suponha que se tem o interesse em investigar o perfil de diversidade nas empresas. Nesse caso, para o 1º nível tem-se os funcionários e para o 2º nível as empresas. Ainda, é possível inserir mais um nível de agregação, incorporando as posições (cargos) de ocupação dentro das empresas. Segundo Laros e Marciano (2008), a estrutura hierárquica também pode ser vista, por exemplo, na pesquisa familiar (membros agrupados em famílias), na pesquisa metodológica (respondentes agrupados por entrevistadores) ou nos estudos transnacionais (indivíduos agrupados em unidades nacionais).

3.3 Distância de Hellinger

A distância de Hellinger (1909) surgiu a partir da integral de Hellinger, que foi definida por Ernst David Hellinger. Tal medida é utilizada em diversas áreas do conhecimento como, por exemplo, Teoria da informação, Estatística, Processamento de imagem, Bioinformática, Econometria e Sensoriamento remoto.

Considerando a conjuntura social, Almeida, Quituisaca-Samaniego e Antamba (2019) exploraram a heterogeneidade da estrutura econômica de pequenas empresas no Equador, relacionando as atividades econômicas em uma unidade geográfica com a representação de cada setor econômico (indústria, comércio e serviços). A distância de Hellinger é utilizada para comparar as estruturas econômicas e as distribuições geográficas dos diferentes setores e atividades da economia, agrupando setores econômicos que tenham distribuições geográficas semelhantes e reconhecendo vinculo entre os setores. Também, Mendoza-Morales, González-Sansón e Aguilar-Betancourt (2016) investigam os fatores ambientais que influenciam na dinâmica de produção de serapilheira. Para isso, utilizou-se uma análise de redundância baseada na distância de Hellinger, permitindo uma compreensão detalhada de como variáveis ambientais específicas impactam nessa produção. Por outro lado, uma aproximação baseada em distâncias para generalizar com qualquer medida de proximidade usando o algoritmo fuzzy k-means é proposta por Oliva et al. (2001) em que, a partir de qualquer matriz de proximidades, calcula-se a distância ao centroide sem necessidade de obter explicitamente as coordenadas dos mesmos.

Na conjuntura de análise de imagem, Cassetti, Gambini e Frery (2013) propõem um método de estimação para dados com ruído speckle, um tipo de ruído multiplicativo, não gaussiano e comum em imagens médicas, de radar e de sensoriamento remoto. A metodologia envolve o uso de medidas de dissimilaridade entre distribuições de probabilidade para ajudar a comparar e ajustar modelos aos dados observados. A estimação é realizada por meio da minimização das distâncias entre a distribuição empírica dos dados e a função de densidade da distribuição teórica, utilizando as distâncias de Hellinger,

Rényi e Triangular. Já Violini e Pasapera (2016) objetivam identificar culturas de verão através de dados de imagens e radar de satélite na área central da província de Córdoba (ARG). Para identificar diferenças entre culturas nas imagens, calculou-se as distâncias de Hellinger entre todas as parcelas plantadas com a mesma cultura e entre aquelas com culturas diferentes.

O objetivo da distância de Hellinger é medir o grau de sobreposição entre duas distribuições de probabilidade. Dessa forma, considerando P e Q duas distribuições de probabilidade discretas, a distância de Hellinger é definida como:

$$d_H(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{c}(\sqrt{p_i} - \sqrt{q_i})^2},$$

em que $P = p_1, p_2, ..., p_c$, $Q = q_1, q_2, ..., q_c$ e $0 \leq d_H \leq 1$. Quanto maior for a distância entre as distribuições, menor será o grau de sobreposição entre elas. Assim, 0 representa que as distribuições são idênticas e 1 que são completamente diferentes.

Vale ressaltar que a distância de Hellinger é considerada uma métrica, pois satisfaz as propriedades apresentadas no Capítulo 2. Além disso, possui como propriedade a associação com a distância de Bhattacharyya (1943) por meio da relação:

$$d_H(p,q) = \sqrt{1 - BC(p,q)}$$

em que $BC = \sum_{i=1}^{c}\sqrt{p_i q_i}$ é o coeficiente de Bhattacharyya. Maiores detalhes dessa relação são encontrados em Derpanis (2008).

3.4 Proposta de Agrupamento multinível

Dados com estrutura multinível possuem uma organização hierárquica natural, onde as observações estão aninhadas em diferentes níveis como visto na Seção 3.2. Essa estrutura tem como premissa que fatores em diferentes níveis exercem certa influência nas unidades de análise e que existe a variação das características das unidades tanto dentro de cada nível quanto entre os níveis, Por exemplo, em um estudo educacional, alunos (nível 1) estão aninhados dentro de turmas (nível 2), que por sua vez estão aninhadas em escolas (nível 3). As características dos alunos podem variar não só entre alunos dentro da mesma turma, mas também entre diferentes turmas e escolas.

A Análise de Agrupamento tem como objetivo encontrar grupos semelhantes de unidades de análise de acordo com algum critério específico. Quando aplicada a dados com estrutura multinível é essencial considerar a complexidade desse tipo de dado para se ter uma compreensão mais fidedigna. Dessa forma, é importante considerar a estrutura

hierárquica e a dependência entre as observações nos diferentes níveis, isto é, a variação tanto dentro quanto entra os níveis.

Como exposto no Capítulo 2, o algoritmo k-médias é um dos mais utilizados na literatura de análise de agrupamento. É um algoritmo simples, que não possui complexidade na sua execução e que tem apresentado bom desempenho tanto em performance quanto em custo computacional nas mais variadas aplicações. No entanto, tal algoritmo tem como pilar o uso da distância Euclidiana, que pode ser utilizada apenas no cenário de variáveis quantitativas.

A nossa proposta de agrupamento multinível consiste em uma adaptação do algoritmo k-médias para dados multiníveis, ou seja, incorporando a estrutura hierárquica dos dados no cálculo das distâncias entre as observações por meio de uma abordagem de ponderação.

A medida mista multinível visa incorporar a estrutura de dependência dos dados multiníveis agregando não só as informações do nível em questão, mas também do nível anterior. Para isso, considerou-se a composição das similaridades Euclidiana, de Jaccard e de Hellinger, de acordo com o exibido na seção 2.3.3. Os termos das similaridades Euclidiana e de Jaccard são utilizados para as variáveis quantitativas e qualitativas, respectivamente, levando-se em conta a proporção de variáveis de cada tipo ao calcular a distância em dados mistos. Já a similaridade de Hellinger refere-se a incorporação da estrutura multinível, em que captura-se para o nível anterior o grau de semelhança entre as distribuições de probabilidade estimada para o referido nível. Dessa forma, para duas observações (i e m) a medida mista multinível é obtida como:

$$s(i,m) = (1-\alpha) \times [w_p \times s_p(i,m) + w_q \times s_q(i,m)] + [\alpha \times s_H(i,m)]$$

em que $w_p = \frac{p}{p+q}$ e $w_q = \frac{q}{p+q}$ são os pesos das parcelas referentes as variáveis quantitativas e qualitativas, s_p, s_q e s_H são as similaridades Euclidiana, de Jaccard e de Hellinger, respectivamente, e $0 \leq \alpha \leq 1$. Vale ressaltar que ajustando o valor do hiperparâmetro α controla-se a importância (peso) relativa das diferentes distâncias na medida final de dissimilaridade, ou seja, a incorporação de informação do nível anterior.

Desse modo, o método heurístico consiste nos seguintes passos:

- Passo 1:

    1. No nível 1, agrega-se as observações por método hierárquico e obtem-se a proporção de observações em cada grupo; ou
    2. Agrega-se as observações em relação a uma variável categórica obtendo-se a proporção em cada classe, a fim de é obter alguma caracterização em grupos.

- Passo 2:

    1. No nível 2, gera-se K centroides (sementes);
    2. Calcula-se as distâncias de cada observação a cada centroide, alocando-as ao grupo mais próximo. Utiliza-se a medida mista multinível apresentada anteriormente;
    3. Atualizam-se os centroides pela média e/ou moda de cada atual grupo;
    4. Repete-se as etapas 2 e 3 até que não haja mudança na formação dos grupos.

Vale destacar que essa estrutura pode ser implementada não apenas para dois níveis, mas para múltiplos níveis replicando a mesma lógica.

Capítulo 4

Estudo de Simulação

Neste capítulo são descritos e apresentados os cenários e principais resultados dos estudos de simulação realizados. Tais estudos foram conduzidos a fim de avaliar o comportamento da medida mista multinível proposta sob diferentes situações e verificar se utilizando-a se obtém, com êxito, a identificação do número de grupos encontrados, uma vez que esse foi estabelecido previamente.

4.1 Estrutura dos dados

Foram realizados experimentos alterando a quantidade de variáveis quantitativas, o número de classes da variável referente ao nível anterior e o hiperparâmetro (α). A Tabela 4.1 apresenta um exemplo da estrutura das bases de dados simuladas.

Tabela 4.1: Exemplificação da estrutura das bases de dados simuladas.

ID	**BIN1**	**BIN2**	**CONT1**	**H_P1**	**H_P2**	**H_P3**	**H_P4**	**GRUPO**
1	1	1	0,4166	0,7240	0,2466	0,0290	0,0004	1
2	0	1	0,4793	0,7323	0,2396	0,0272	0,0009	1
3	0	1	0,3962	0,7332	0,2408	0,0251	0,0009	1
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
1.500	0	1	0,5251	0,0009	0,0286	0,2375	0,7330	3

A estrutura dos dados é composta por informações referentes ao nível 2 (macro) e informações referentes ao nível 1 (micro). As variáveis binárias e contínuas são relativas ao nível 2 e para o nível 1 tem-se as proporções para cada categoria de alguma variável qualitativa que separe os grupos no respectivo nível. Por exemplo, em um estudo envolvendo alunos (nível 1) e escolas (nível 2) as informações referentes ao nível 2 poderiam ser tipo de rede de ensino e quantidade de polos de ensino e referentes ao nível 1 a raça dos alunos através da proporção de alunos classificados como pretos, brancos e indígenas. Nesse caso, teríamos as variáveis H_P1, H_P2 e H_P3 referentes ao nível anterior.

As variáveis binárias (BIN1 e BIN2) foram geradas a partir da distribuição de Bernoulli em que BIN1 $\sim$ Bernoulli($p = \frac{1}{2}$) e BIN2 $\sim$ Bernoulli($p = \frac{2}{3}$). Em todas as bases de dados simuladas o número de variáveis binárias geradas foi o mesmo. As variáveis contínuas foram geradas a partir da distribuição Normal univariada ou multivariada, de acordo com a especificidade do experimento. As variáveis referentes ao nível anterior foram geradas a partir da distribuição Binomial, com número de tentativas igual ao número de classes desejadas de Hellinger -1 e probabilidade de sucesso p_0, de acordo com a especificidade do experimento. As informações referentes ao nível anterior são obtidas por meio das proporções de cada classe.

4.2 Simulação

Para todos os experimentos desse cenário foram considerados três grupos ($k = 3$), duas variáveis binárias ($p = 2$), BIN1 e BIN2, bem como 500 observações para cada grupo ($n = 500$).

Experimento 1

Para esse experimento tem-se uma variável contínua (X_1) e quatro classes de Hellinger (C_H) que foram gerados com os seguintes parâmetros:

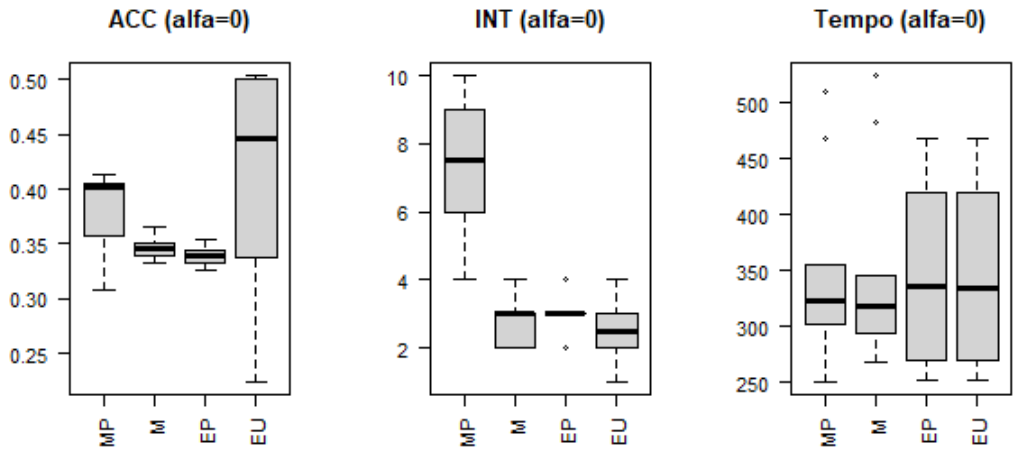
- Grupo 1:
  - $X_1 \sim N(0; 1)$
  - $C_H \sim Bin(3; 0, 1)$
- Grupo 2:
  - $X_1 \sim N(1; 4)$
  - $C_H \sim Bin(3; 0, 5)$
- Grupo 3:
  - $X_1 \sim N(2; 2)$
  - $C_H \sim Bin(3; 0, 9)$

Através de $Bin(3, p_0)$ gera-se as variáveis com informações relativas ao nível anterior. Dessa forma, temos que $P_0 = P(C_H = 0)$, $P_1 = P(C_H = 1)$, $P_2 = P(C_H = 2)$ e $P_3 = P(C_H = 3)$ representa a probabilidade de cada categoria da desejada, sendo $P_0 + P_1 + P_2 + P_3 = 1$.

A identificação do número de grupos encontrados foi realizada por meio da medida de acurácia (ACC), que representa a taxa de acerto do grupo da observação. Também foi verificado o o número de interações até não haver mais mudança na formação dos grupos (INT) e o tempo de execução do agrupamento (TEMPO). As medidas consideradas foram a mista proposta (MP), mista padrão (M), euclidiana ponderada (EP) e euclidiana padrão (EU).

Para $\alpha = 0$ (Figura 4.1), em relação à acuracidade, a distância euclidiana apresentou maiores valores, mas também uma grande variabilidade. Seguidamente tem-se a distância mista multinível proposta, que apresentou uma variabilidade menor que a euclidiana. Já em relação a quantidade de interações realizadas, a medida mista proposta foi a que obteve maior quantidade. Por outro lado, em relação ao tempo, a medida mista multinível proposta e a medida mista padrão apresentaram menores tempos de execução do algoritmo.

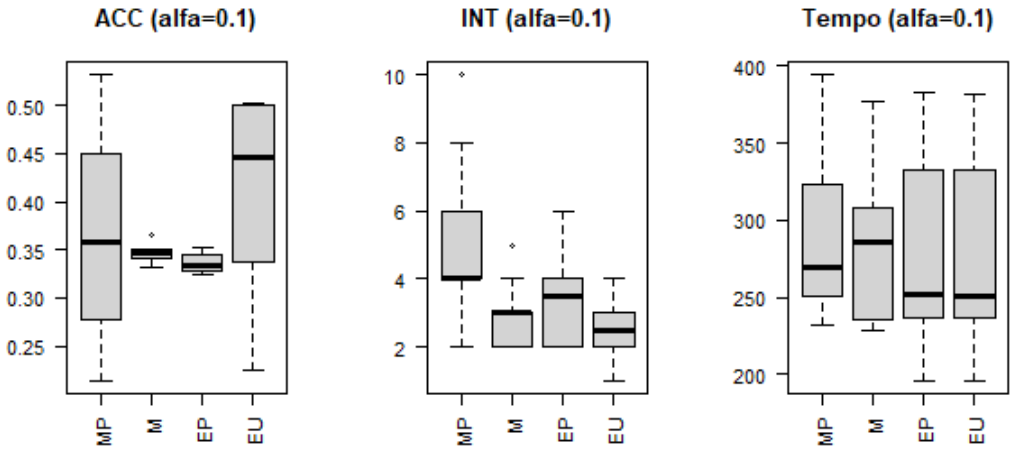
Figura 4.1: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0$ (Experimento 1).



