UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ÍCARO JOSIAS FERREIRA PAIVA

Método de Particionamento Utilizando Distância de Mahalanobis para Intervalos

Recife

2024

ÍCARO JOSIAS FERREIRA PAIVA

Método de Particionamento Utilizando Distância de Mahalanobis para Intervalos

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

Área de Concentração: Inteligência Computacional.

Supervisor: Renata Maria Cardoso Rodrigues de Souza

Co-supervisor: Leandro Carlos de Souza

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Paiva, Ícaro Josias Ferreira.
Método de particionamento utilizando distância de Mahalanobis para intervalos / Ícaro Josias Ferreira Paiva. - Recife, 2024.
61f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-graduação em Ciência da Computação, 2024.
Orientação: Renata Maria Cardoso Rodrigues de Souza.
Coorientação: Leandro Carlos de Souza.
Inclui bibliografia.

1. Agrupamento; 2. Dados Intervalares; 3. Distância de Mahalanobis. I. Souza, Renata Maria Cardoso Rodrigues de. II. Souza, Leandro Carlos de. III. Título.

UFPE-Biblioteca Central

AGRADECIMENTOS

Primeiramente, gostaria de expressar minha gratidão à minha família, por seu amor incondicional e apoio contínuo. Agradeço aos meus amigos e amigas que sempre acreditaram no meu potencial e me incentivaram a seguir em frente.

Gostaria de agradecer a Dra. Renata de Souza, minha orientadora, pela orientação, paciência e apoio durante todo o processo de elaboração desta dissertação. Sua expertise e conselhos foram fundamentais para a realização deste trabalho.

Agradeço, ainda, aos meus parceiros de pesquisa e colaboradores por suas contribuições valiosas e por compartilharem seu conhecimento e experiência comigo.

A todos, meu sincero obrigado.

RESUMO

Este trabalho investiga métodos de particionamento utilizando a distância de Mahalanobis para dados intervalares. Os dados intervalares apresentam-se como uma alternativa aos dados clássicos, pois permitem capturar mais variabilidade e incerteza nos dados. Nesse sentido, é possível fazer uso desses atributos para organizar dados similares em grupos e separar dados distintos, minimizando a distância intra-grupos e maximizando a distância inter-grupos. Em muitas aplicações reais, como na análise de dados climáticos, biométricos ou financeiros, os dados não estão disponíveis em um valor pontual, mas sim como intervalos que refletem incertezas ou flutuações. Nesses casos, o uso de métodos tradicionais de análise pode ser limitado, pois não capturam adequadamente essa variabilidade. Ao avaliar as possibilidades de disposição dos conjuntos e a relação entre seus atributos, pode-se obter informações importantes para melhorar os resultados do agrupamento, tornando a distância de Mahalanobis uma aliada poderosa por considerar a correlação das variáveis. A distância de Mahalanobis se destaca nesses cenários por utilizar a matriz de variâncias e covariâncias em seus cálculos e adaptar-se bem à análise de dados intervalares. Com isso, o método proposto é capaz de formar agrupamentos robustos em dados intervalares, mesmo na presença de assimetria entre os limites inferiores e superiores. Para validar esta abordagem, o método desenvolvido foi avaliado e comparado com outras técnicas da literatura que utilizam abordagens similares. Os resultados indicam uma melhoria na qualidade dos agrupamentos, particularmente em cenários com presença de correlação entre variáveis, ainda que a variabilidade dos dados seja significativamente divergente entre seus limites. Além disso, este trabalho traz uma aplicação em um conjunto de dados climáticos brasileiros, demonstrando a capacidade de encontrar padrões entre as estações meteorológicas. Conclui-se que a distância de Mahalanobis para dados intervalares, desenvolvida pelo método proposto, pode proporcionar agrupamentos mais precisos e significativos, tornando-se uma ferramenta poderosa para análise de dados com alta variabilidade e incerteza.

Palavras-chave: Agrupamento; Dados Intervalares; Distância de Mahalanobis.

ABSTRACT

This study investigates partitioning methods using the Mahalanobis distance for interval data. Interval data presents itself as an alternative to classical data, as it allows capturing more variability and uncertainty in the data. In this sense, these attributes can be leveraged to organize similar data into groups and separate distinct data, minimizing intra-group distance and maximizing inter-group distance. In many real-world applications, such as climate, biometric, or financial data analysis, the data is not available as a single value but rather as intervals that reflect uncertainties or fluctuations. In such cases, traditional analysis methods may be limited, as they do not adequately capture this variability. By evaluating the possible arrangements of the sets and the relationships between their attributes, important insights can be gained to improve clustering results, making Mahalanobis distance a powerful ally, as it considers variable correlations. Mahalanobis distance stands out in these scenarios by utilizing the variance-covariance matrix in its calculations and adapting well to interval data analysis. Thus, the proposed method is capable of forming robust clusters in interval data, even in the presence of asymmetry between the lower and upper bounds. To validate this approach, the developed method was evaluated and compared with other techniques in the literature that use similar approaches. The results indicate an improvement in clustering quality, particularly in scenarios with correlated variables, even when the data variability is significantly divergent between its bounds. Additionally, this work presents an application to a set of Brazilian climate data, demonstrating the method's ability to find patterns among meteorological stations. It is concluded that Mahalanobis distance for interval data, as developed in the proposed method, can provide more accurate and meaningful clusters, becoming a powerful tool for data analysis with high variability and uncertainty.

Keywords: Clustering; Interval Data; Mahalanobis Distance.

LISTA DE FIGURAS

LISTA DE TABELAS

CONTEÚDO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Uma das habilidades mais básicas dos seres vivos envolve o agrupamento de objetos semelhantes para produzir uma classificação. A ideia de agrupar coisas semelhantes em categorias é claramente primitiva, uma vez que os primeiros humanos, por exemplo, devem ter sido capazes de perceber que muitos objetos individuais compartilham certas propriedades, como serem comestíveis, venenosos, ou ferozes, e assim por diante (EVERITT LANDAU; STAHL, 2011). A metodologia de agrupamento é fundamental em situações onde há pouca ou nenhuma informação prévia sobre a distribuição dos dados. Nessas condições, os métodos de agrupamento permitem identificar a estrutura dos dados com base em similaridades, sem depender de pressupostos estatísticos rígidos, facilitando uma análise eficiente das relações entre os pontos amostrais (JAIN; MURTY; FLYNN, 1999).

O processo de agrupamento constitui uma técnica estatística de aprendizado não supervisionado, cuja finalidade é identificar estruturas subjacentes sem a necessidade de rótulos previamente definidos. No contexto de mineração de dados, esse processo enfrenta três desafios fundamentais: (i) a escalabilidade computacional em grandes volumes de dados, (ii) a alta dimensionalidade, que aumenta a complexidade de processamento devido ao elevado número de atributos, e (iii) a heterogeneidade dos tipos de atributos, que impõe a necessidade de técnicas mais sofisticadas. Esses fatores acarretam um aumento significativo nas exigências computacionais, limitando a eficácia dos algoritmos tradicionais de agrupamento (BERKHIN, 2006).

Portanto, a extração de conhecimento a partir de grandes bases de dados é, atualmente, um aspecto fundamental para a tomada de decisões. Na prática, o objetivo consiste em revelar novas percepções a partir de uma base de dados por meio da utilização de uma tabela de dados padrão, na qual as entradas são compostas por um conjunto de unidades descritas por variáveis categóricas ou quantitativas finitas (DIDAY; NOIRHOMME-FRAITURE, 2008). Na Ciência de Dados, quando os conjuntos de dados são enormes, uma das primeiras abordagens é agregar indivíduos em classes que representam as unidades, com o objetivo de reduzir o tamanho inicial da população, resumindo assim os dados. (DIDAY, 2016). Com isso, é importante explorar formas de representação que permitam estudar as classes sem ignorar a complexidade do comportamento dos indivíduos integrantes.

Diday (1989) introduziu um novo paradigma para lidar com dados mais complexos, a Análise de Dados Simbólicos (ADS). O principal objetivo da ADS é desenvolver métodos para tratamento de dados mais complexos como intervalos, conjuntos e distribuição de probabilidades ou de pesos. ADS inicia com a agregação/redução de bases de dados clássicos em uma estrutura mais complexa chamada de dados simbólicos, pois eles contêm variação interna e são estruturados.

Dados intervalares são frequentemente utilizados para representar fenômenos com incerteza ou variabilidade. Cada observação é descrita por um intervalo $[x_L, x_U]$, onde x_L é o limite inferior e x_U o limite superior. Esse formato é útil em áreas como análise financeira e engenharia de qualidade, onde os dados apresentam variação inerente. Algumas áreas já utilizam constantemente esse tipo de estrutura, seja para captar a incerteza intrínseca do fenômeno avaliado, seja pela necessidade de observar diretamente os limites. Por exemplo: (a) na análise financeira, as projeções de preço e risco são representadas como intervalos para capturar a incerteza do mercado; (b) na engenharia de qualidade, as tolerâncias de fabricação são expressas como intervalos para garantir a conformidade dos produtos; (c) da mesma forma, em estudos climáticos, as medições de temperatura, precipitação e umidade frequentemente são representadas como intervalos devido à conveniência no dia a dia.

Dentro desse contexto, surgem novas oportunidades para explorar formas efetivas de utilizar estruturas de dados intervalares com o objetivo de obter agrupamentos mais homogêneos. Ao trabalhar com dados intervalares, observamos que os limites superiores e inferiores de uma mesma classe podem apresentar padrões e comportamentos divergentes. Representar adequadamente as informações desses limites é fundamental para determinar a qualidade dos agrupamentos obtidos.

A literatura sobre algoritmo de nuvens dinâmicas de dados intervalares explora diversas abordagens baseadas em distâncias para intervalo. Chavent e Lechevallier (2002) introduziram o uso da distância de Hausdorff, enquanto Souza, Carvalho e Silva (2004) propuseram a distância City-Block, adaptada para dados intervalares. Posteriormente, Carvalho e Lechevallier (2009a) exploraram o agrupamento dinâmico baseado em adaptações únicas (City-Block e Hausdorff). Carvalho, Brito e Bock (2006) sugeriram a distância L_2, acompanhada de técnicas de padronização e Souza, Souza e Amaral (2020) propuseram uma distância híbrida baseada em L_q. Outras abordagens introduziram aprimoramentos como a distância Wasserstein, proposta por Irpino e Verde (2008) e adaptada por Sun et

al. (2022), que considera a distribuição dos intervalos, oferecendo maior sensibilidade à dispersão interna.

Outras linhas de pesquisa utilizam a correlação como uma alternativa para a melhoria de agrupamento aplicada também a dados intervalares. Essa métrica de similaridade leva em consideração as covariância entre variáveis e a dispersão dos dados, oferecendo uma medida mais robusta e informativa do que métricas como a distância euclidiana. Souza, Carvalho e Tenorio (2004) apresentaram a distância quadrática adaptativa, que permite considerar a correlação entre os limites das variáveis. Esse método utiliza matrizes de covariância separadas para os limites inferior e superior, permitindo que a distância seja calculada de forma híbrida. Carvalho e Lechevallier (2009b) também propuseram distâncias quadráticas adaptativas em que as matrizes de covariância limites inferior e superior são combinadas para ambos os limites.

Essas abordagens apresentaram bons resultados, especialmente em situações onde há correlação entre os limites das variáveis, demonstrando a eficiência da distância de similaridade que utilizam a covariância como parâmetro em seus cálculos. Sendo assim, a distância de Mahalanobis é uma alternativa poderosa para a análise de covariância e pode ser aplicada também a dados intervalares. Essa métrica de similaridade leva em consideração as correlações entre variáveis e a dispersão dos dados, oferecendo uma medida mais robusta e informativa do que métricas como a distância euclidiana. No entanto, os desafios associados à representação e manipulação de intervalos, bem como à estimação da matriz de covariância para esses dados, exigem abordagens inovadoras e rigorosas.

Muitos métodos de agrupamento assumem que os limites inferior e superior dos intervalos têm a mesma importância, o que pode gerar problemas práticos, já que um dos limites pode ser menos relevante, mas ainda assim influenciar o resultado do agrupamento. Para superar essa limitação, Rodríguez e Carvalho (2022) propôs uma abordagem baseada nas distâncias Euclidiana e City-Block, que lida de maneira mais flexível com os limites intervalares, independentemente de sua simetria. Esse avanço foi possível por utilizar os pesos da distância adaptativa conjuntamento, permitindo uma nova forma de calcular a dissimilaridade.

Esta dissertação apresenta um método de agrupamento para dados intervalares, utilizando a distância de Mahalanobis para lidar com a assimetria entre os limites superior e inferior. A mudança central está na utilização pesos da distancia adaptativa conjuntamente, o que permite ao método acomodar discrepâncias significativas entre esses limites,

melhorando a qualidade dos agrupamentos por meio de uma integração da covariância. A pesquisa aborda tanto os aspectos teóricos quanto práticos, oferecendo uma contribuição importante ao campo, com novas ferramentas para análise de dados intervalares aplicáveis em diversos domínios.

1.2 OBJETIVO

O objetivo principal deste trabalho é desenvolver um método de agrupamento particional adaptando a distância de Mahalanobis para dados intervalares. Para atingir esse o objetivo principal, ao longo deste trabalho os objetivos específicos dessa dissertação são:

- Propor um novo método de partição de dados intervalares com distancia de Mahalanobis;
- avaliar o método proposto com experimentos em dados sintéticos e dados reais;
- realizar analise comparativa utilizando medidas de avaliação de desempenho da qualidade do agrupamento;
- realizar uma aplicação do método proposto em um problema com dados climáticos brasileiro.

1.3 ORGANIZAÇÃO

Além do atual capítulo introdutório, este trabalho é organizado da seguinte maneira:

Capítulo 2 - Referencial Teórico: Serão apresentados os conceitos necessários que norteiam premissas básicas do método como agrupamento, medidas de dissimilação e dados intervalares. Além disso, serão apresentado métodos da literatura para agrupamento de intervalar.

Capítulo 3 - Método Proposto: Será descrito as metodologias e técnicas utilizadas nas etapas do método proposto (DMCB). Sendo esse método visando fazer agrupamento de dados intervalares utilizando a distância de mahalanobis.