Fonte: Próprio autor.

Para $\alpha = 0,1$ (Figura 4.2), a distância euclidiana apresentou maiores valores de ACC, mas também uma grande dispersão. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve maior quantidade. Por outro lado, em relação ao tempo, todas as medidas apresentaram grandes variações nos tempos de execução do algoritmo, no entanto a euclidiana padrão e ponderada apresentaram menores medianas de tempo.

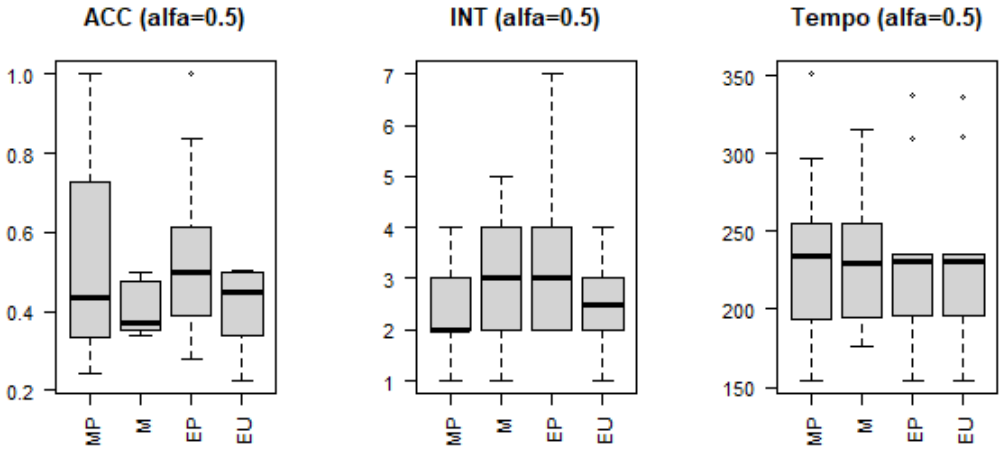
Figura 4.2: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,1$ (Experimento 1).



Fonte: Próprio autor.

Para $\alpha = 0,5$ (Figura 4.3), em relação à acuracidade, a distância mista multinível proposta apresentou grande variação nos valores de ACC, enquanto que a euclidiana ponderada apresentou menor variação. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve menor quantidade, seguida da euclidiana padrão. Por outro lado, em relação ao tempo, as medidas euclidiana ponderada e padrão apresentaram melhores resultados, com menores tempos e dispersão.

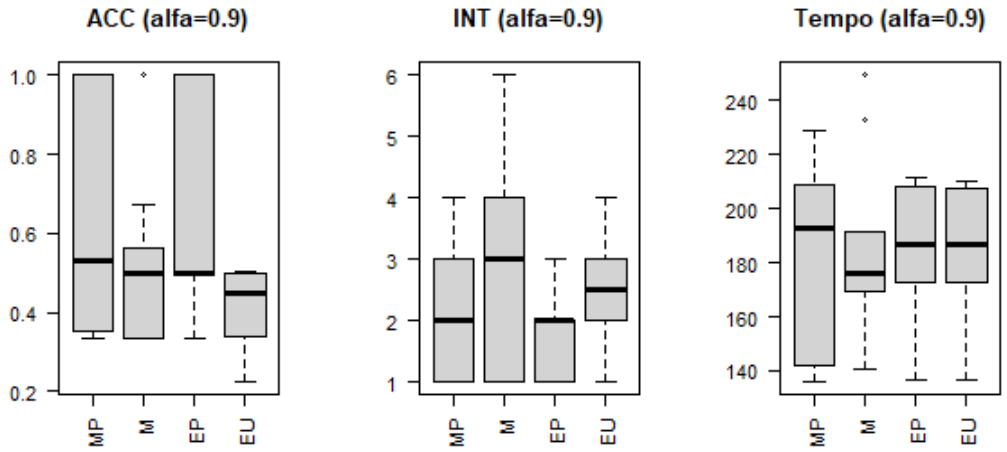
Figura 4.3: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,5$ (Experimento 1).



Fonte: Próprio autor.

Para $\alpha = 0,9$ (Figura 4.4), a distância euclidiana ponderada e mista multinível proposta apresentaram maiores valores de ACC, mas com uma grande dispersão. Já em relação a quantidade de interações realizadas, a medida euclidiana ponderada foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, a medida mista ponderada padrão apresentou melhores resultados, com menores tempos e dispersão nos tempos de execução do algoritmo.

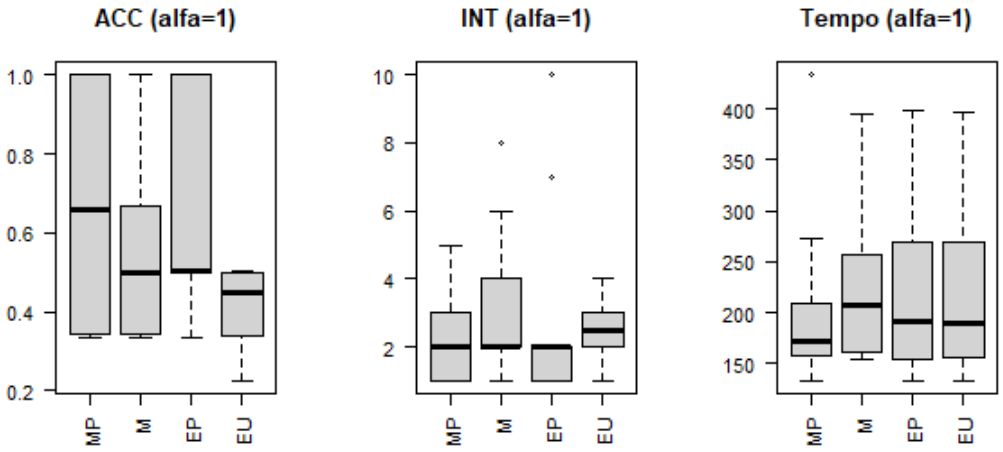
Figura 4.4: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,9$ (Experimento 1).



Fonte: Próprio autor.

Para $\alpha = 1$ (Figura 4.5), em relação à acuracidade, a distância euclidiana ponderada e mista proposta apresentaram maiores valores de ACC, mas com uma grande dispersão. Já em relação a quantidade de interações realizadas, a medida euclidiana ponderada foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, a medida mista proposta apresentou melhores resultados, com menores tempos e dispersão nos tempos de execução do algoritmo.

Figura 4.5: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 1$ (Experimento 1).



Fonte: Próprio autor.

Para o experimento 1, os maiores valores de acurácia foram obtidos com as distâncias mista multinível proposta e euclidiana ponderada para alfas intermediários e altos, ou seja, a partir de $\alpha = 0,5$. Já para a quantidade de interações, a distância euclidiana é a que apresenta menores valores. Em relação a menor tempo de execução, as distâncias euclidiana e euclidiana ponderada apresentaram menores valores para $\alpha = 0,1$ e $\alpha = 0,5$, enquanto que para os demais alfas foram as distâncias mista proposta e/ou mista padrão.

Experimento 2

Para esse experimento tem-se 10 variáveis contínuas e quatro classes de Hellinger que foram gerados com os seguintes parâmetros:

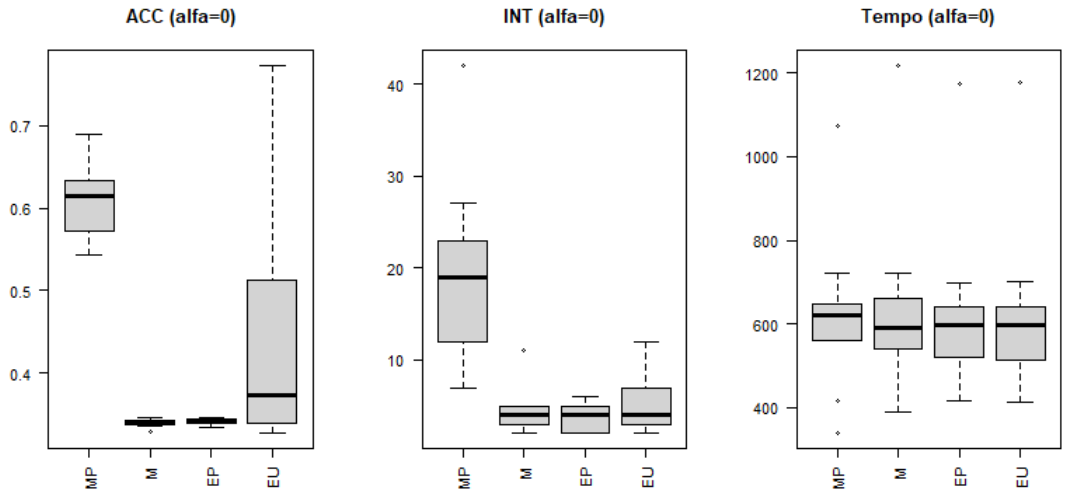
- Grupo 1:
  - $\mathbf{X} \sim N(\mathbf{0}; diag(1))$
  - $C_H \sim Bin(3; 0, 1)$
- Grupo 2:
  - $\mathbf{X} \sim N(\mathbf{1}; diag(4))$
  - $C_H \sim Bin(3; 0, 5)$
- Grupo 3:
  - $\mathbf{X} \sim N(\mathbf{2}; diag(2))$

– $C_H \sim Bin(3; 0, 9)$

Através de $Bin(3, p0)$ gera-se as variáveis relativas a $P_0 = P(C_H = 0)$, $P_1 = P(C_H = 1)$, $P_2 = P(C_H = 2)$ e $P_3 = P(C_H = 3)$, sendo $P_0 + P_1 + P_2 + P_3 = 1$.

Para $\alpha = 0$, em relação à acuracidade, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve maior quantidade. Por outro lado, em relação ao tempo, a medida mista ponderada padrão apresentou melhores resultados, com menores tempos e dispersão nos tempos de execução do algoritmo.

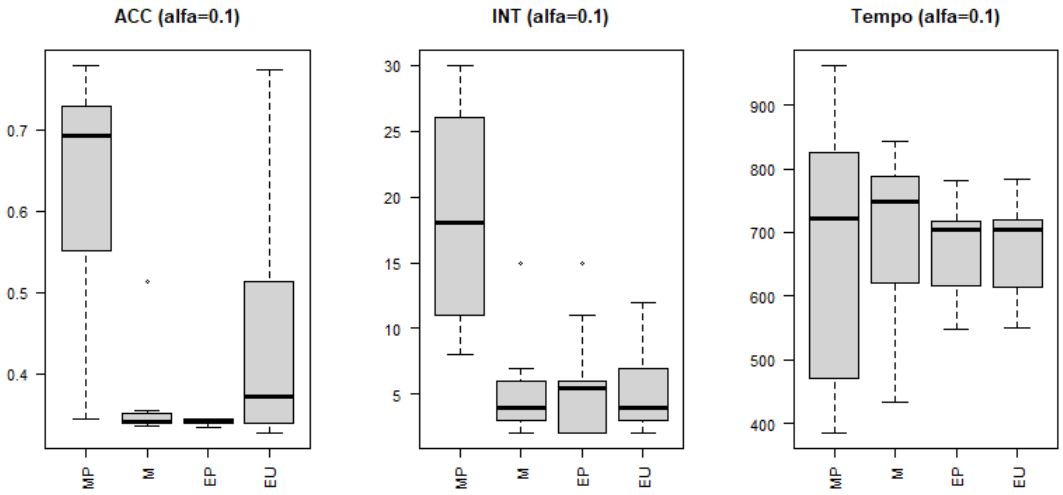
Figura 4.6: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0$ (Experimento 2).



Fonte: Próprio autor.

Para $\alpha = 0,1$, em relação à acuracidade, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve maior quantidade. Por outro lado, em relação ao tempo, as medidas apresentaram tempos medianos próximos, no entanto a medida mista ponderada proposta teve maior variabilidade nos tempos de execução do algoritmo.

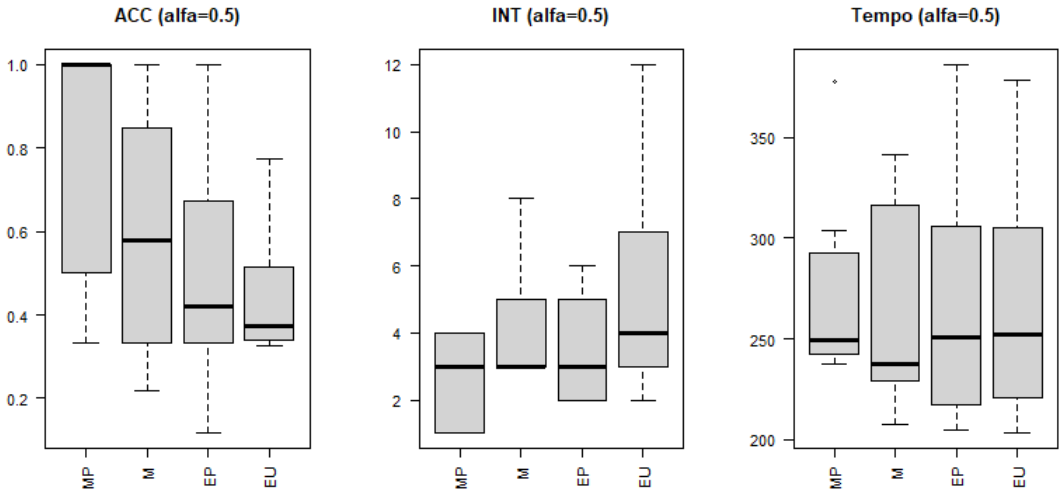
Figura 4.7: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,1$ (Experimento 2).



Fonte: Próprio autor.

Para $\alpha = 0.5$, em relação à acuracidade, a distância mista multinível proposta apresentou melhores valores de ACC, apesar da grande variação. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve menor quantidade, seguida da euclidiana ponderada. Por outro lado, em relação ao tempo, a medida mista multinível proposta apresentou melhor resultado, com menor variação no tempo de execução.

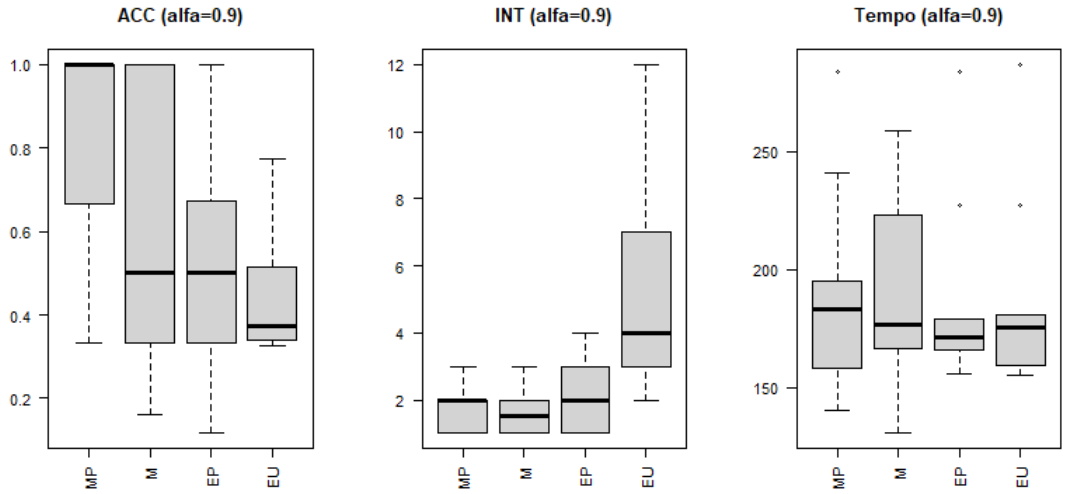
Figura 4.8: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,5$ (Experimento 2).



Fonte: Próprio autor.

Para $\alpha = 0.9$, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida mista padrão foi a que obteve menor quantidade, seguida pela mista multinível proposta. Por outro lado, em relação ao tempo, a medida euclidiana ponderada apresentou melhores resultados, com menores tempos mais estáveis.