Capítulo 4 - Experimentos e Resultados: Nesta etapa, Será elaborado a criação de conjunto de dados sintéticos apresentado conjuntos de dados reais para a avaliação.

Esses experimentos serão feitos tanto com o métodos da literatura quanto o método proposto e avaliado os resultado obtidos nos índices descritos.

Capítulo 5 - Aplicação em Dados Climáticos: Neste capítulo, será apresentado um conjunto de dados reais climáticos. Será apresentado as etapas de pre-processamento do conjunto, métricas de avaliação e resultados do agrupamento obtidos pelo método proposto.

Capítulo 6 - Conclusão: Avaliação final do trabalho e analise de trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo serão discutidos os principais conceitos utilizados para o desenvolvimento deste trabalho.

2.1 ANALISE DE AGRUPAMENTO

Agrupamento refere-se à divisão de dados em grupos de objetos semelhantes. Cada grupo, ou *cluster*, consiste em objetos semelhantes entre si e diferentes de objetos de outros grupos (EVERITT LANDAU; STAHL, 2011).

O agrupamento de dados é uma técnica para análise estatística de dados, utilizada em muitos campos, incluindo aprendizado de máquina, mineração de dados, reconhecimento de padrões, análise de imagens e bioinformática. Agrupar refere-se à tarefa de dividir um conjunto de dados em grupos (*clusters*) com base nas semelhanças entre os objetos, sem que haja rótulos ou categorias predefinidos, de modo que os dados em cada subconjunto (idealmente) compartilhem compartilhem características similares, frequentemente proximidade de acordo com alguma medida de distância definida (ABONYI; FEIL, 2007).

A análise de agrupamentos é a organização de uma coleção de padrões (geralmente representados como um vetor de medidas ou um ponto em um espaço multidimensional) em *clusters* com base na similaridade. Intuitivamente, os padrões dentro de um *cluster* válido são mais semelhantes entre si do que são a um padrão pertencente a um *cluster* diferente (JAIN; MURTY; FLYNN, 1999).

Os métodos baseados em distância são muito populares na literatura, porque podem ser usados com praticamente qualquer tipo de dados, desde que seja criada uma função de distância apropriada para esse tipo de dados. Assim, o problema de agrupamento pode ser reduzido ao problema de encontrar uma função de distância para aquele tipo de dados. Portanto, o projeto de funções de distância tornou-se uma importante área de pesquisa para mineração de dados por si só (KUMAR, 2007).

2.1.1 Medidas de Similaridade

Como a similaridade é fundamental para a definição de um *cluster*, uma medida da similaridade entre dois padrões retirados do mesmo espaço de características é essencial para a maioria dos procedimentos de agrupamento. Devido à variedade de tipos e escalas de características, a medida de distância (ou medidas) deve ser escolhida cuidadosamente. É mais comum calcular a dissimilaridade entre dois padrões usando uma medida de distância definida no espaço de características. (ABONYI; FEIL, 2007)

Quando todas as variáveis registradas são contínuas, as proximidades entre os indivíduos são normalmente quantificadas por medidas de dissimilaridade ou medidas de distância. Uma variedade de medidas foram propostas para derivar uma matriz de dissimilaridade a partir de um conjunto de observações multivariadas contínuas (EVERITT LANDAU; STAHL, 2011).

Segundo (MAO; JAIN, 1996), se os *clusters* em um conjunto de dados forem suficientemente 'compactos' e 'isolados', no sentido de que a variação entre *clusters* seja significativamente maior do que a variação dentro de cada *cluster*, a maioria dos métodos de agrupamento será capaz de identificar esses *clusters*. No entanto, em muitos conjuntos de dados reais, os *clusters* não possuem uma separação clara ou uma forma geométrica simples, como a forma hiperesférica frequentemente assumida em técnicas que utilizam a distância euclidiana. Quando os *clusters* são alongados ou possuem formatos irregulares, a escolha da medida de distância pode impactar substancialmente os resultados do agrupamento. Por exemplo, algoritmos de agrupamento particional e técnicas baseadas em redes neurais que utilizam a distância euclidiana tendem a dividir *clusters* grandes e alongados em certas circunstâncias.

Seja $\mathbf{x}_i$ e $\mathbf{x}_j$ dois vetores P-dimensional e $\mathbf{A}$ uma matriz P x P.

$$d_Q^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j) \tag{2.1}$$

A equação 2.1 generalista da distância na forma quadrática tem como o fator $\mathbf{A}$ que pode receber diferentes termos que modificam seu cálculo. Cada uma dessas métricas pode ser mais apropriada para uma distribuição dos dados do conjunto como observado na fig. 2.1.

- Quando temos o termo igual a matriz identidade, $\mathbf{A} = \mathbf{I}$, então a distância Eucli-

diana. Muitos algoritmos de agrupamento particional são adequados para detectar *clusters* em formato hiperesférico, porque a distância euclidiana é a mais comumente usada para calcular a distância entre um padrão e o centro do *cluster* atribuído.

$$d_E^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) \tag{2.2}$$

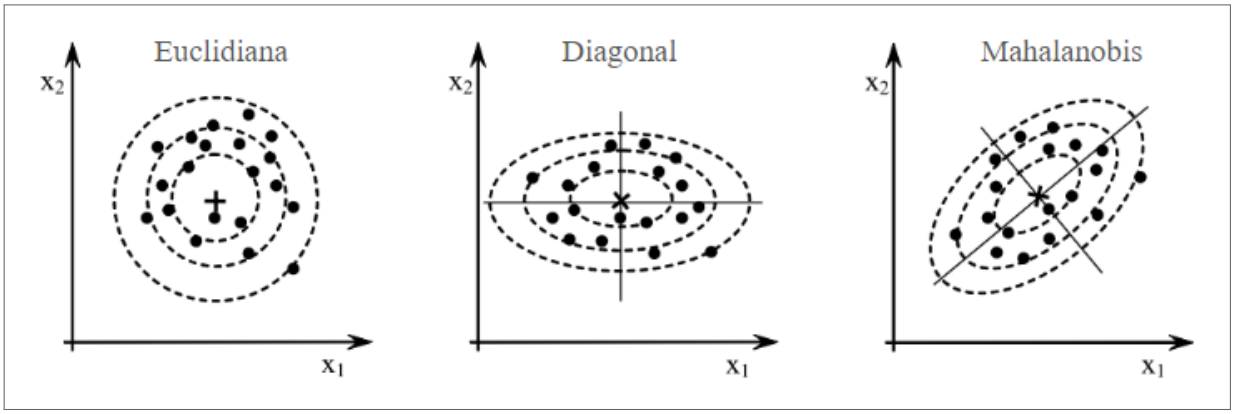
- Analogamente podemos criar a distancia euclidiana ponderada onde a diagonal no lugar receber valor unitário, como a euclidiana, passa a receber valores que funcionam como pesos, $\mathbf{A} = \mathbf{W}$ com $\mathbf{W} = (w_1, w_2, \cdots, w_P) \cdot \mathbf{I}$. A distância ponderada lida com formato hiperelipsoidal, porém apenas variadas nos paralela as coordenadas e não relevam a covariância entre as variáveis.

$$d_P^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} (\mathbf{x}_i - \mathbf{x}_j) \tag{2.3}$$

- A distância de Mahalanobis também pode ser utilizada como critério de agrupamento para lidar com aglomerados de formato hiperelipsoidal. Assim, a distância entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_j$ é dada por uma fórmula que envolve a matriz inversa da covariância dos pontos, $\mathbf{A} = \mathbf{\Sigma}^{-1}$. Essa abordagem permite atribuir pesos diferentes às características com base em suas variâncias e nas correlações lineares entre pares de variáveis

$$d_M^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{2.4}$$

Figura 1 – Medidas de similaridade.