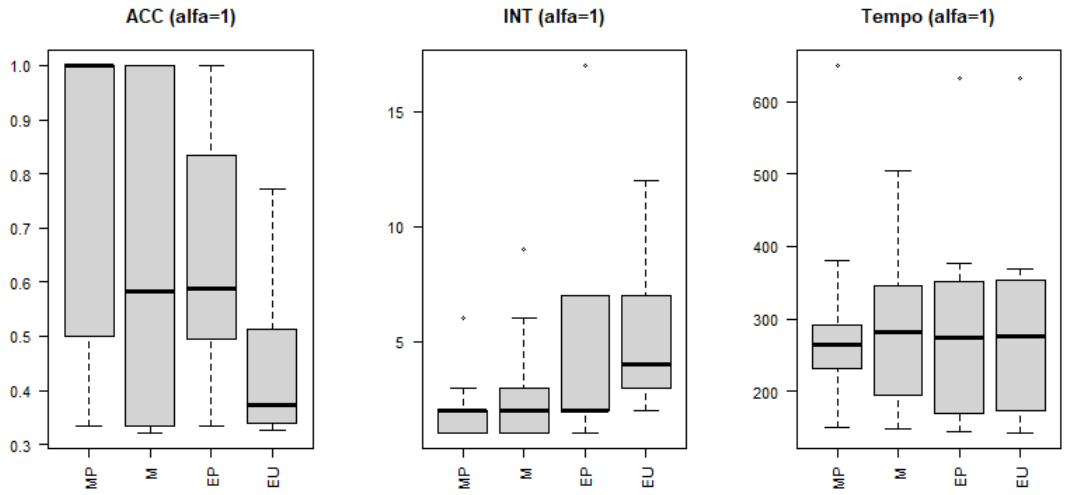
Figura 4.9: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,9$ (Experimento 2).



Fonte: Próprio autor.

Para $\alpha = 1$, em relação à acuracidade, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida mista padrão foi a que obteve menor quantidade, sendo seguida pela mista multinível proposta. Por outro lado, em relação ao tempo, a medida mista multinível proposta apresentou melhores resultados, com menores tempos nos tempos de execução do algoritmo.

Figura 4.10: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 1$ (Experimento 2).



Fonte: Próprio autor.

Para o experimento 2, os maiores valores de acurácia foram obtidos com a distância mista multinível proposta. Já para a quantidade de interações, as distâncias mista multinível proposta e padrão apresentaram menores valores. Em relação a menor tempo de execução, não se teve uma predominância de alguma das distâncias.

Experimento 3

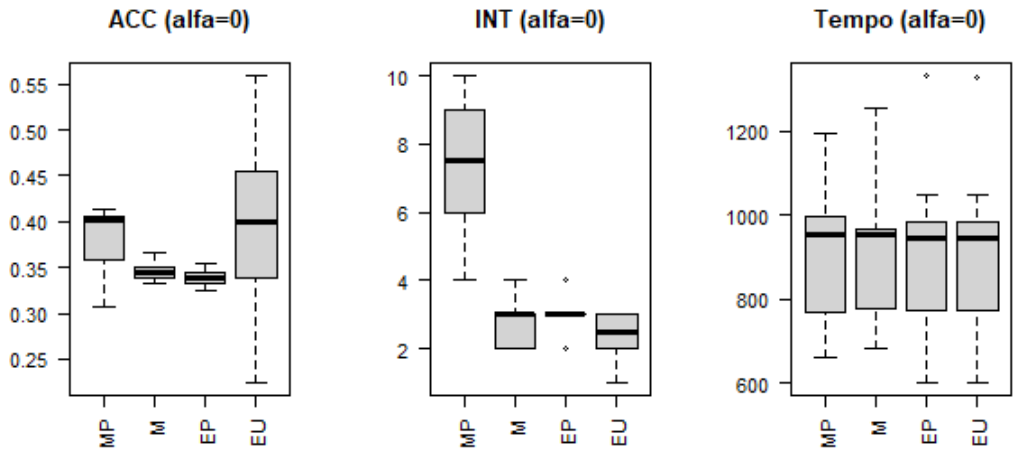
Para esse experimento tem-se uma variável contínua e 10 classes de Hellinger que foram gerados com os seguintes parâmetros:

- Grupo 1:
  - $X_1 \sim N(0; 1)$
  - $C_H \sim Bin(3; 0, 1)$
- Grupo 2:
  - $X_1 \sim N(1; 4)$
  - $C_H \sim Bin(3; 0, 5)$
- Grupo 3:
  - $X_1 \sim N(2; 2)$
  - $C_H \sim Bin(3; 0, 9)$

Através de $Bin(3, p0)$ gera-se as variáveis relativas a $P_0 = P(C_H = 0)$, $P_1 = P(C_H = 1)$, $P_2 = P(C_H = 2)$ e $P_3 = P(C_H = 3)$, sendo $P_0 + P_1 + P_2 + P_3 = 1$.

Para $\alpha = 0$, em relação à acuracidade, a distância euclidiana apresentou maiores valores de ACC, apesar da grande dispersão. Já em relação a quantidade de interações realizadas, a medida euclidiana foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, as distâncias euclidiana ponderada e padrão apresentaram melhores resultados.

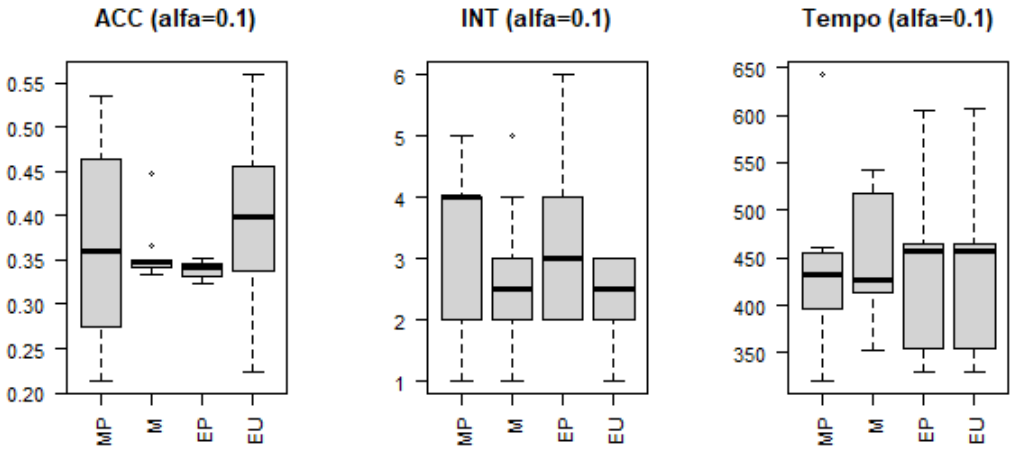
Figura 4.11: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0$ (Experimento 3).



Fonte: Próprio autor.

Para $\alpha = 0,1$, em relação à acuracidade, a distância euclidiana apresentou maiores valores, seguidamente pela mista multinível proposta. Já em relação a quantidade de interações realizadas, a medida euclidiana também foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, a medida mista multinível proposta apresentou menores tempos.

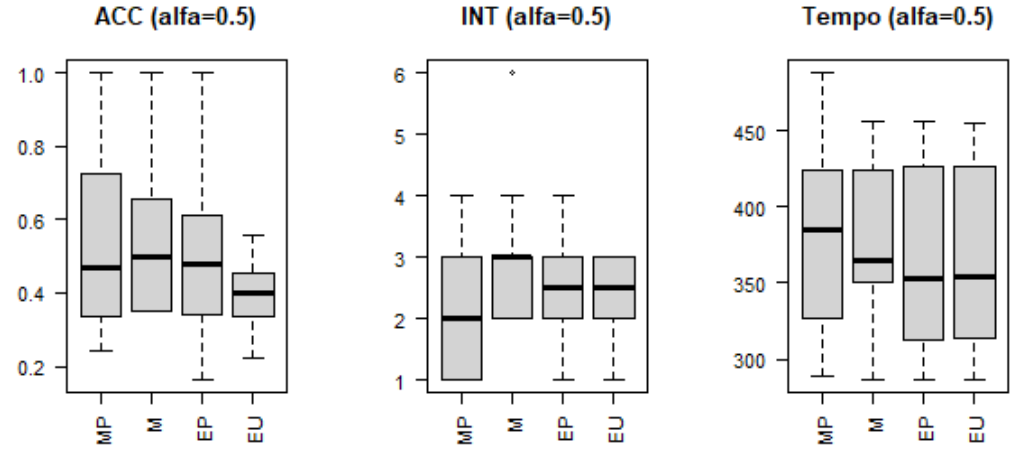
Figura 4.12: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,1$ (Experimento 3).



Fonte: Próprio autor.

Para $\alpha = 0.5$, em relação à acuracidade, a distância mista multinível proposta apresentou melhores valores de ACC, apesar da grande variação. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve menor quantidade, seguida da euclidiana. Por outro lado, em relação ao tempo, as medidas euclidiana ponderada e padrão apresentaram menores valores, apesar da grande variação.

Figura 4.13: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,5$ (Experimento 3).

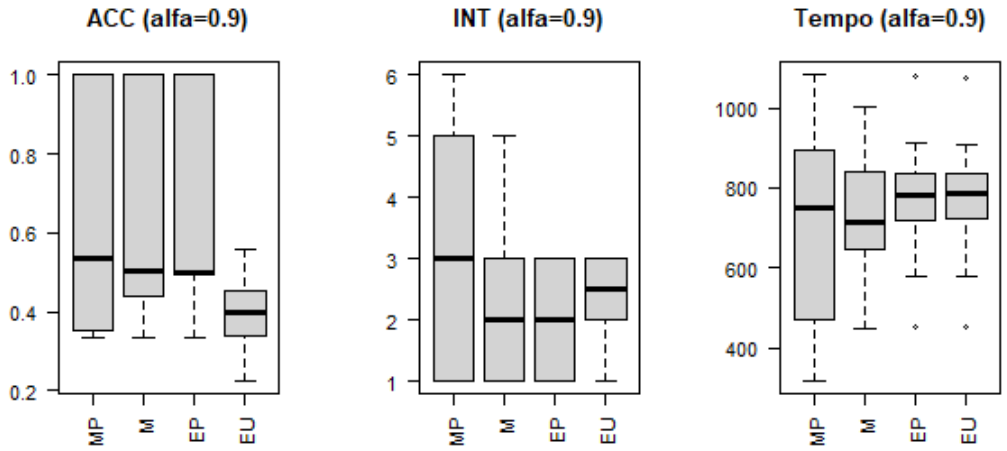


Fonte: Próprio autor.

Para $\alpha = 0.9$, a distância mista multinível proposta apresentou maiores valores de

ACC. Já em relação a quantidade de interações realizadas, a medida euclidiana ponderada foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, a medida mista multinível proposta apresentou menores resultados.

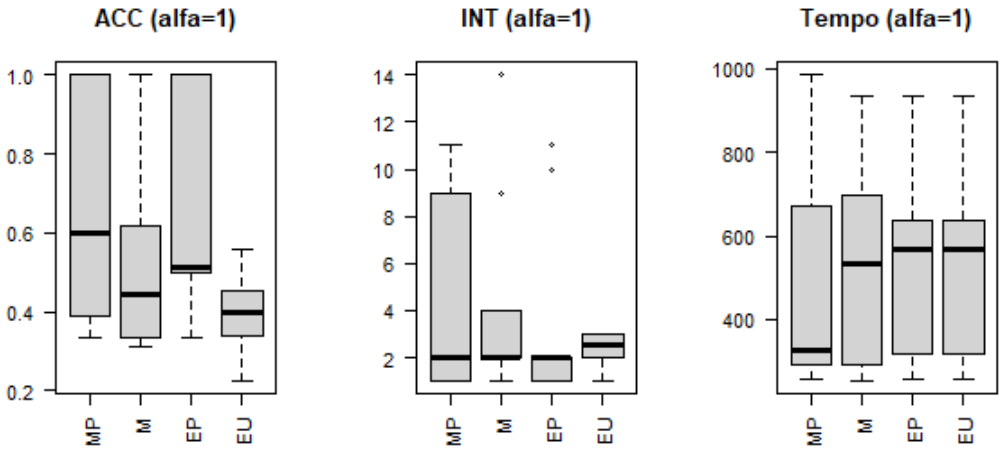
Figura 4.14: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,9$ (Experimento 3).



Fonte: Próprio autor.

Para $\alpha = 1$, em relação à acuracidade, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida euclidiana ponderada foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, a medida mista multinível proposta apresentou melhores resultados, com menores tempos nos tempos de execução do algoritmo.

Figura 4.15: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 1$ (Experimento 3).



Fonte: Próprio autor.

Para o experimento 3, os maiores valores de acurácia foram obtidos com a distância euclidiana, para alfas baixos (0, 0.1) e mista/mista propsta para alfas intermediários e altos (0.5, 0.8, 1). Já para a quantidade de interações, a distância euclidiana ponderada apresentou menores valores. Em relação a menor tempo de execução, a distância mista multinível proposta apresentou menores valores.

Experimento 4

Para esse experimento tem-se 10 variáveis contínuas e 10 classes de Hellinger que foram gerados com os seguintes parâmetros:

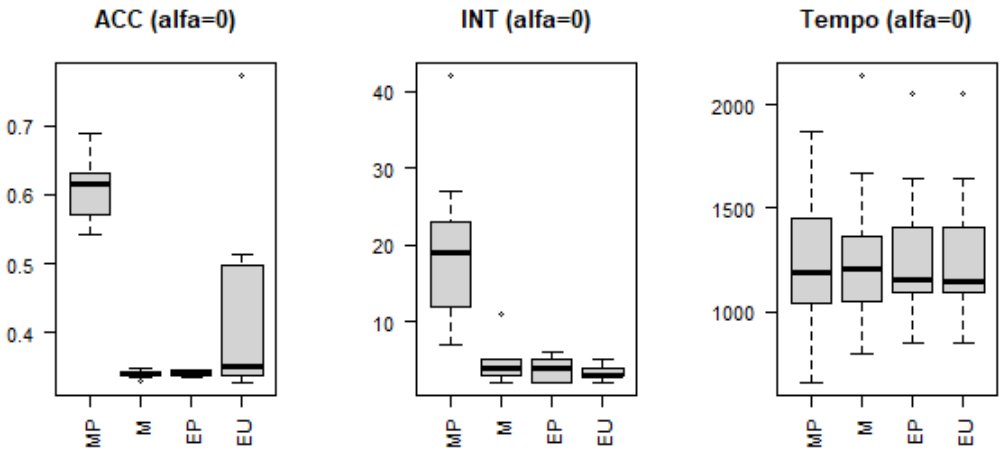
- Grupo 1:
  - $\mathbf{X} \sim N(\mathbf{0}; diag(1))$
  - $C_H \sim Bin(3; 0, 1)$
- Grupo 2:
  - $\mathbf{X} \sim N(\mathbf{1}; diag(4))$
  - $C_H \sim Bin(3; 0, 5)$
- Grupo 3:
  - $\mathbf{X} \sim N(\mathbf{2}; diag(2))$

– $C_H \sim Bin(3; 0, 9)$

Através de $Bin(3, p0)$ gera-se as variáveis relativas a $P_0 = P(C_H = 0)$, $P_1 = P(C_H = 1)$, $P_2 = P(C_H = 2)$ e $P_3 = P(C_H = 3)$, sendo $P_0 + P_1 + P_2 + P_3 = 1$.

Para $\alpha = 0$, em relação à acuracidade, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve maior quantidade e a euclidiana a menor. Por outro lado, em relação ao tempo, a medida mista ponderada padrão apresentou melhores resultados, com menores tempos e dispersão nos tempos de execução do algoritmo.

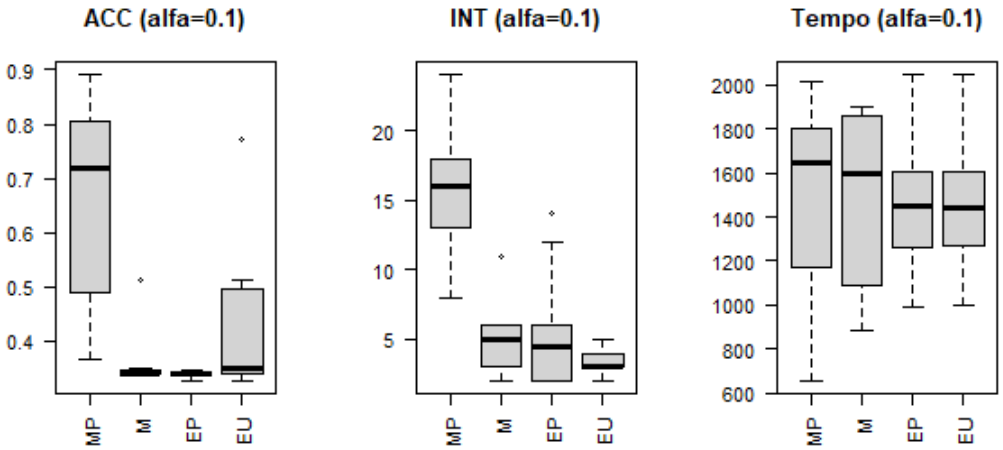
Figura 4.16: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0$ (Experimento 4).



Fonte: Próprio autor.

Para $\alpha = 0,1$, em relação à acuracidade, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve maior quantidade e a euclidiana a menor. Por outro lado, em relação ao tempo, as medidas apresentaram tempos medianos próximos.

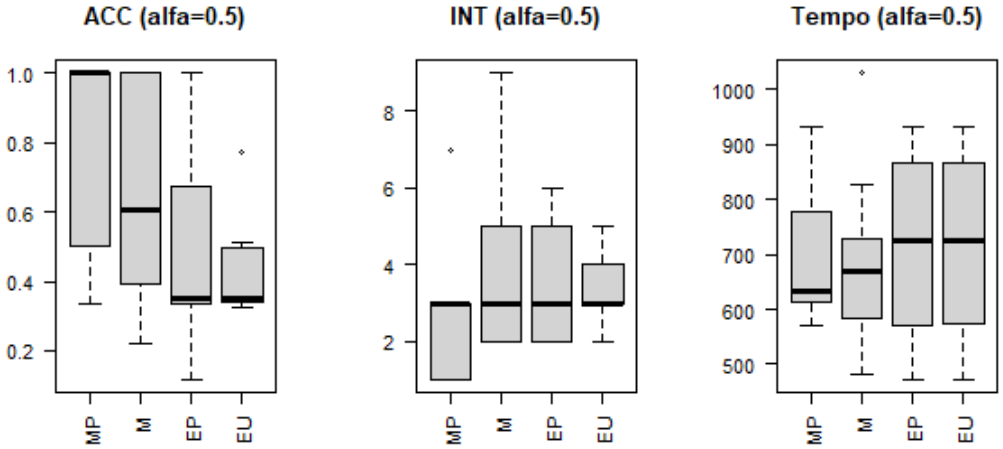
Figura 4.17: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,1$ (Experimento 4).



Fonte: Próprio autor.

Para $\alpha = 0.5$, em relação à acuracidade, a distância mista multinível proposta apresentou melhores valores de ACC, apesar da grande variação. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, a medida mista multinível proposta apresentou melhor resultado, com menor variação no tempo de execução.

Figura 4.18: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,5$ (Experimento 4).

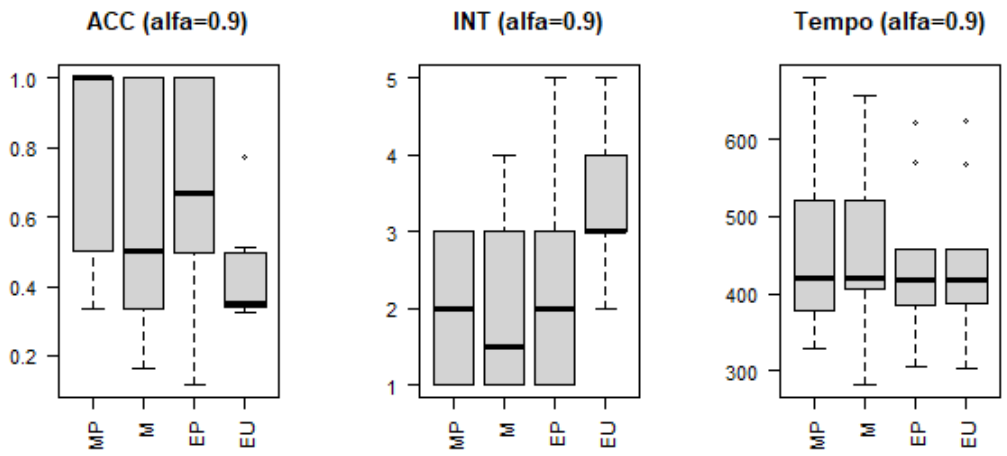


Fonte: Próprio autor.

Para $\alpha = 0.9$, a distância mista multinível proposta apresentou maiores valores

de ACC. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve menor quantidade, apesar da dispersão. Por outro lado, em relação ao tempo, a medida mista padrão apresentou melhores resultados, com menores tempos.

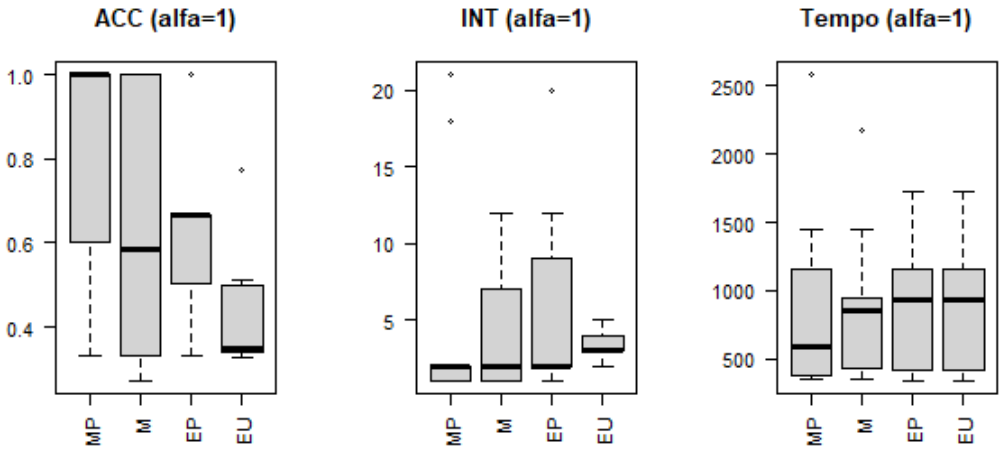
Figura 4.19: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 0,9$ (Experimento 4).



Fonte: Próprio autor.

Para $\alpha = 1$, em relação à acuracidade, a distância mista multinível proposta apresentou maiores valores de ACC. Já em relação a quantidade de interações realizadas, a medida mista multinível proposta foi a que obteve menor quantidade. Por outro lado, em relação ao tempo, a medida mista multinível proposta apresentou melhores resultados, com menores tempos nos tempos de execução do algoritmo.

Figura 4.20: Acurácia, quantidade de interações e tempo de execução do algoritmo para diferentes medidas de distância e $\alpha = 1$ (Experimento 4).



Fonte: Próprio autor.

Para o experimento 4, os maiores valores de acurácia foram obtidos com a distância mista multinível proposta. Já para a quantidade de interações, as distâncias mista ponderada e euclidiana padrão apresentaram menores valores. Em relação a menor tempo de execução, não se teve uma predominância de alguma das distâncias.

4.3 Comentários gerais

De forma geral nos experimentos, a ACC independe da quantidade de classes de Hellinger. A medida mista proposta apresentou melhores valores, exceto quando se teve apenas uma variável numérica. Para esse caso e $\alpha \leq 0,5$ a medida euclidiana apresentou melhores resultados.

Em relação ao tempo de execução, para $\alpha = 0$, as medidas apresentaram comportamentos muito semelhantes. Para o caso com poucas classes de Hellinger e muitas variáveis numéricas a medida mista proposta apresentou melhores tempos. Já com apenas uma variável numérica a medida depende do valor de α. Para $\alpha \geq 0,5$ a medida mista proposta teve menores valores e para $\alpha < 0,5$ a medida euclidiana. Para os casos com muitas classes de Hellinger o padrão observado é trocado. Para uma variável numérica a medida mista proposta apresentou melhores resultados e com muitas variáveis numéricas oscila entre as medidas euclidiana e mista proposta.

Quanto a quantidade de interações, para poucas classes de Hellinger e variáveis numéricas a medida euclidiana apresentou menores quantidades, e para muitas classes de Hellinger e variáveis numéricas tanto a medida euclidiana quanto a mista proposta

apresentaram menores valores.

Capítulo 5

Aplicação

Neste capítulo são apresentadas duas aplicações para ilustrar o método proposto na Seção 3.4. Para cada aplicação tem-se informações sobre o contexto, os dados, resultados e conclusões das respectivas análises. Assim como no Capítulo 4, toda a análise foi realizada no *software* R versão 4.2.1 (R Core Team, 2022).

5.1 Aplicação 1: Social

Esta aplicação foi construída a partir de uma colaboração com o Centro de Integração de Dados e Conhecimentos para Saúde (Cidacs/Fiocruz Bahia), o qual tinha um projeto em que um dos objetivos era identificar os diferentes perfis de pobreza das famílias com crianças menores de cinco anos inscritas no Cadastro Único (CadÚnico) no Brasil. Nessa aplicação considerou-se apenas 2 níveis: famílias (1º nível) e municípios (2º nível).

Primeiramente duas bases de dados foram concedidas pela instituição, uma contendo indicadores municipais e outra referente ao agrupamento realizado no nível 1 (famílias), tendo-se o número total de famílias de cada grupo por município. Desse modo, obteve-se a proporção de cada grupo do nível 1 para cada município e apenas os indicadores municipais referentes ao ano de 2010 para a montagem da base de dados de interesse.

Seguidamente, após juntar as duas bases de dados considerando como variável-chave o código do município de acordo com o IBGE (Instituto Brasileiro de Geografia e Estatística), fez-se a seleção de variáveis para cada grupo do nível 1 através do método de GENUER R.; POGGI (2010). Esse método tem como base a importância da variável de modelos *Random Forest*. A lista completa das variáveis municipais utilizadas pode ser vista no Apêndice.

Dessa maneira, as variáveis mais importantes foram:

- *coverage_bf_mun*: Cobertura do Programa Bolsa Família em relação à população total;

- *sewage*: % da população com esgotamento sanitário inadequado;
- *txurb*: Taxa de urbanização;
- *extr_pobres_bf*: Taxa de extrema pobreza (linhas do Bolsa Família);
- *water*: % da população com água encanada.

Seguidamente realizou-se o agrupamento do nível 2 (municípios) utilizando a metodologia de agrupamento multinível proposta e considerando $\alpha = 0,9$, o que apresentou melhores resultados das métricas de avaliação de agrupamentos. A Figura 5.1 apresenta os padrões das variáveis em cada grupo obtido e de forma geral. Observa-se que o grupo 1 apresenta alta população com esgotamento sanitário inadequado, cobertura do PBF e taxa de extrema pobreza, além de baixa taxa de urbanização e população com água encanada. O grupo 2 apresenta padrão similar, porém com magnitude menor. Já o grupo 3, apresenta baixa população com esgotamento sanitário inadequado, cobertura do PBF e extrema pobreza e uma maior taxa de urbanização e de população com água encanada. Em relação ao grupo 4, tem-se certa regularidade quanto aos indicadores e o grupo 5 apresentou uma baixa população com esgotamento sanitário inadequado, cobertura PBF, extrema pobreza, urbanização e população com água encanada.

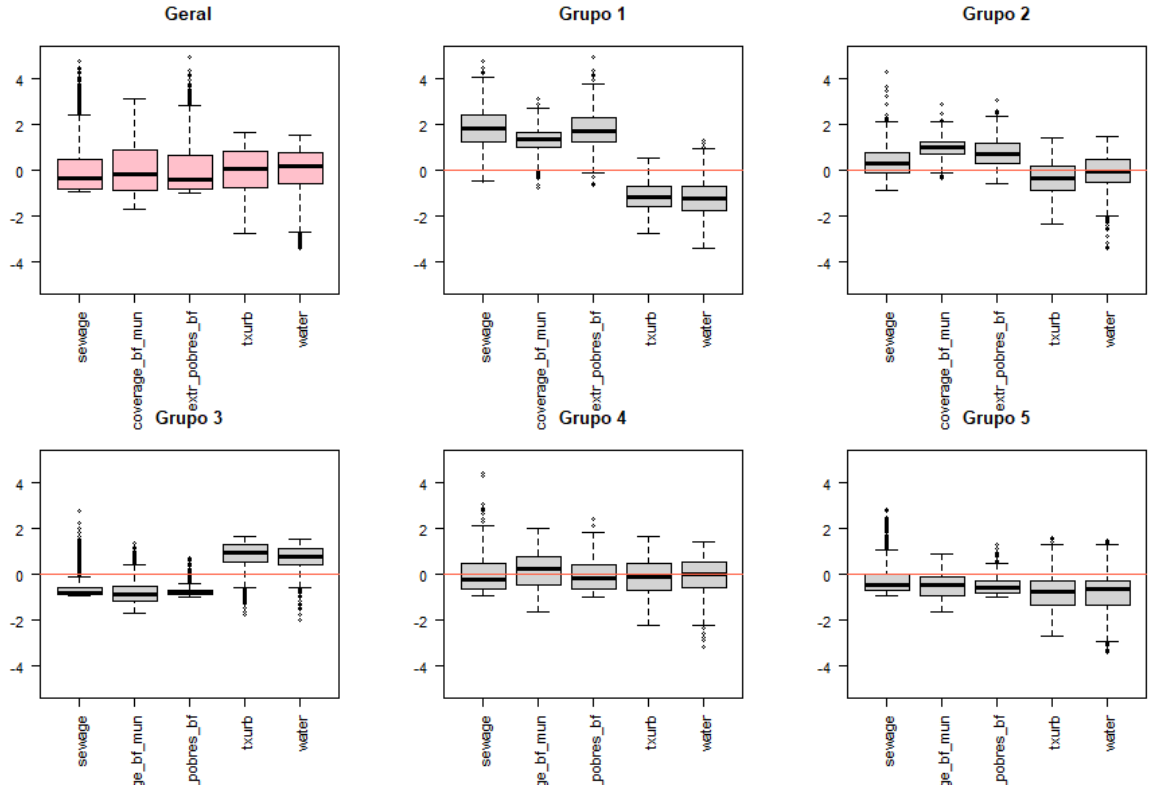


Figura 5.1: Comportamento das variáveis em cada grupo.

Posteriormente foi elaborado o mapa com o agrupamento obtido para o nível 2. Por meio da Figura 5.2 percebe-se que na região do Centro-Oeste há predomínio dos grupos 3 e 5. Já o Nordeste apresenta a predominância dos grupos 1, 2 e 3, tendo uma maior concentração do grupo 1 mais ao noroeste da região e dos grupos 2 e 3 nas demais áreas. Em relação ao Norte percebe-se padrões bem distintos. O estado de Rôndonia engloba os grupos 4 e 5, enquanto que os estados do Acre, Amazonas e Roraima possuem os grupos 1 e 2. Além disso, o Pará apresenta os grupos 1 e 5, e nos estados do Amapá e Tocantins os grupos estão distribuídos de forma mais regular. Sobre o Sudeste, o grupo 3 tem predominância. No entanto, o norte do estado de Minas Gerais detém mais o grupo 2 e no Espírito Santo o grupo 5 é o que mais aparece. Quanto ao Sul, apresenta padrão mais estável com os grupos 3 e 5. Ainda, nota-se que o grupo 4 tem presença em todo o território nacional e que o grupo 5 tanto ao Sul quanto ao Centro-Oeste e Norte do país.