Fonte: ABONYI; FEIL (2007)

2.1.2 Métodos de Agrupamento

Uma das razões para a diversidade de algoritmos é que não existe uma definição geral de *cluster* e, na verdade, existem vários tipos deles: *clusters* esféricos, *clusters* alongados, *clusters* lineares e assim por diante. Além disso, diferentes aplicações fazem uso de diferentes tipos de dados, como variáveis contínuas, variáveis discretas, similaridades e dissimilaridades. Portanto, é necessário ter diferentes métodos de agrupamento para se adaptar ao tipo de aplicação e ao tipo de *clusters* desejados (KAUFMAN; ROUSSEEUW, 2009).

A escolha de um algoritmo de agrupamento depende tanto do tipo de dados disponíveis quanto do propósito específico. Às vezes, vários algoritmos são aplicáveis, e argumentos a priori podem não ser suficientes para reduzir a escolha a um único método (KAUFMAN; ROUSSEEUW, 2009).

Os métodos de partição precisam ser fornecidos com um conjunto de sementes iniciais (ou *clusters*) que são então melhoradas iterativamente. Os métodos hierárquicos, por outro lado, podem começar com pontos de dados individuais em *clusters* únicos e construir o *clusters*. O papel da métrica de distância também é diferente em ambos os algoritmos. No agrupamento hierárquico, a métrica de distância é aplicada inicialmente nos pontos de dados no nível base e depois aplicada progressivamente nos sub-*clusters*, escolhendo pontos representativos absolutos para os sub-*clusters*. Porém, no caso de métodos particionais, em geral, os pontos representativos escolhidos em diferentes iterações podem ser pontos virtuais como o centroide do *clusters* (KUMAR, 2007).

Os métodos de agrupamento podem divergir em sua classificados segundo os seguintes aspectos:

Partição versus Hierárquico: O agrupamento hierárquico é um método de agrupamento que organiza os dados em uma hierarquia de *clusters*, onde cada *clusters* contém objetos de dados similares. Esses agrupamentos são então combinados em *clusters* maiores e assim por diante, formando uma estrutura em forma de árvore chamada dendrograma (PENG; LI, 2006). Por outro lado, os algoritmos de partição visam descobrir agrupamentos nos dados otimizando uma função objetivo específica e aprimorando iterativamente a qualidade das partições. Durante essas iterações, pontos representativos, como centroides de *clusters* (que podem não corresponder a

nenhum ponto de dados real), são escolhidos para melhorar a qualidade dos agrupamentos encontrados (KUMAR, 2007).

Rígido versus Difusa: Em uma partição, cada objeto do conjunto de dados é atribuído a um e apenas um *clusters*. Portanto, os métodos de particionamento às vezes são ditos para produzir um agrupamento rígido, porque tomam uma decisão clara para cada objeto. Por outro lado, um método de agrupamento difuso permite alguma ambiguidade nos dados, o que frequentemente ocorre (KAUFMAN; ROUSSEEUW, 2009).

2.2 ANÁLISE DE DADOS SIMBÓLICOS

Quando avaliamos as classes, sua descrição não pode ser meramente expressa por valores numéricos e categóricos. Isso se deve à natureza da variabilidade dos indivíduos dentro de cada classe. Essa variabilidade é melhor compreendida por meio de intervalos, histogramas, distribuições de probabilidade, gráficos de barras, sequências de valores categóricos ou numéricos. Esses tipos de dados são denominados "simbólicos", pois não podem ser reduzidos a números sem perda significativa de informação. As variáveis simbólicas associam a cada classe um valor simbólico (DIDAY, 2016).

Existe três principais formas de se obter dados simbólicos: Primeira ordem - quando a observação do individuo apresenta naturalmente mais de um valor para determinada variável; Segunda ordem - quando é necessário analisar não indivíduos isolados (ou "objetos de primeira ordem"), mas sim classes mais ou menos homogêneas de indivíduos ("objetos agregados", "super-indivíduos", "objetos de segunda ordem"); Dados incertos - quando os estudos que não podem ser baseados em resultados experimentais ou de entrevistas definitivos e únicos, mas devem levar em consideração alguma imprecisão, incerteza e plausibilidade ao registrar um 'valor' das variáveis subjacentes. Isso inclui dados probabilísticos ou possibilísticos, dados difusos, medições com intervalos de tolerância, etc (BOCK; DIDAY, 2012).

Assim como em analises de dados clássicos os dados simbólicos podem ser encontrados em variáveis categóricas e numéricas.

2.2.1 Dados Intervalares

Dentre todos os formatos de dados simbólicos, os dados com valores de intervalo desempenham um papel importante não apenas porque são os mais comumente vistos, mas também porque as técnicas para analisá-los muitas vezes podem ser prontamente generalizadas para outros tipos de dados (XU, 2010). Dados intervalares são descritos por um grupo de variáveis, cada uma das quais contém um intervalo de valores contínuos em vez do tradicional valor contínuo ou discreto único. A análise de dados tradicional simplesmente substitui cada intervalo por seu representante (por exemplo, centro ou média) e ignora a informação estrutural dos intervalos (PENG; LI, 2006).

Uma variável aleatória simbólica com valor de intervalo $\mathbf{Y}$ é aquela que assume valores em um intervalo; ou seja $\mathbf{Y} = \xi = [a, b] \in \mathbb{R}^1$ com $a \leq b$ e $a, b \in \mathbb{R}^1$. O intervalo pode ser aberto ou fechado em ambas extremidades (a, b), $[a, b)$, $(a, b]$ ou $[a, b]$ (BILLARD, 2006).

A Tabela 1 apresenta os valores para as variáveis aleatórias relacionadas ao tamanho das espécies de cogumelos, a saber, Y_1 = Largura do Chapéu do Píleo, Y_2 = Comprimento do Estipe e Y_3 = Largura do Estipe. Em particular, para a espécie Arorae, $Y_1 = [3.0, 8.0]$; ou seja, a largura do chapéu do píleo assume valores no intervalo $[3, 8]$. Um $Y_{\text{clássico}} = a$ é um $Y_{\text{simbólico}} = [a, b]$ (BILLARD, 2006).