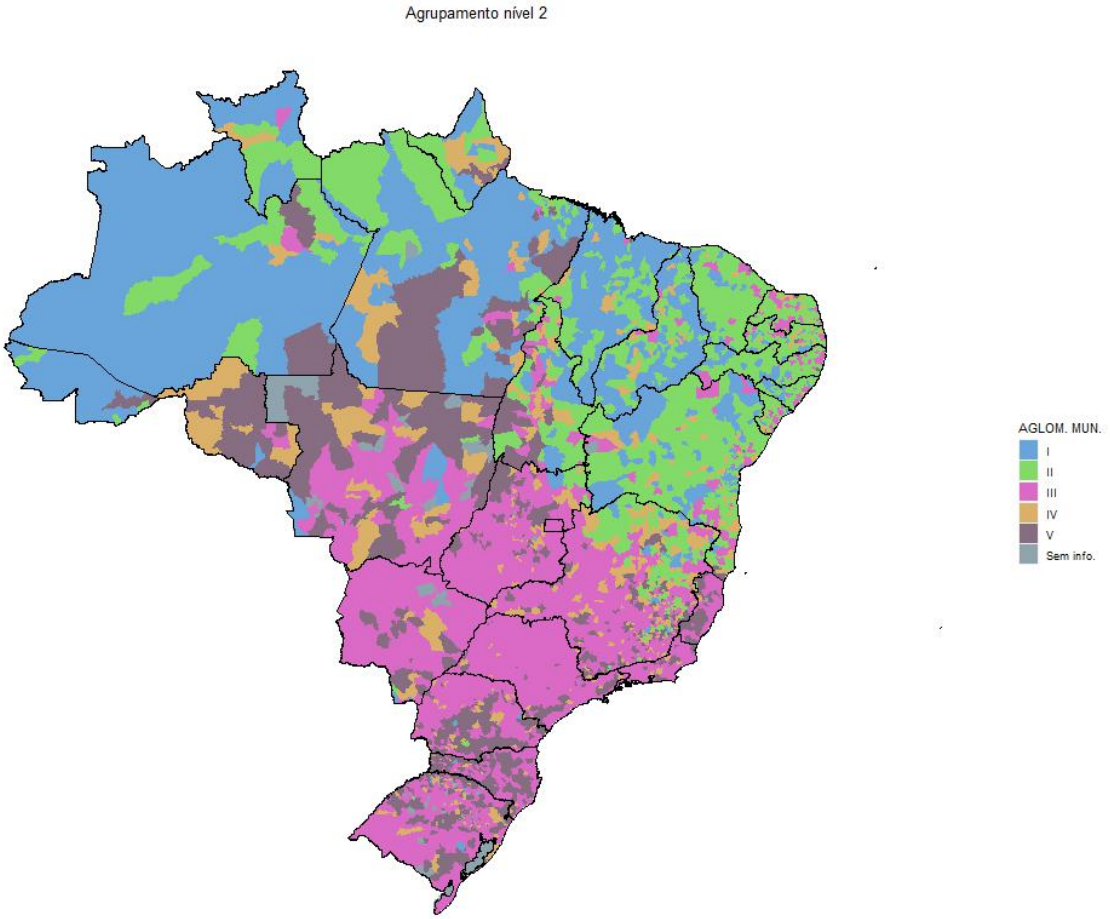


Figura 5.2: Mapa do agrupamento multinível

Por meio da Tabela 5.1, temos que:

- Grupo 1: Acentuada cobertura do PBF, não urbanização, não recursos hídricos e pobreza extrema;

- Grupo 2: Alta cobertura do PBF, pouca urbanização, pouco recursos hídricos e alta pobreza extrema;
- Grupo 3: Baixa cobertura do PBF, maior urbanização, mais recursos hídricos e baixa pobreza extrema;
- Grupo 4: Moderada cobertura do PBF, urbanização, recursos hídricos e pobreza extrema;
- Grupo 5: Baixa cobertura do PBF, moderada urbanização, recursos hídricos e baixa pobreza extrema.

Grupo	n	Cobertura PBF	Esg. san. inadequado	Urbanização	Extrema pobreza	Água encanada
1	615	**57,84**	**43,02**	**38,91**	**32,41**	**43,12**
2	1.305	51,29	20,04	56,61	20,46	66,66
3	2.228	**17,67**	**4,40**	**82,81**	**2,58**	**83,36**
4	483	35,83	14,27	61,56	11,13	67,13
5	876	22,85	11,13	46,75	5,62	52,40
Brasil	5565	32,54	14,35	64,10	11,38	68,56

Tabela 5.1: Variáveis pelos grupso

Dessa maneira, tem-se que o grupo 1 concentra os piores cenários dos indicadores sociais, o grupo 3 os melhores cenários e o grupo 4 o cenário intermediário, ou seja, apresentando indicadores com melhores e piores cenários.

5.1.1 Antes de 2010

Dividindo a análise por meio do ano de 2010, temos que a Figura 5.3 apresenta os padrões das variáveis em cada grupo obtido e de forma geral antes de 2010. Observa-se que o grupo 1 apresenta alta população com esgotamento sanitário inadequado, cobertura do PBF e taxa de extrema pobreza, além de baixa taxa de urbanização e população com água encanada. O grupo 4 apresenta padrão similar, porém com magnitude menor. Já o grupo 3, apresenta baixa população com esgotamento sanitário inadequado, cobertura do PBF e extrema pobreza e uma maior taxa de urbanização e de população com água encanada. Em relação ao grupo 2, tem-se um leve aumento da cobertura do PBF e taxa de extrema pobreza e população com água encanada e certa regularidade quanto a população com esgotamento sanitário inadequado e taxa de urbanização e o grupo 5 apresentou uma baixa população com esgotamento sanitário inadequado, cobertura PBF, extrema pobreza, urbanização e população com água encanada.

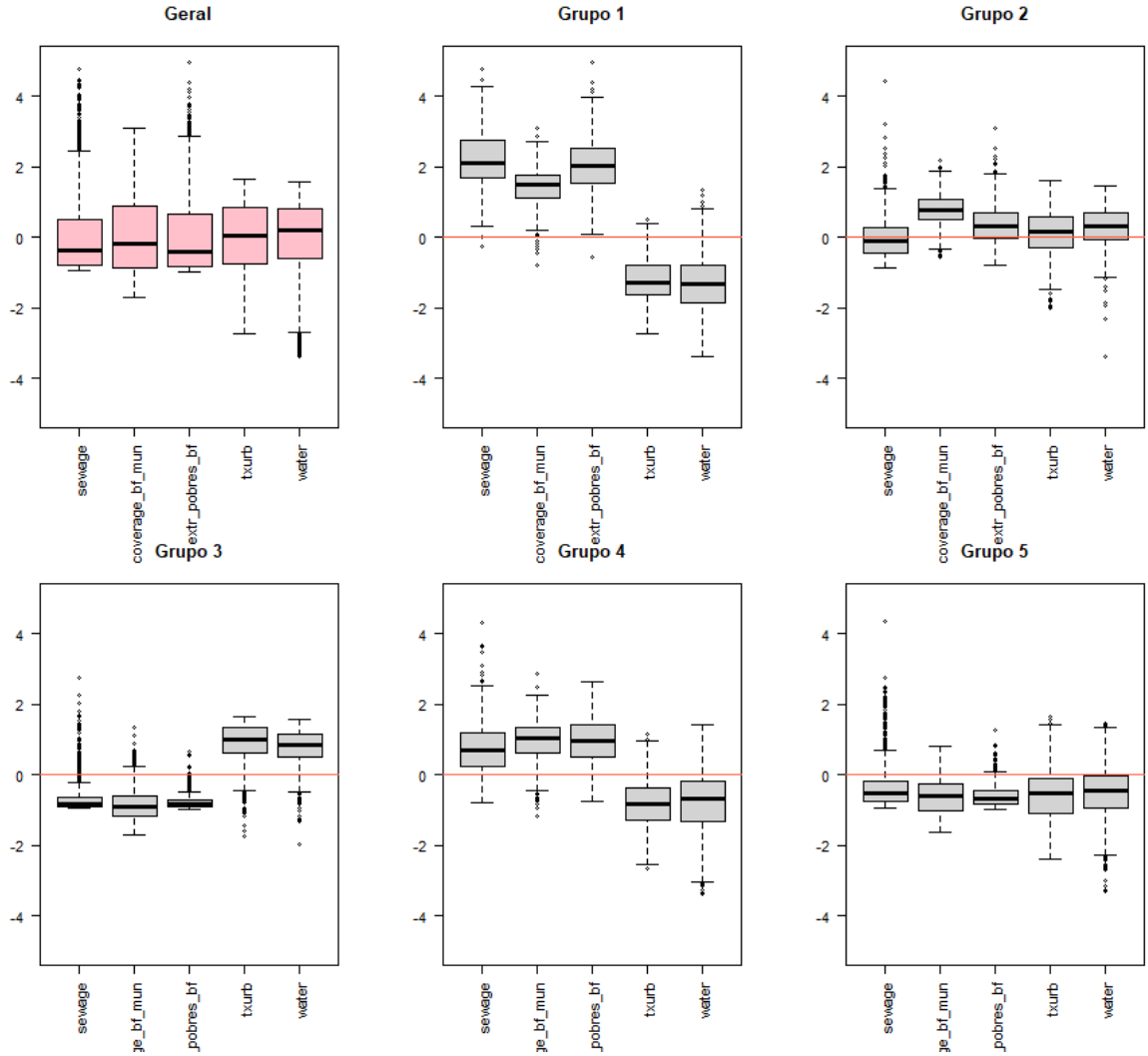


Figura 5.3: Comportamento das variáveis de forma geral e em cada grupo (antes de 2010).

Posteriormente foi elaborado o mapa com o agrupamento obtido para o nível 2. Por meio da Figura 5.4 percebe-se que na região do Centro-Oeste há predomínio dos grupos 3 e 5. Já o Nordeste apresenta a predominância dos grupos 1, 2 e 4, tendo uma maior concentração do grupo 1 mais ao noroeste da região e dos grupos 2 e 4 nas demais áreas. Em relação ao Norte percebe-se padrões bem distintos. O estado de Rondônia engloba os grupos 4 e 5, enquanto que os estados do Acre, Amazonas e Roraima possuem os grupos 1 e 4. Além disso, o Pará apresenta os grupos 1 e 4, e nos estados do Amapá e Tocantins os grupos estão distribuídos de forma mais regular. Sobre o Sudeste, o grupo 3 tem predominância. No entanto, o norte do estado de Minas Gerais detém mais o grupo 2. Quanto ao Sul, apresenta padrão mais estável com os grupos 3 e 5, e o grupo 4 no Paraná. Ainda, nota-se que o grupo 1 aparece mais ao Norte do país, o grupo 2 no Nordeste, o grupo 3 no Centro-Oeste, Sudeste e Sul, o grupo 4 no Norte, Nordeste e Sul e o grupo 5 no Centro-Oeste e Sul.

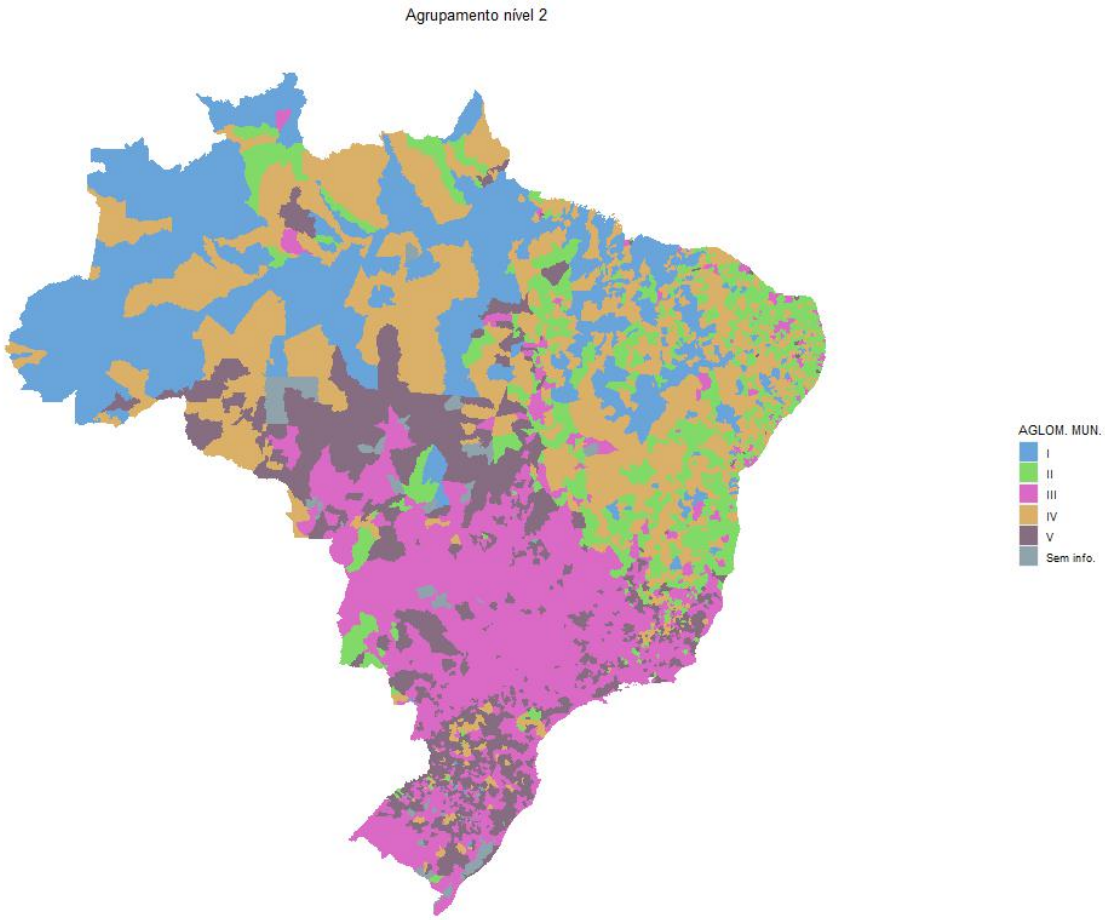


Figura 5.4: Mapa do agrupamento multinível (antes de 2010).

Por meio da Tabela 5.2, temos que:

- Grupo 1: Acentuada cobertura do PBF, não urbanização, não recursos hídricos e alta pobreza extrema;
- Grupo 2: Alta cobertura do PBF, urbanização, recursos hídricos e pobreza extrema;
- Grupo 3: Baixa cobertura do PBF, maior urbanização, mais recursos hídricos e baixa pobreza extrema;
- Grupo 4: Alta cobertura do PBF, urbanização, pouco recursos hídricos e pobreza extrema.
- Grupo 5: Baixa cobertura do PBF, moderada urbanização, recursos hídricos e e baixa pobreza extrema.

Grupo	n	Cobertura PBF	Esg. san. inadequado	Urbanização	Extrema pobreza	Água encanada
1	412	**59,48**	**48,27**	**37,22**	**35,30**	**41,30**
2	1008	47,20	14,23	66,65	15,71	74,73
3	2026	**16,44**	**4,12**	**84,33**	**2,14**	**84,28**
4	1049	*50,53*	*25,60*	*45,50*	*22,45*	*52,84*
5	1012	20,55	9,50	51,28	4,36	58,33
Brasil	5565	32,54	14,35	64,10	11,38	68,56

Tabela 5.2: Variáveis pelos grupos pré 2010

Dessa maneira, tem-se que o grupo 1 possui os piores cenários dos indicadores sociais, o grupo 3 os melhores cenários e o grupo 2 o cenário intermediário, ou seja, apresentando indicadores com melhores e piores cenários.

5.1.2 Depois de 2010

Temos que a Figura 5.3 apresenta os padrões das variáveis em cada grupo obtido e de forma geral depois de 2010. Observa-se que o grupo 1 apresenta alta população com esgotamento sanitário inadequado, cobertura do PBF e taxa de extrema pobreza, além de baixa taxa de urbanização e população com água encanada. O grupo 3 apresenta padrão similar, porém com magnitude menor. Já o grupo 4, apresenta baixa população com esgotamento sanitário inadequado, cobertura do PBF e extrema pobreza e uma maior taxa de urbanização e de população com água encanada. Em relação ao grupo 5, tem-se certa regularidade quanto aos indicadores e o grupo 2 apresentou uma baixa população com esgotamento sanitário inadequado, cobertura PBF, extrema pobreza, urbanização e população com água encanada.

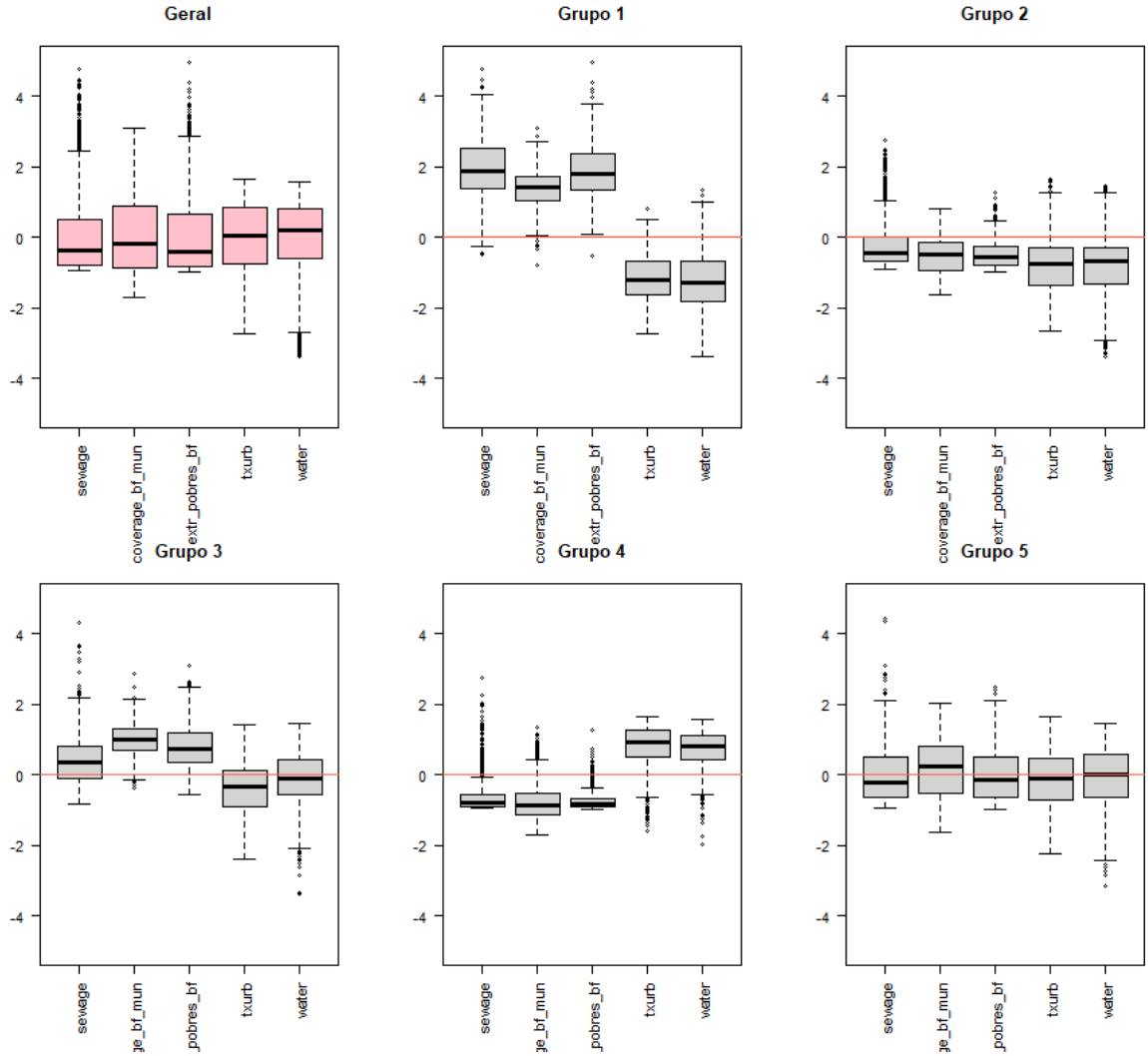


Figura 5.5: Comportamento das variáveis de forma geral e em cada grupo (depois de 2010).

Posteriormente foi elaborado o mapa com o agrupamento obtido para o nível 2. Por meio da Figura 5.6 percebe-se que na região do Centro-Oeste há predomínio dos grupos 2 e 4. Já o Nordeste apresenta a predominância dos grupos 1 e 3, tendo uma maior concentração do grupo 1 mais ao noroeste da região e do grupos 3 nas demais áreas. Em relação ao Norte percebe-se padrões bem distintos. O estado de Rondônia engloba os grupos 2 e 5, enquanto que os estados do Acre, Amazonas e Roraima possuem os grupos 1 e 4. Além disso, o Pará apresenta os grupos 1 e 3, e nos estados do Amapá e Tocantins os grupos estão distribuídos de forma mais regular. Sobre o Sudeste, o grupo 4 tem predominância. No entanto, o norte do estado de Minas Gerais detém mais o grupo 3. Quanto ao Sul, apresenta padrão mais estável com os grupos 2 e 4. Ainda, nota-se que o grupo 1 aparece mais ao norte do país, o grupo 2 no Centro-Oeste e Sul, o grupo 3 no nordeste, o grupo 4 no Sudeste e Sul e o grupo 5 nas diversas regiões.

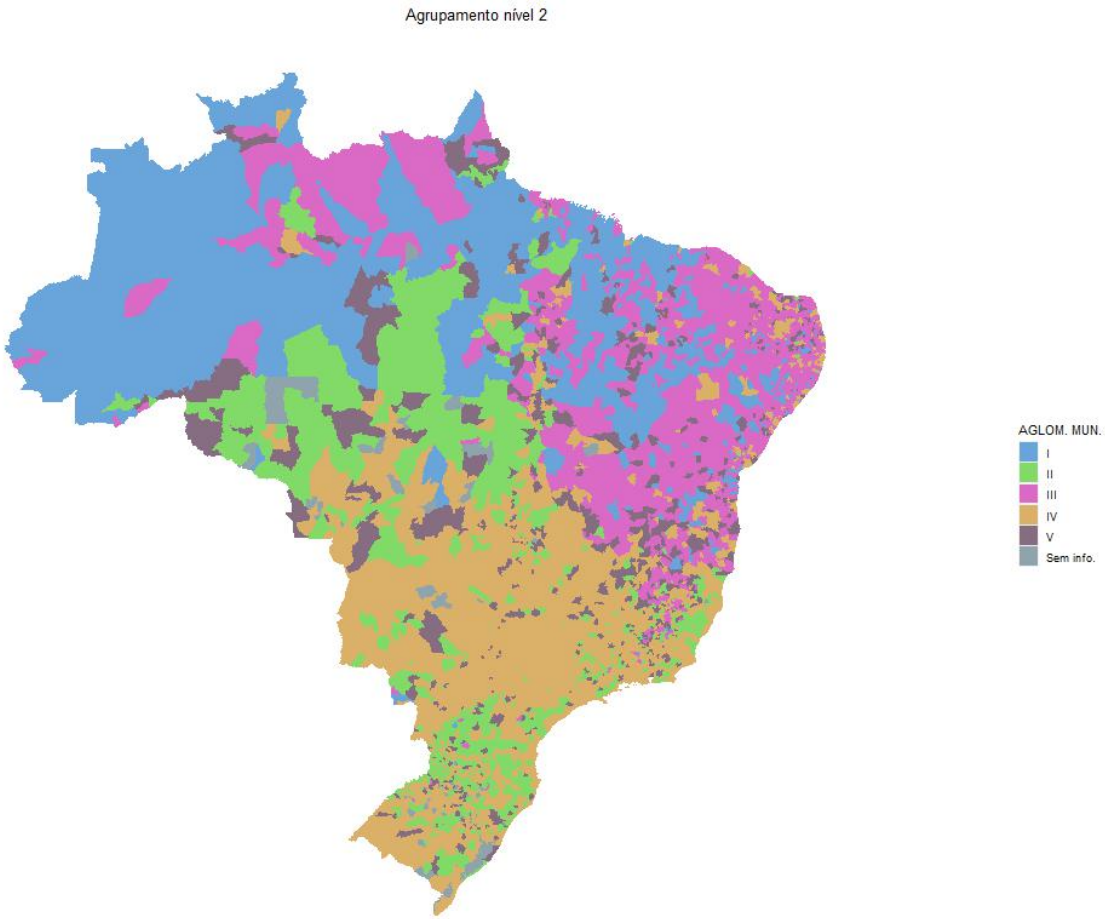


Figura 5.6: Mapa do agrupamento multinível (depois de 2010).

Por meio da Tabela 5.3, temos que:

- Grupo 1: Acentuada cobertura do PBF, não urbanização, não recursos hídricos e alta pobreza extrema;
- Grupo 2: Baixa cobertura do PBF, moderada urbanização, recursos hídricos e baixa pobreza extrema;
- Grupo 3: Alta cobertura do PBF, urbanização, recursos hídricos e pobreza extrema;
- Grupo 4: Baixa cobertura do PBF, maior urbanização, mais recursos hídricos e baixa pobreza extrema;
- Grupo 5: Baixa cobertura do PBF, maior urbanização, recursos hídricos, e pobreza extrema.

Grupo	n	Cobertura PBF	Esg. san. inadequado	Urbanização	Extrema pobreza	Água encanada
1	549	**58,28**	**44,33**	**38,88**	**33,11**	**42,71**
2	844	22,64	11,06	46,75	5,59	51,77
3	1344	51,38	20,66	55,78	20,72	65,83
4	2231	**17,80**	**4,41**	**82,60**	**2,60**	**83,40**
5	538	35,86	14,43	61,15	11,36	66,64
Brasil	5565	32,54	14,36	64,10	11,38	68,57

Tabela 5.3: Variáveis pelos grupos pós 2010

Dessa maneira, tem-se que o grupo 1 possui os piores cenários dos indicadores sociais, o grupo 4 os melhores cenários e o grupo 5 o cenário intermediário, ou seja, apresentando indicadores com melhores e piores cenários.

5.2 Aplicação 2: Tipificação dos cursos de graduação com enfoque para a área de Ciência de dados

Na área da Educação, um encadeamento natural pode ser visto como Alunos → Turmas → Disciplinas → Cursos → Áreas do conhecimento → IES (Instituições de Ensino Superior) → Organização. Atualmente temos uma crescente na área nomeada de Ciência de dados, onde é comum encontrar profissionais com formação nas áreas de Estatística e Ciência da computação. Cursos com enfoque para essa grande área tem sido criados recentemente, conforme pode ser visto em ARA et al. que estudam os primeiros cursos de graduação em universidades brasileiras focados para Ciência de dados. Nesse contexto, nosso interesse é tipificar os cursos de graduação com enfoque para a área de Ciência de dados, afim de entender o que os diferenciam.

Os dados utilizados nessa aplicação são provenientes dos microdados do Censo da Educação Superior (CES) do ano de 2022. O CES é uma pesquisa estatística sobre instituições de ensino, cursos, alunos e professores do ensino superior que tem como objetivo avaliar a situação da educação superior no Brasil. É realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) anualmente e pode ser acessado em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior.

O conjunto de dados utilizado foi o MICRODADOS_CADASTRO_CURSOS_2022 referente aos cadastros dos cursos de graduação. Os cursos com enfoque para a área de análise de dados foram obtidos através das palavras-chaves: dado, dados, *data*, estatistica, ciência e artificial, obtendo-se 24 cursos que foram: Estatística, Estatística e Ciência de dados, Matemática aplicada e computacional com habilitação em Estatística econômica, Ciência de dados, Ciência dos dados, Ciências de dados, *Data science*, *Data science analytics*, *Marketing* digital e *Data science*, Ciências de dados e Análise de comportamento, Ciência de dados e Inteligência artificial, Ciência de dados e Big data, Ciência

de dados e *Machine learning*, Ciência de dados para negócios, Ciência de dados e Inteligência analítica, Ciência da computação, Análise de dados, Inteligência de mercado e Análise de dados, *Big data* e Inteligência analítica, *Big data* e Inteligência artificial, *Big data* no agronegócio, *Big data* para negócios, Inteligência artificial, e Inteligência artificial aplicada.

No que se refere as variáveis, 6 (seis) foram utilizadas sendo elas:

- Referentes ao primeiro nível (alunos):
  - TX_CONCLUSAO_BRA: Taxa de conclusão de curso de brancos obtida por meio do quociente entre a quantidade de concluintes de cor branca e a quantidade de ingressantes de cor branca;
  - TX_CONCLUSAO_PP: Taxa de conclusão de curso de pretos e pardos obtida por meio do quociente entre a quantidade de concluintes de cor preta e parda e a quantidade de ingressantes de cor preta e parda.
- Referentes ao segundo nível (cursos):
  - QTD_LOCAIS: Quantidade de locais de oferta do curso obtida por meio da contagem de aparecimento do código do curso;
  - REL_CAND_VAGA: Relação candidato-vaga obtida por meio do quociente entre a quantidade total de inscritos no curso e a quantidade total de vagas;
  - TP_REDE: Tipo de rede de ensino (pública, privada);
  - TP_MODALIDADE_ENSINO: Tipo da modalidade de ensino do curso (presencial, curso a distância).

Tendo em vista que existem muitos cursos de Ciência da computação, o que pode influenciar a análise de agrupamento dominando-a, considerou-se duas estratégias para analisar os dados: uma considerando tal curso e outra sem considerá-lo. Tais análises são apresentadas a seguir.

Com Ciência da computação

Para esta situação, tem-se que $n = 586$ observações e realizou-se o agrupamento multinível dos dados considerando diferentes valores de α e k, afim de identificar a combinação que produz os melhores resultados. A combinação ótima foi obtida com $\alpha = 0,5$ e $k = 7$, e a Figura 5.7 apresenta o resultado das medidas de avaliação de agrupamentos. Nota-se que para a medida WSS existe uma maior redução para $k = 7$ e, também, maior valor para DU nessa mesma quantidade de grupos. Ainda, para a silhueta, verifica-se indícios de sobreposição nos dados.

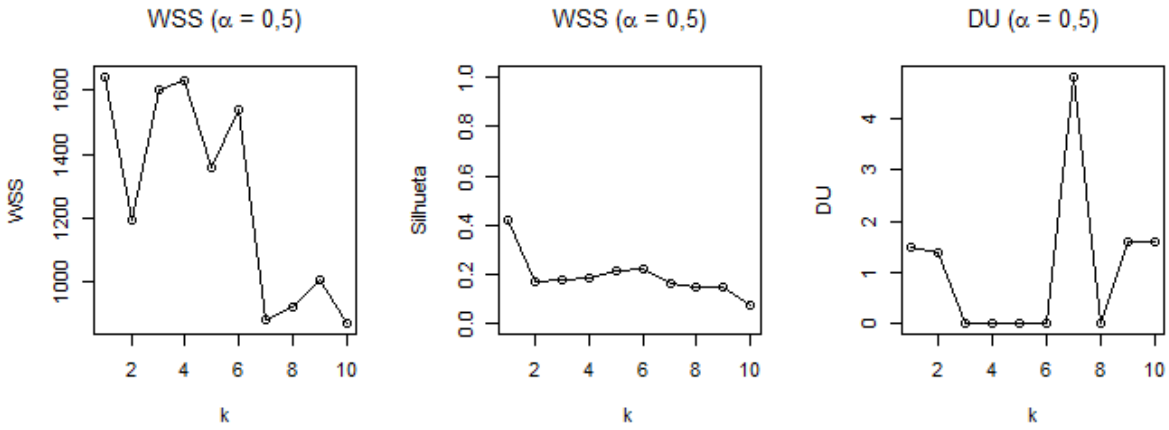


Figura 5.7: Medidas de avaliação de agrupamentos da aplicação 2 (com Ciência da computação).