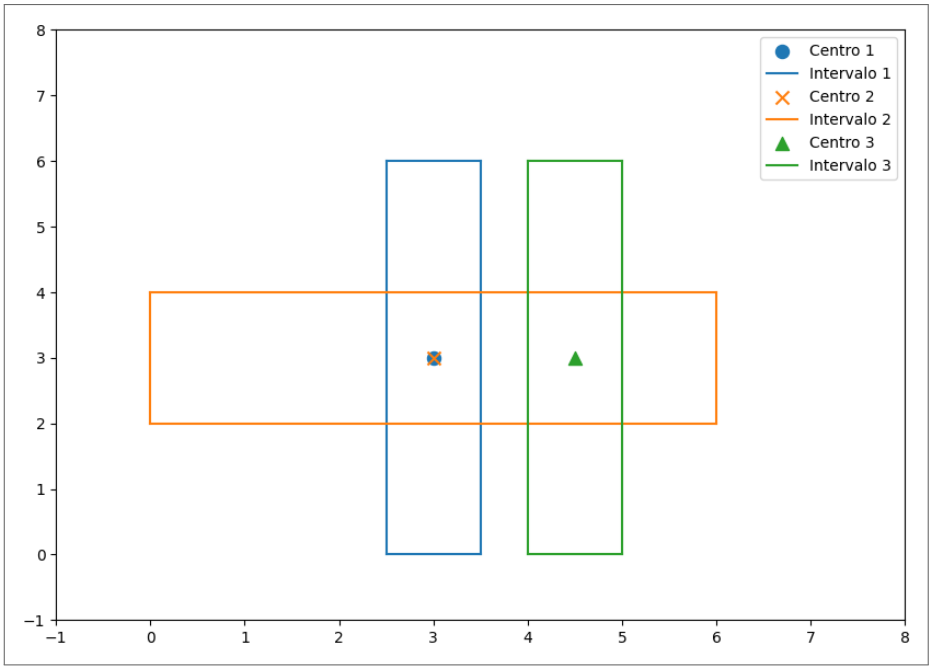
Tabela 1 – Dados de Cogumelos.

Espécie	**Pileus Comprimento**	**Estipe Largura**	**Estipe Comprimento**
arorae	[3.0, 8.0]	[4.0, 9.0]	[0.5, 2.5]
arvenis	[6.0, 21.0]	[4.0, 14.0]	[1.0, 3.5]
...	...	...	...

Fonte: Adaptado de BILLARD (2006)

Conjunto de dados tradicionais podem ser facilmente transpostas a tipos de dados de intervalo substituindo cada intervalo por um representante (por exemplo, a média dos pontos no intervalo). No entanto, essa abordagem ignora a informação estrutural do intervalo (PENG; LI, 2006). Como apresentado na Figura 2, se usarmos a média para representar os intervalos, não conseguiremos distinguir entre observação 1 e Observação 2. Com isso, mesmo a Obervação 3 tendo comportamento mais próximo ao comportamento da Observação 2 dependendo do tipo de dado usado podemos chegar a conclusões distintas.

Figura 2 – Limitação de uso médias para representar intervalos.



Fonte: Autor (2024)

2.2.2 Agrupamento de Dados Intervalares

A análise de agrupamentos de dados simbólicos é uma área de estudo importante, semelhante à análise de dados clássicos. A abordagem de dados simbólicos abre novas possibilidades para avaliar o agrupamento de dados, ampliando o espaço para o desenvolvimento de novas técnicas e adaptações dos métodos clássicos. Além disso, os dados intervalares oferecem uma maior facilidade de adaptação dos métodos já estabelecidos na literatura. Todas as classificações dos métodos mencionados na Seção 2.1.2 podem ser estendidas para os dados intervalares. Com isso, no agrupamento do tipo partição, a maior mudança ocorre no cálculo da similaridade entre os pontos intervalares e, naturalmente, na notação dos conjuntos e de seus representantes.

Os dados intervalares podem ser representados por um vetor de valores de intervalo. Seja A um conjunto de n objetos descritos por p variáveis intervalares. Cada i-ésimo objeto pode ser representado por um vetor $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^p)$, onde existem p valores de intervalo que $x_i^j = \left[a_i^j, b_i^j\right]$ e $a_i^j \leq b_i^j$. Suponha que queremos atribuir os objetos simbólicos em A a K grupos $C = (C_1, C_2, \ldots, C_K)$, onde C_k, $1 \leq k \leq K$ denota o k-ésimo grupo. Também utilizamos $i \in C_k$ para denotar que o i-ésimo objeto está no *cluster* C_k. Os K *clusters* têm suas representações ou protótipos correspondentes onde cada protótipo é

representado por um vetor de valores de intervalo, de modo que $\mathbf{g}_k = (g_k^1, g_k^2, \ldots, g_k^p)$ e $g_k^j = \left[w_k^j, y_k^j\right])$ e $w_k^j \leq y_k^j$.

Essa outra forma de organizar os dados permite também avaliar novas dinâmicas de relação entre as variáveis. Se para dados clássicos avaliar a covariância entre variáveis já está bastante já está bastante estruturado, quando falamos de dados simbólicos ainda existem linhas de pesquisas que buscam encontrar esses elementos da forma mais efetiva.

Estudos modernos de agrupamento de dados simbólicos tem avançado baseado no método de agrupamento dinâmico desenvolvido por Diday e colaboradores exemplificam essa abordagem, fornecendo uma estrutura e algoritmos para uma identificação precisa das classes e suas representações. Este método tem sido aplicado em diversas estruturas interclasses e modos de representação, como partições, hierarquias, sementes, leis de probabilidade, entre outros. Diday também propôs uma representação lógica de *clusters* em seu trabalho de 1976. Além disso, surgiram algoritmos baseados no método de agrupamento dinâmico, explorados por pesquisadores pioneiros como Diday, Govaert, Lechevallier, Sidi, entre outros (BOCK; DIDAY, 2012).

2.2.2.1 Métodos de Nuvens Dinâmicas para Intervalos usando uma Distância Quadrática Baseada em uma Matriz de Covariância para cada Limite dos Intervalos

Souza, Carvalho e Tenorio (2004) apresentaram o primeiro método de agrupamento baseado em nuvens dinâmicas usando uma distância quadrática para intervalos. Uma versão desse método com distância com pesos nas variáveis, chamada distância adaptativa, foi também proposto. Ambos os métodos consideraram que a distância entre vetores de intervalos é dada pela soma de duas distâncias quadráticas aplicadas aos limites inferior e superior dos intervalos, respectivamente.

Os métodos fornecem uma partição e um protótipo para cada grupo, otimizando um critério de adequação que mede o ajuste entre os grupos e seus representantes. As distâncias quadráticas adaptativas mudam a cada iteração do algoritmo e podem ser diferentes de um *cluster* para outro.