A Figura 5.8 apresenta os padrões das variáveis quantitativas de forma geral e em cada grupo obtido. Observa-se que o grupo 1 apresenta baixíssimas quantidades de locais de oferta do curso e relação candidato-vaga moderada. Por outro lado, o grupo 3 possui cursos com acentuada quantidade de locais de oferta e relações candidato-vaga baixíssimas. Os grupos 5 e 6 apresentam o mesmo padrão de comportamento do grupo 3, porém com proporções consideravelmente mais reduzidas. Já o grupo 7 apresenta alta variabilidade quanto a relação candidato-vaga e baixa quantidade de locais de oferta, apresentando um valor destoante dos demais, e o grupo 2 apresenta mesmo padrão de comportamento, porém com menores proporções. E, o grupo 4, apresenta baixíssima relação candidato-vaga e baixas quantidades de locais de oferta dos cursos.

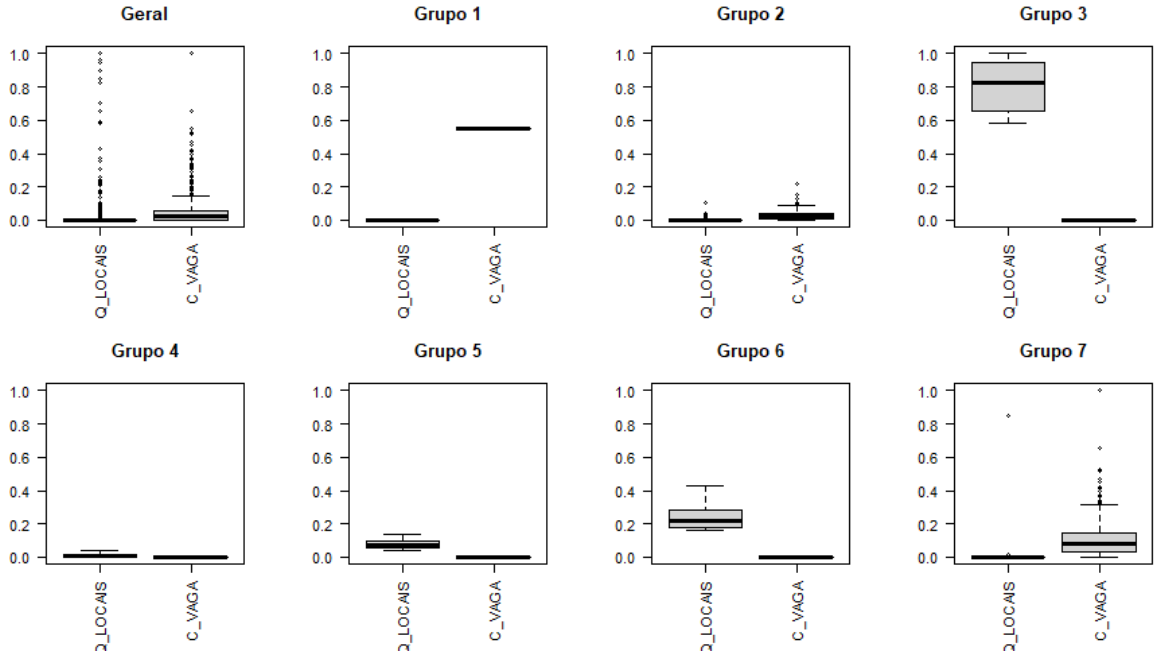


Figura 5.8: Comportamento das variáveis quantitativas de forma geral e em cada grupo da aplicação 2 (com Ciência da computação).

A Tabela 5.4 sumariza estatísticas descritivas sobre as variáveis para cada grupo. A partir dessas estatísticas pode-se notar que no grupo 1 tem-se apenas um curso, apresentando mais candidatos do que vaga. Já o grupo 3 possui os cursos com maiores quantidades de locais de oferta e, também, muito mais vagas do que candidatos. Os grupos 5 e 6 apresentam o mesmo padrão que o grupo 3, no entanto se diferenciam entre si pela quantidade de locais de oferta, pois o grupo 6 tem cursos com maiores quantidades de locais do que o grupo 5. O grupo 7 apresenta alta variação na quantidade de locais de oferta e o grupo 2 apresenta o mesmo padrão, mas com menores proporções. E, o grupo 4, apresenta as menores relações candidato-vaga e quantidade de locais de oferta.

Grupo	n	Gratuito	Rede	Modalidade	Qtd. locais oferta	Relação cand.-vaga
1	1	Não (100%)	Privada (100%)	Presencial (100%)	1; 0 [1; 1]	26,5; 0 [26,5; 26,5]
2	289	Não (289)	Privada (289)	Presencial (272)	1,62; 3,38 [1; 43]	1,31; 1,31 [0; 10,69]
3	9	Não (9)	Privada (9)	A distância (9)	328,67; 68,07 [240; 413]	0; 0 [0; 0]
4	56	Não (56)	Privada (56)	A distância (56)	5,02; 4,33 [1; 17]	0; 0 [0; 0]
5	20	Não (20)	Privada (20)	A distância (20)	32,15; 10,25 [19; 57]	0; 0 [0; 0]
6	20	Não (20)	Privada (20)	A distância (20)	98; 32,32 [68; 177]	0; 0 [0; 0]
7	191	Sim (184)	Pública (191)	Presencial (186)	2,93; 25,40 [1; 352]	5,92; 6,47 [0; 48,53]

Tabela 5.4: Estatísticas descritivas das variáveis da aplicação 2 (com Ciência da computação).

Por meio desses resultados nós podemos depreender que:

- Grupo 1: Apenas o curso de Ciência da computação, presencial, não gratuito, da rede privada e ofertado em Brasília (DF) com 1.166 inscritos para 44 vagas;

- Grupo 2: Cursos com poucos locais de oferta, relação candidato-vaga moderada, não gratuitos, da rede privada e maioria presencial. Engloba, principalmente, cursos de Ciência da computação (85%) de diversos locais como, por exemplo, Recife (PE), Passo Fundo (RS) e Ji-Paraná (RO);

- Grupo 3: Cursos com abundância nos locais de oferta, muito mais vagas do que candidatos, não gratuitos, da rede privada e a distância. Engloba, principalmente, cursos de Ciência da computação (33,33%) e Ciência de dados (22,22%);

- Grupo 4: Cursos com poucos locais de oferta, mais vagas do que candidatos, não gratuitos, da rede privada e a distância. Engloba, principalmente, cursos de Ciência da computação (23,21%), Ciência de dados (16,07%), Inteligência artificial (16,07%) e Big data e Inteligência analítica (14,29%);

- Grupo 5: Cursos com razoáveis locais de oferta, mais vagas do que candidatos, não gratuitos, da rede privada e a distância. Engloba, principalmente, cursos de Ciência da computação (45%), Ciência de dados (20%) e Ciências de dados (10%).

- Grupo 6: Cursos com mais locais de oferta do que o grupo 5 e menos que o grupo 3, mais vagas do que candidatos, não gratuitos, da rede privada e a distância. Engloba, principalmente, cursos de Ciência de dados (35%), Ciência da computação (25%), Big data e Inteligência analítica (10%) e Estatística (10%);

- Grupo 7: Cursos com poucos locais de oferta com valor atípico para o curso de Ciência de dados, relação candidato-vaga vasta, maioria gratuitos, da rede pública e predominantemente presencial. Engloba, principalmente, cursos de Ciência da computação (74,87%) e Estatística (16,75%).

Sem Ciência da Computação

Para esta situação, tem-se que $n = 164$ observações e realizou-se o agrupamento multinível dos dados considerando diferentes valores de α e k, afim de identificar a combinação que produz os melhores resultados. A combinação ótima foi obtida com $\alpha = 0,5$ e $k = 5$, e a Figura 5.9 apresenta o resultado das medidas de avaliação de agrupamentos. Nota-se que para a medida WSS existe uma maior redução para $k = 5$ e, também, maior valor para DU nessa mesma quantidade de grupos. Ainda, para a silhueta, verifica-se indícios de sobreposição nos dados.

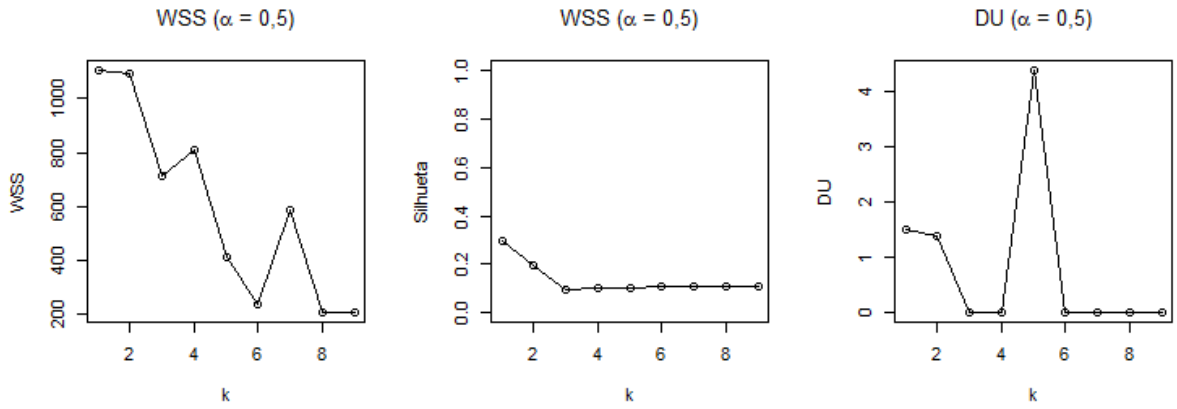


Figura 5.9: Medidas de avaliação de agrupamentos da aplicação 2 (sem Ciência da computação).

A Figura 5.8 apresenta os padrões das variáveis quantitativas de forma geral e em cada grupo obtido. Observa-se que o grupo 1 apresenta baixíssimas quantidades de locais de oferta do curso e relação candidato-vaga. Por outro lado, o grupo 4 possui cursos com acentuada quantidade de locais de oferta e relações candidato-vaga baixíssimas. Os grupos 3 e 5 apresentam padrão de comportamento similar, apresentando baixa relação candidato-vaga e quantidades de locais de oferta de cursos mais dispersa para o grupo 5. Já o grupo 2 apresenta baixíssimas quantidades de locais de oferta e baixa relação candidato vaga, apresentando alguns valores destoantes.

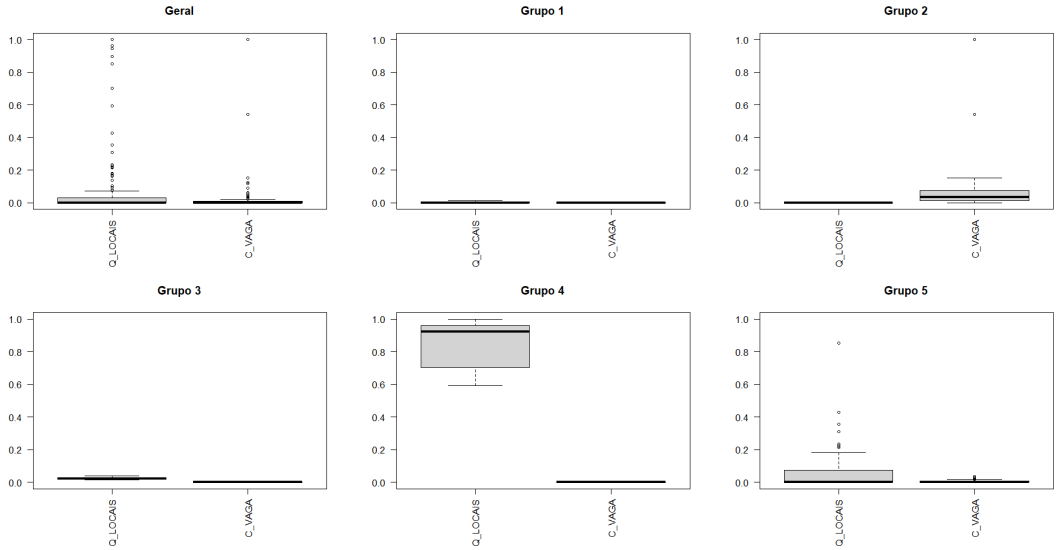


Figura 5.10: Comportamento das variáveis quantitativas de forma geral e em cada grupo da aplicação 2 (sem Ciência da Computação).

A Tabela 5.5 sumariza estatísticas descritivas sobre as variáveis para cada grupo. A partir dessas estatísticas pode-se notar que no grupo 1 tem-se muito mais vaga do que candidato e poucos locais de oferta dos cursos. Já o grupo 2 possui os cursos ofertados

apenas em um local e com relação candidato-vaga vasta. Os grupos 3 e 4 apresentam o mesmo padrão de comportamento entre sim, exceto pelos locais de oferta. O grupo 3 possui baixos locais de oferta enquanto o grupo 4 apresenta alta variação na quantidade de locais de oferta. E, o grupo 5, apresenta ampla quantidade de locais de oferta e mais vagas do que candidatos.

Grupo	n	Gratuito	Rede	Modalidade	Qtd. locais oferta	Relação cand.-vaga
1	27	Não (27)	Privada (27)	A distância (27)	2,37; 1,71 [1; 6]	0; 0 [0; 0]
2	24	Não (24)	Privada (24)	Presencial (24)	1; 0 [1; 1]	5,21; 10,98 [0; 50]
3	13	Não (13)	Privada (13)	A distância (13)	10,31; 2,93 [7; 17]	0; 0 [0; 0]
4	6	Não (6)	Privada (6)	A distância (6)	351; 67,64 [245; 413]	0; 0 [0; 0]
5	94	Sim (48)	Pública (48)	A distância (48)	25,23; 49,32 [1; 352]	0,23; 0,35 [0; 1,79]

Tabela 5.5: Estatísticas descritivas das variáveis da aplicação 2 (sem Ciência da Computação).

Por meio desses resultados nós podemos depreender que:

- Grupo 1: Cursos a distância, não gratuitos, da rede privada, muito mais vagas do que candidatos e pouquíssimos locais de oferta. Engloba, principalmente, cursos de Big data e Inteligência analítica (22,22%), Ciência de dados (22,22%) e Inteligência artificial (22%);

- Grupo 2: Cursos com apenas 1 local de oferta, relação candidato-vaga baixa, mas com alguns valores destoantes com mais candidatos do que vagas, não gratuitos, da rede privada e presencial. Engloba, principalmente, cursos de Ciência de dados e Inteligência artificial (33,33%) e Ciência de dados (20,83%);

- Grupo 3: Cursos com poucos locais de oferta, mais vagas do que candidatos, não gratuitos, da rede privada e a distância. Engloba, principalmente, Ciência de dados (23,08%), Estatística (23,08%) e Inteligência artificial (23,08%);

- Grupo 4: Cursos com alta variação na quantidade de locais de oferta, com mais vagas do que candidatos, não gratuitos, da rede privada e a distância. Engloba, principalmente, cursos de Ciência de dados (33,33%), Big data e Inteligência analítica (16,67%) e Ciências de dados (16,67%);

- Grupo 5: Cursos com razoáveis quantidades de locais de oferta, mais vagas do que candidatos e equilíbrio quanto a gratuidade, tipo de rede (privada e pública) e a modalidade (presencial e a distância). Engloba, principalmente, cursos de Estatística (37,23%), Ciência de dados (26,60%) e Big data e inteligência analítica (10,64%).

Capítulo 6

Considerações Finais

Nos capítulos iniciais desta dissertação são apresentados o contexto e o problema de pesquisa, discorrendo sobre a análise de agrupamento e a estrutura de dados multiníveis. Seguidamente tem-se uma revisão de literatura e a conceituação da abordagem proposta, bem como a verificação desta em dados simulados e reais.

Estudos de simulação foram realizados com o objetivo de avaliar o comportamento da abordagem proposta. As simulações apresentaram resultados satisfatórios, uma vez que a identificação do grupo correto ocorreu de acordo com o esperado. Tais estudos mostraram que a taxa de acerto da metodologia proposta independe da quantidade de classes referentes a medida de Hellinger, ou seja, as distribuições referentes ao nível anterior, e que a medida mista multinível proposta apresentou melhores valores quando se teve mais de uma variável quantitativa.

Para o estudo em dados reais foram utilizados dois bancos de dados para ilustrar a aplicabilidade do método, em que os resultados dos agrupamentos foram satisfatórios também. No entanto, vale ressaltar que é importante realizar mais estudos de simulação com diferentes valores de k e quantidades de variáveis qualitativas, ou seja, variando ainda mais os cenários. Neste trabalho nos restringimos a dois níveis de estrutura multinível e variáveis quantitativas binárias.

Como trabalho futuro, pretende-se realizar pesquisas mais extensas e aprofundadas na literatura a fim de confirmar a unicidade da proposta desenvolvida. Deseja-se ainda realizar mais estudos de simulação para estruturas diversas, com diferentes configurações de variáveis e valores de k. Apesar do índice silhueta não ter se mostrado tão útil, pretende-se explorar esta situação mais a fundo podendo, inclusive, surgir uma métrica de avaliação mais específica para o caso. Como outro tema de pesquisa futura sugere-se o estudo do tempo computacional e da complexidade algorítmica do método proposto.

Apêndice A

Tabela das notações utilizadas

Notação	Descrição
$\mathbf{X}$	Conjunto de dados
n	Quantidade de unidades de análises (observações)
p	Quantidade de variáveis
i	Índice para a linha $(1 \leq i \leq m \leq n)$
j	Índice para a coluna e $(1 \leq j \leq r \leq p)$
q	Quantidade de variáveis qualitativas
f	Quantidade de variáveis quantitativas
x_{ij}	i-ésima observação da j-ésima coluna
$\mathbf{x}_{i.}$	vetor da i-ésima observação contendo todas as variáveis
$\mathbf{x}_{.j}$	vetor da j-ésima variável contendo todas as observações
$\mathbf{D}$	Matriz de dissimilaridade
$\mathbf{S}$	Matriz de similaridade
$s(x_i, x_m) = s_{im}$	similaridade entre duas observações
$d(x_i, x_m) = d_{im}$	dissimilaridade entre duas observações
Σ	Matriz de variância-covariância
K	Número total de grupos
G_k	k-ésimo grupo
$\mathbf{c}_k$	centróide do k-ésimo grupo

Apêndice B

Dicionário de variáveis da aplicação 1

Variável	**Indicadores**
pop_child	População de crianças ($\leq$ 14 anos)
pobinfant05	% de crianças entre 0 e 4 anos que moram em domicílios pobres
pobinfant14	% de crianças entre 5 a 14 anos que moram em domicílios pobres
expobinfant05	% de crianças entre 0 e 4 anos que moram em domicílios extremamente pobres
expobinfant14	% de crianças entre 5 a 14 anos que moram em domicílios extremamente pobres
naomatricula05_14	% de crianças entre 5 a 14 anos que não frequenta escola
trabinf	Taxa de trabalho infantil
poptcu	População total TCU
fam_benef_bf	Número de beneficiários do Bolsa Família
txurb	Taxa de urbanização
water	% da população com água encanada
sewage	% da população com esgotamento sanitário inadequado
alfab	Taxa de Alfabetização
gini	Taxa de desigualdade (índice de GINI)
pobres_bf	Taxa de pobreza (linhas do Bolsa Família)
extr_pobres_bf	Taxa de extrema pobreza (linhas do Bolsa Família)
povgap_bf	Poverty gap
extpovgap_bf	Extreme poverty gap
income_nom_pc	Rendimento domiciliar per capita (Nominal)
income_real_pc	Rendimento domiciliar per capita (Real: Deflacionada)
coverage_bf_target	Cobertura do Programa Bolsa Família em relação à população elegível
coverage_bf_mun	Cobertura do Programa Bolsa Família em relação à população total
pib	Produto Interno Bruto (PIB) - valores correntes (R$ 1.000)
pibpc	Produto Interno Bruto (PIB) per capita - valores correntes (R$ 1.000)
deflator	Deflator do PIB
pib_d	Produto Interno Bruto (PIB) - valores constantes (ano base 2002)
pibpc_d	Produto Interno Bruto (PIB) per capita - valores constantes (ano base 2002)

Referências Bibliográficas

ALBUQUERQUE, M. A. et al. Comparação entre coeficientes similaridade um aplicação em ciências florestais. *Matemática e Estatística em Foco*, v. 4, n. 2, p. 102–114, 2016.

ALMEIDA, C.; QUITUISACA-SAMANIEGO, L.; ANTAMBA, L. Concentración/especialización económica en el ecuador. *Perfiles Económicos*, 2019.

AMORIM, L. B. de; CAVALCANTI, G. D.; CRUZ, R. M. The choice of scaling technique matters for classification performance. *Applied Soft Computing*, Elsevier, v. 133, p. 109924, 2023.

ANDRADE, J. M. d.; LAROS, J. A. Fatores associados ao desempenho escolar: estudo multinível com dados do saeb/2001. *Psicologia: teoria e pesquisa*, SciELO Brasil, v. 23, p. 33–41, 2007.

ARA, A. et al. Ciência de dados: uma descrição dos primeiros cursos de graduação em universidades brasileiras. SciELO Preprints.

BARBARA, D. An introduction to cluster analysis for data mining. *Retrieved November*, v. 12, p. 2003, 2000.

BARBOSA, L. d. S. As distâncias de manhattan e chebyshev na avaliação da acurácia posicional com feições lineares em produtos cartográficos. Universidade Federal de Viçosa, 2022.

BENIN, G. et al. Comparações entre medidas de dissimilaridade e estatísticas multivariadas como critérios no direcionamento de hibridações em aveia. *Ciência Rural*, SciELO Brasil, v. 33, n. 4, p. 657–662, 2003.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, Elsevier, v. 10, n. 2-3, p. 191–203, 1984.

BHATTACHARYYA, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, v. 35, p. 99–109, 1943.

BUSSAB, W.; MORETTIN, P. *Estatística Básica*. [S.l.]: Saraiva, 2017.

CASSETTI, J.; GAMBINI, J.; FRERY, A. Estimación de parámetros utilizando distancias estocásticas para datos con ruido speckle. In: *14th Argentine Symposium on Technology, AST*. [S.l.: s.n.], 2013.

CHEN, Z.-G. et al. An efficient privacy protection in mobility social network services with novel clustering-based anonymization. *EURASIP journal on Wireless communications and networking*, Springer, v. 2016, n. 1, p. 1–9, 2016.

CRISPIM, D. L.; FERNANDES, L. L.; ALBUQUERQUE, R. d. O. Aplicação de técnica estatística multivariada em indicadores de sustentabilidade nos municípios do marajó-pa. *Revista Principia-Divulgação Científica e Tecnológica do IFPB,(46)*, p. 145–154, 2019.

DAS, A. K.; CHAKI, R.; BISWAS, A. Power aware cluster based routing (pacbr) protocol for wireless sensor network. In: SPRINGER. *IFIP International Conference on Computer Information Systems and Industrial Management*. [S.l.], 2013. p. 289–300.

DERPANIS, K. G. The bhattacharyya measure. *Mendeley Computer*, v. 1, n. 4, p. 1990–1992, 2008.

DEZA, M. M.; DEZA, E. Encyclopedia of distances. In: *Encyclopedia of distances*. [S.l.]: Springer, 2009. p. 1–583.

EMYGDIO, B. M. et al. Eficiência de coeficientes de similaridade em genótipos de feijão mediante marcadores rapd. *Pesquisa Agropecuária Brasileira*, SciELO Brasil, v. 38, p. 243–250, 2003.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FACELI, K. e. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.: s.n.], 2011.

FÁVERO, L. P. et al. *Análise de dados: modelagem multivariada para tomada de decisões*. [S.l.]: Rio de janeiro: Elsevier, 2009.

FENG ZHONGYAN; FENG Z.;WAN, J.-X. L. X. Design and implementation of global name space in multilevel cluster file system. In: *Jisuanji Gongcheng/Computer Engineering*. [S.l.: s.n.], 2006. p. 67–69+72.

FERRARI, D. G. et al. Seleção de algoritmos para a tarefa de agrupamento de dados: uma abordagem via meta-aprendizagem. Universidade Presbiteriana Mackenzie, 2014.

FILHO, A. C. et al. Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão. *Ciência Rural*, SciELO Brasil, v. 38, p. 2138–2145, 2008.

GENUER R.; POGGI, J. T.-M. C. Variable selection using random forests. *Pattern Recognition Letters*, Elsevier, v. 31, n. 14, p. 2225–2236, 2010.

GOMES, D. A. et al. Caracterização de genótipos de mandioca por técnicas multivariadas. *Research, Society and Development*, v. 9, n. 7, p. e252974181–e252974181, 2020.

GOMES, D. A. et al. Análise multivariada para classificaçao da velocidade do vento no estado do ceará. *Sigmae*, v. 8, n. 2, p. 90–97, 2019.

GOWER, J. C. A general coefficient of similarity and some of its properties. *Biometrics*, JSTOR, p. 857–871, 1971.

HARTUV, E.; SHAMIR, R. A clustering algorithm based on graph connectivity. *Information processing letters*, Elsevier, v. 76, n. 4-6, p. 175–181, 2000.

HELLINGER, E. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, De Gruyter, v. 1909, n. 136, p. 210–271, 1909.

HERRERO, J.; VALENCIA, A.; DOPAZO, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, Oxford University Press, v. 17, n. 2, p. 126–136, 2001.

HUANG, Z. Clustering large data sets with mixed numeric and categorical values. In: CITESEER. *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. [S.l.], 1997. p. 21–34.

HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, Springer, v. 2, n. 3, p. 283–304, 1998.

INZOLI, S.; GIUDICI, M.; HUISMAN, J. Estimation of alluvial sediments properties with a multilevel cluster analyses of spectral induced polarization data. In: EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS. *Near Surface Geoscience 2015-21st European Meeting of Environmental and Engineering Geophysics*. [S.l.], 2015. v. 2015, n. 1, p. 1–5.

JAIN, A.; DUBES, R. *Algorithms for Clustering Data*. Prentice Hall, 1988. (Prentice Hall advanced reference series). ISBN 9780130222787. Disponível em: ⟨https://books.google.com.br/books?id=7eBQAAAAMAAJ⟩.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010.

JIANG, D.; TANG, C.; ZHANG, A. Cluster analysis for gene expression data: a survey. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 16, n. 11, p. 1370–1386, 2004.

KANG, L.; WU, L.; YANG, Y.-H. A novel unsupervised approach for multilevel image clustering from unordered image collection. *Frontiers of Computer Science*, Springer, v. 7, n. 1, p. 69–82, 2013.

KIM, D.-W.; LEE, K. H.; LEE, D. Fuzzy clustering of categorical data using fuzzy centroids. *Pattern recognition letters*, Elsevier, v. 25, n. 11, p. 1263–1271, 2004.

KOHONEN, T. *Self-Organizing Maps*. [S.l.]: Springer Series in Information Sciences, 2001.

LAROS, J. A.; MARCIANO, J. L. P. Análise multinível aplicada aos dados do nels: 88. *Estudos em avaliação educacional*, v. 19, n. 40, p. 263–278, 2008.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MAHALANOBIS, P. C. On the generalized distance in statistics. In: NATIONAL INSTITUTE OF SCIENCE OF INDIA. [S.l.], 1936.

MCGARIGAL, K.; CUSHMAN, S. A.; STAFFORD, S. *Multivariate statistics for wildlife and ecology research*. [S.l.]: Springer Science & Business Media, 2000.

MENDOZA-MORALES, A. J.; GONZÁLEZ-SANSÓN, G.; AGUILAR-BETANCOURT, C. Producción espacial y temporal de hojarasca del manglar en la laguna barra de navidad, jalisco, méxico. *Revista de biología tropical*, v. 64, n. 1, p. 259–273, 2016.

MERLO, J. et al. A brief conceptual tutorial on multilevel analysis in social epidemiology: interpreting neighbourhood differences and the effect of neighbourhood characteristics on individual health. *Journal of Epidemiology & Community Health*, BMJ Publishing Group Ltd, v. 59, n. 12, p. 1022–1029, 2005.

MEYER, A. d. S. *Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes.* Tese (Doutorado) — Universidade de São Paulo, 2002.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. In: *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. [S.l.: s.n.], 2007. p. 1–295.

NASCIMENTO, K. K. F. do et al. Análise multivariada de casos de leishmaniose visceral no cariri cearense nos anos de 2008 a 2017. *Sigmae*, v. 8, n. 2, p. 248–256, 2019.

NETO, A. S. A.; NEGREIROS, M. Avaliação da performance de índices de similaridade aplicados ao agrupamento de objetos textuais. *Revista Brasileira de Computação Aplicada*, v. 9, n. 4, p. 43–59, 2017.

OLIVA, F. et al. Contribuciones desde una perspectiva basada en proximidades al fuzzy k-means clustering. In: *XXVI Congreso Nacional de Estadística e Investigación Operativa*. [S.l.: s.n.], 2001.

PEREIRA, T. M. Discriminação de populações com diferentes graus de similaridade por redes neurais artificiais. Universidade Federal de Viçosa, 2009.

PEUGH, J. L. A practical guide to multilevel modeling. *Journal of school psychology*, Elsevier, v. 48, n. 1, p. 85–112, 2010.

PITA, R. D. d. R. *Clustering categorical data using the frequency factor*. Dissertação (Mestrado) — Universidade Federal da Bahia, 2019.

PUNJ, G.; STEWART, D. W. Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, SAGE Publications Sage CA: Los Angeles, CA, v. 20, n. 2, p. 134–148, 1983.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: ⟨https://www.R-project.org/⟩.

RDUSSEEUN, L.; KAUFMAN, P. Clustering by means of medoids. In: *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland*. [S.l.: s.n.], 1987. v. 31.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Embrapa Informática Agropecuária-Artigo em periódico indexado (ALICE)*, Revista de Sistema de Informação da FSMA, Macaé, n. 7, p. 7-21, 2011., 2011.

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

SAITTA, S.; RAPHAEL, B.; SMITH, I. F. A bounded index for cluster validity. In: SPRINGER. *International workshop on machine learning and data mining in pattern recognition*. [S.l.], 2007. p. 174–187.

SALTON, G. Developments in automatic text retrieval. *science*, American Association for the Advancement of Science, v. 253, n. 5023, p. 974–980, 1991.

SCOLTOCK, J. A survey of the literature of cluster analysis. *The Computer Journal*, The British Computer Society, v. 25, n. 1, p. 130–134, 1982.

SHARAN, R.; SHAMIR, R. Click: a clustering algorithm with applications to gene expression analysis. In: *Proc Int Conf Intell Syst Mol Biol*. [S.l.: s.n.], 2000. v. 8, n. 307, p. 16.

SPEECE, D. L.; MCKINNEY, J. D.; APPELBAUM, M. I. Classification and validation of behavioral subtypes of learning-disabled children. *Journal of Educational Psychology*, American Psychological Association, v. 77, n. 1, p. 67, 1985.

TAXT, T.; FLYNN, P. J.; JAIN, A. K. Segmentation of document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 11, n. 12, p. 1322–1329, 1989.

VIOLINI, S.; PASAPERA, J. *Uso de suelo agrícola en la zona central de Córdoba: análisis de datos espaciales multisensor para su estudio y gestión*. [S.l.]: Argentina: IG-CONAE/UNC, 2016.

WILSCHUT, T. et al. Multilevel flow-based markov clustering for design structure matrices. *Journal of Mechanical Design*, American Society of Mechanical Engineers, v. 139, n. 12, p. 121402, 2017.

XU, R.; WUNSCH, D. *Clustering*. [S.l.]: John Wiley & Sons, 2008. v. 10.