Com base em $\mathbf{x}_i$ e $\mathbf{g}_k$ descritos na seção 2.2.2, O critério a ser otimizado considerando distâncias adaptativas é dados por

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k) \tag{2.5}$$

Considerando os seguintes vetores: $\mathbf{x}_{iL} = (a_{i1}, \ldots, a_{ip})$ e $\mathbf{x}_{iU} = (b_{i1}, \ldots, b_{ip})$ para o $i-$ésimo objeto e $\mathbf{g}_{kL} = (y_{k1}, \ldots, y_{kp})$ e $\mathbf{g}_{kU} = (w_{k1}, \ldots, w_{kp})$ para o $k-$ésimo *cluster*. Temos então a distância adaptativa é dados por

$$d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k) = (\mathbf{x}_{iL} - \mathbf{g}_{kL})^T \mathbf{M}_{kL} (\mathbf{x}_{iL} - \mathbf{g}_{kL}) + (\mathbf{x}_{iU} - \mathbf{g}_{kU})^T \mathbf{M}_{kU} (\mathbf{x}_{iU} - \mathbf{g}_{kU}) \tag{2.6}$$

O algoritmo visa encontrar uma partição em k *clusters*, um conjunto de protótipos e um conjunto de distâncias minimizando J. Nesse caso, os protótipos $\{\mathbf{g}_k = (g^1_k, g^2_k, \cdots, g^p_k)\}$ são vetores de médias e as matrizes $\mathbf{M}_{kL}$ e $\mathbf{M}_{kU}$ são obtidas minimizando J considerando as restrições que $|\mathbf{M}_{kL}| = 1$ e $|\mathbf{M}_{kU}| = 1$. A solução é obtida pelo método multiplicadores de Lagranje e é dada por

$$\mathbf{M}_{kL} = [det\,(\mathbf{Q}_{kL})]^{1/p}\, \mathbf{Q}^{-1}_{kL} \text{ e } \mathbf{M}_{kU} = [det\,(\mathbf{Q}_{kU})]^{1/p}\, \mathbf{Q}^{-1}_{kU} \tag{2.7}$$

Três situações para as matrizes podem ser adotadas:

1. $\mathbf{Q}_{kL}$ e $\mathbf{Q}_{kU}$ são matrizes identidade. Nesse caso, a distância não é adaptativa e $d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k)$ é a distancia Euclidiana para intervalos (CARVALHO; BRITO; BOCK, 2006)

$$d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k) = (\mathbf{x}_{iL} - \mathbf{g}_{kL})^T (\mathbf{x}_{iL} - \mathbf{g}_{kL}) + (\mathbf{x}_{iU} - \mathbf{g}_{kU})^T (\mathbf{x}_{iU} - \mathbf{g}_{kU}) \tag{2.8}$$

2. $\mathbf{Q}_{kL}$ e $\mathbf{Q}_{kU}$ são matrizes diagonais dadas como segue e $d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k)$ é a distância a Euclidiana adaptativa por *cluster* e variável (SOUZA; CARVALHO; SILVA, 2004)

$$\begin{cases} \mathbf{Q}_{kL} = \text{Diag}\left(\sum_{i \in C_k} \left[(\mathbf{x}_{iL} - \mathbf{g}_{kL})\,(\mathbf{x}_{iL} - \mathbf{g}_{kL})^T\right]\right) \\ \mathbf{Q}_{kU} = \text{Diag}\left(\sum_{i \in C_k} \left[(\mathbf{x}_{iU} - \mathbf{g}_{kU})\,(\mathbf{x}_{iU} - \mathbf{g}_{kU})^T\right]\right) \end{cases} \tag{2.9}$$

3. $\mathbf{Q}_{kL}$ e $\mathbf{Q}_{kU}$ são matrizes completas e $d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k)$ é definida como duas Mahalanobis adaptativa para os limites inferiores e superiores dos intervalos, respectivamente

$$\begin{cases} \mathbf{Q}_{kL} = \sum_{i \in C_k} \left[(\mathbf{x}_{iL} - \mathbf{g}_{kL})\,(\mathbf{x}_{iL} - \mathbf{g}_{kL})^T\right] \\ \mathbf{Q}_{kU} = \sum_{i \in C_k} \left[(\mathbf{x}_{iU} - \mathbf{g}_{kU})\,(\mathbf{x}_{iU} - \mathbf{g}_{kU})^T\right] \end{cases} \tag{2.10}$$

2.2.2.2 *Métodos de Nuvens Dinâmicas para Intervalos usando uma Distância Quadrática Baseado em uma Matriz de Covariâncias Combinada*

Carvalho e Lechevallier (2009b) também apresentam métodos de agrupamento dinâmico de particionamento para dados com valor de intervalo baseados em distâncias quadráticas adaptativas. Esses métodos fornecem uma partição e um protótipo para cada *cluster*, otimizando um critério de adequação que mede o ajuste entre os *clusters* e seus representantes. Essas distâncias quadráticas adaptativas mudam a cada iteração do algoritmo e podem ser diferentes de um *cluster* para outro.

Em comparação ao trabalho apresentado em Souza, Carvalho e Tenorio (2004), os autores propuseram uma abordagem para a distancia quadrática adaptativa assumindo uma matriz combinada e a mesma para ambos os limites dos intervalos. O critério é dado por

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k) \tag{2.11}$$

com

$$d^2_{M_k}(\mathbf{x}_i, \mathbf{g}_k) = (\mathbf{x}_{iL} - \mathbf{g}_{kL})^T \mathbf{M}_k (\mathbf{x}_{iL} - \mathbf{g}_{kL}) + (\mathbf{x}_{iU} - \mathbf{g}_{kU})^T \mathbf{M}_k (\mathbf{x}_{iU} - \mathbf{g}_{kU}) \tag{2.12}$$

A matriz $\mathbf{M}_k$ é obtida minimizando J considerando a restrição que $|\mathbf{M}_k| = 1$. A solução é obtida pelo método multiplicadores de Lagranje e é dada por

$$\mathbf{M}_k = [det(\mathbf{Q}_k)]^{1/p} \mathbf{Q}_k^{-1} \tag{2.13}$$

1. Assumindo que as variáveis são independentes. A distância quadrática é também uma versão da distância Euclidiana adaptativa para intervalos. A matriz $\mathbf{Q}_k$ é definida como

$$\mathbf{Q}_k = \text{Diag}\left(\sum_{i \in C_k} \left[(\mathbf{x}_{iL} - \mathbf{g}_{kL})(\mathbf{x}_{iL} - \mathbf{g}_{kL})^T + (\mathbf{x}_{iU} - \mathbf{g}_{kU})(\mathbf{x}_{iU} - \mathbf{g}_{kU})^T\right]\right) \tag{2.14}$$

2. Considerando as covariâncias entre as variáveis, a matriz $\mathbf{Q}_k$ é como segue

$$\mathbf{Q}_k = \sum_{i \in C_k} \left[(\mathbf{x}_{iL} - \mathbf{g}_{kL})(\mathbf{x}_{iL} - \mathbf{g}_{kL})^T + (\mathbf{x}_{iU} - \mathbf{g}_{kU})(\mathbf{x}_{iU} - \mathbf{g}_{kU})^T\right] \tag{2.15}$$

3 MÉTODO PROPOSTO

Os métodos de nuvens dinâmicas com distâncias quadráticas adaptativas da literatura de ADS medem as diferenças entre elementos e *clusters* considerando uma combinação (hibridismo) através da soma de duas distâncias quadráticas: uma para os valores de limite inferior e outra para os valores de limite superior. Em Souza, Carvalho e Tenorio (2004) consideram uma matriz de variância-covariância para cada limite dos intervalos (Equação 2.7) e em Carvalho e Lechevallier (2009b) a matriz de variância-covariância utilizada para ambos os limites é a mesma (unificada), mas essa matriz resulta de uma combinação das matrizes de covariância calculadas separadamente para os limites inferior e superior dos intervalos (Equação 2.13).

A principal contribuição deste trabalho é apresentar um método de nuvens dinâmicas para intervalos que utiliza a distância de Mahalanobis. Diferentemente de abordagens anteriores, essa distância não é definida como uma combinação de duas dissimilaridades aplicadas aos limites dos intervalos (hibridismo), como apresentado em Souza, Carvalho e Tenorio (2004). Além disso, essa distância permite levar em conta as variância-covariâncias de cada limite dos intervalos assumindo que os limites dos intervalos podem ter variabilidades diferentes ao contrário do método proposto por Carvalho e Lechevallier (2009b) que usa variância-covariâncias combinadas.

Seja $A = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ um conjunto de n objetos descritos por p variáveis intervalares para ser agrupado em K *clusters*. Seja $P = \{C_1, \ldots, C_K\}$ uma partição de A em K *clusters*. Cada i-ésimo objeto de A sendo representado por um vetor p valores de intervalo $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^p)$, com $x_i^j = \left[a_i^j, b_i^j\right]$ e $a_i^j \leq b_i^j$. Seja $G = \{\mathbf{g}_1, \ldots, \mathbf{g}_K\}$ um conjunto de representantes (protótipos) dos *clusters* de P. Cada *cluster* é representado por um vetor de p variáveis intervalares $\mathbf{g}_k = (g_k^1, g_k^2, \ldots, g_k^p)$ e $g_k^j = \left[y_k^j, w_k^j\right])$ e $y_k^j \leq w_k^j$. Seja um conjunto de distâncias representado por um conjunto de K matrizes de covariância $M = \{\mathbf{M}_1, \ldots, \mathbf{M}_K\}$ associadas a P.

Considere $\mathbf{z}_i = (\mathbf{x}_{iL}, \mathbf{x}_{iU})$ $(i = 1, \ldots, n)$ e $\mathbf{r}_k = (\mathbf{g}_{kL}, \mathbf{g}_{kU})$ $(k = 1, \ldots, K)$ com $\mathbf{x}_{iL} = (a_{i1}, \ldots, a_{ip})$ e $\mathbf{x}_{iU} = (b_{i1}, \ldots, b_{ip})$ para o $i-$ésimo objeto e $\mathbf{g}_{kL} = (y_{k1}, \ldots, y_{kp})$ e $\mathbf{g}_{kU} = (w_{k1}, \ldots, w_{kp})$ para o $k-$ésimo *cluster*.

3.1 DISTÂNCIA DE MAHALANOBIS ADAPTATIVA PARA INTERVALOS

A distância de Mahalanobis adaptativa para intervalos é dada por

$$d^2_{M_k}(\mathbf{z}_i, \mathbf{r}_k) = (\mathbf{z}_i - \mathbf{r}_k)^T \mathbf{M}_k (\mathbf{z}_i - \mathbf{r}_k) \tag{3.1}$$

Os parâmetros dessa distância podem mudar iterativa por *cluster* e variável e o critério de agrupamento a ser otimizado é dado por

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} d^2_{M_k}(\mathbf{z}_i, \mathbf{r}_k) \tag{3.2}$$

A matriz $\mathbf{M}_k$ é de tamanho $2p \times 2p$ e é obtida minimizando J considerando a restrição $|\mathbf{M}_k| = 1$. A distância em Equação 3.1 pode ser vista como uma distância de Mahalanobis adaptativa clássica aplicada a vetores de valores numéricos. Nesse caso, o método proposto aqui é o método de nuvens dinâmicas com distâncias adaptativas introduzido por (GOVAERT, 1975) e a solução é apresentada como segue

$$\mathbf{M}_k = (|\mathbf{Q}_k|)^{1/2p} \mathbf{Q}_k^{-1} \tag{3.3}$$

Contudo como este trabalho objetiva propor um método de nuvens dinâmicas com distâncias quadráticas adaptativas para intervalos em que os elementos da matriz $\mathbf{M}_k$ são obtidos considerando as covariâncias das variáveis conjuntamente, a matriz $\mathbf{Q}_k$ de tamanho $2p \times 2p$ tem a seguinte estrutura

$$\mathbf{Q}_k = \begin{pmatrix} \mathbf{Q}_k^L & \mathbf{0}_{p\times p} \\ \mathbf{0}_{p\times p} & \mathbf{Q}_k^U \end{pmatrix}$$

Em comparação com o método correspondente para intervalos introduzido em Carvalho e Lechevallier (2009b), o método proposto nesse trabalho também considera uma única matriz adaptativa $\mathbf{M}_k$ porém essa matriz não é definida como uma matriz de variâncias-covariâncias combinadas. Aqui a matriz $\mathbf{M}_k$ é uma matriz definida em blocos baseada em variâncias-covariâncias conjuntamente. Esse caminho permite melhorar a qualidade do agrupamento levando em conta as covariâncias entre variáveis. A solução gera uma matriz $\mathbf{Q}_k$ diferente da matriz obtida pelo método de Carvalho e Lechevallier (2009b).

Quando $\mathbf{Q}_k^L$ e $\mathbf{Q}_k^U$ são matrizes identidade (ou seja assumindo $\mathbf{\Sigma}_k = \mathbf{I}$ para $k = 1, \ldots, K$) a distância em Equação 3.1 pode ser reescrita como segue e essa é mesma em Equação 2.8

$$d^2_{M_k}(\mathbf{z}_i, \mathbf{r}_k) = (\mathbf{z}_i - \mathbf{r}_k)^T (\mathbf{z}_i - \mathbf{r}_k) = (\mathbf{x}_{iL} - \mathbf{g}_{kL})^T (\mathbf{x}_{iL} - \mathbf{g}_{kL}) + (\mathbf{x}_{iU} - \mathbf{g}_{kU})^T (\mathbf{x}_{iU} - \mathbf{g}_{kU}) \tag{3.4}$$

3.2 PROBLEMA DE OTIMIZAÇÃO

O algoritmo inicia com uma partição inicial P e executa três passos iterativos até a convergência quando o critério J alcança um valor estacionário que representa um mínimo local. Assim, o problema de otimização pode ser solucionado por iterativamente resolver os seguintes problemas:

1. $[P_1]$: Fixado a partição $P = \hat{P}$ e o conjunto de matrizes $M = \hat{M}$ encontrar conjunto de protótipos $\hat{G}$.
2. $[P_2]$: Fixado a partição $P = \hat{P}$ e conjunto de protótipos $G = \hat{G}$ encontrar M.
3. $[P_3]$ Fixado o conjunto de protótipos $G = \hat{G}$ e o conjunto de matrizes $M = \hat{M}$ encontrar P.

Proposição 3.1. *O problema* $[P_1]$ *é solucionado calculando* $g_k^j = [y_k^j, w_k^j]$ *com*

$$y_k^j = \frac{\sum_{i \in C_K} a_i^j}{n_k} \;\; e \;\; w_k^j = \frac{\sum_{i \in C_K} b_i^j}{n_k} \tag{3.5}$$

sendo n_k *o tamanho do cluster* C_k.

Demonstração. Seja $b = (\sum_{i \in C_k} z_i)/n_k$. Temos que $\sum_{i \in C_k} (\mathbf{z}_i - \mathbf{b})^T = \sum_{i \in C_k} (\mathbf{z}_i - \mathbf{b}) = 0$ e o critério associado ao *cluster* k pode ser reescrito por

$$\sum_{i \in C_k} (\mathbf{z}_i - \mathbf{r}_k)^T \mathbf{M}_k (\mathbf{z}_i - \mathbf{r}_k) = \sum_{i \in C_k} (\mathbf{z}_i - \mathbf{b})^T \mathbf{M}_k (\mathbf{z}_i - \mathbf{b}) + n_k (\mathbf{b} - \mathbf{r}_k)^T \mathbf{M}_k (\mathbf{b} - \mathbf{r}_k)$$

Podemos observar que o critério associado ao *cluster* C_k é minimizado quando $\mathbf{r}_k = \mathbf{b}$. Como $\mathbf{r}_k = (\mathbf{y}_k, \mathbf{w}_k) = (y_k^1, \ldots, y_k^p, w_k^1, \ldots, w_k^p)$ temos que o protótipo do *cluster* C_k tem

os limites dos intervalos $g_k^j = [y_k^J, w_k^j]$ $(k = 1, \ldots, K); (j = 1, \ldots, p)$ calculado por $y_k^j = (\sum_{i \in C_k} a_i^j)/n_k$ e $w_k^j = (\sum_{i \in C_k} b_i^j)/n_k$. Como $\sum_{i \in C_k} a_i^j \leq \sum_{i \in C_k} b_i^j$ isso implica que $y_k^j \leq w_k^j$. □

Proposição 3.2. *O problema* $[P_2]$ *é solucionado calculando* $\boldsymbol{Q}_{kL}$ *e* $\boldsymbol{Q}_{kU}$*, ambas de tamanho* $p \times p$*, por*

$$\begin{cases} \mathbf{Q}_k^L = \sum_{i \in C_k} \left[(\mathbf{x}_{iL} - \mathbf{g}_{kL}) (\mathbf{x}_{iL} - \mathbf{g}_{kL})^T \right] \\ \mathbf{Q}_k^U = \sum_{i \in C_k} \left[(\mathbf{x}_{iU} - \mathbf{g}_{kU}) (\mathbf{x}_{iU} - \mathbf{g}_{kU})^T \right] \end{cases} \tag{3.6}$$

Demonstração. Seja $g(\mathbf{M}_k) = 1 - |\mathbf{M}_k|$. Como o critério é adtitivo com respeito a variabilidade de cada *cluster*, ou seja, $J = \sum_{k=1}^K J_k$ a ideia é determinar o valor extremo do critério J_k com respeito ao *cluster* C_k considerando a restrição $g(\mathbf{M}_k) = 0$. Seja a função Lagrangiana

$$\begin{aligned} L(\mathbf{M}_k, \beta) &= J_k(\mathbf{M}_k) + \beta g(\mathbf{M}_k) \\ &= \sum_{i \in C_k} [(\mathbf{z}_i - \mathbf{r}_k)^T \mathbf{M}_k (\mathbf{z}_i - \mathbf{r}_k) + \beta (1 - |\mathbf{M}_k|) \end{aligned}$$

Obtendo a derivada de $L(\mathbf{M}_k, \beta)$ com respeito a $\mathbf{M}_k$ temos,

$$\frac{dL(\mathbf{M}_k, \beta)}{d\mathbf{M}_k} = \sum_{i \in C_k} (\mathbf{z}_i - \mathbf{r}_k)^T \mathbf{M}_k (\mathbf{z}_i - \mathbf{r}_k) - \beta \, |\mathbf{M}_k| \mathbf{M}_k^{-1} = 0$$

Uma vez que $|\mathbf{M}_k| = 1$, temos que $\mathbf{M}_k^{-1} = (1/\beta) \, \mathbf{Q}_k$ com $\mathbf{Q}_k = \sum_{i \in C_k} [(\mathbf{z}_i - \mathbf{r}_k)(\mathbf{z}_i - \mathbf{r}_k)^T]$. Como $|\mathbf{M}_k^{-1}| = 1/|\mathbf{M}_k| = 1$ e $\mathbf{M}_k^{-1} = 1/\beta \, \mathbf{Q}_k$ temos

$$|\mathbf{M}_k^{-1}| = 1/(\beta^{2p}) \, |\mathbf{Q}_k| = 1 \Rightarrow \beta = |\mathbf{Q}_k|^{1/2p}$$

.

Como $\mathbf{M}_k^{-1} = 1/\beta \, \mathbf{Q}_k$ temos então que $\mathbf{M}_k^{-1} = 1/|\mathbf{Q}_k|^{1/2p} \, \mathbf{Q}_k$, assim temos que

$$\mathbf{M}_k = (|\mathbf{Q}_k|)^{1/2p} \, \mathbf{Q}_k^{-1}$$

.

É conhecido que $(\mathbf{z}_i - \mathbf{r}_k)^T \mathbf{M}_k (\mathbf{z}_i - \mathbf{r}_k) =$ traço $\left[(\mathbf{z}_i - \mathbf{r}_k)^T (\mathbf{z}_i - \mathbf{r}_k) \mathbf{M}_k \right]$ uma vez que $(\mathbf{z}_i - \mathbf{r}_k)^T \mathbf{M}_k (\mathbf{z}_i - \mathbf{r}_k)$ é um escalar. Assim, $J_k(\mathbf{M}_k)$ =traço$[\sum_{i \in C_k} (\mathbf{z}_i - \mathbf{r}_k)(\mathbf{z}_i -$

$\mathbf{r}_k)^T\mathbf{M}_k]$ =traço$[\mathbf{Q}_k\,\mathbf{M}_k]$. Com $\mathbf{M}_k = (|\mathbf{Q}_k|)^{1/2p}\,\mathbf{Q}_k^{-1}$ o critério $J_k(\mathbf{M}_k)$ alcança um valor extremo e esse valor é $J_k(\mathbf{M}_k)$ =traço$[\mathbf{Q}_k\ (|\mathbf{Q}_k|)^{1/2p}\,\mathbf{Q}_k^{-1}] = (2p)\,|\mathbf{Q}_k|)^{1/2p}$.

Por outro lado temos, $J_k(\mathbf{I})$ = traço$[\mathbf{Q}_k\mathbf{I}]$ = traço$[\mathbf{Q}_k]$. Como uma matriz simétrica definida positiva, temos $\mathbf{Q}_k = \mathbf{P}\boldsymbol{\Delta}\mathbf{P}^T$ (conforme procedimento de decomposição valor singular) sendo: $\mathbf{P}\mathbf{P}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}$, $\boldsymbol{\Delta}$= diag$(\delta_1, \ldots, \delta_{2p})$, and δ_j $(j = 1, \ldots, 2p)$ são auto-valores de $\mathbf{Q}_k$. Então, $J_k^2(\mathbf{I})$ = traço$(\mathbf{P}\boldsymbol{\Delta}\mathbf{P}^T)$ = traço$(\boldsymbol{\Delta})$= $\sum_{j=1}^{2p}\delta_j$. Além disso, $|\mathbf{Q}_k| = |\mathbf{P}\boldsymbol{\Delta}\mathbf{P}^T| = |\boldsymbol{\Delta}| = \Pi_{j=1}^{2p}\delta_j$.

É conhecido que a média aritmética é maior que a média geométrica, ou seja, $1/(2p)\,(\delta_1 + \ldots + \delta_{2p}) > (\delta_1 \times \ldots \times \delta_{2p})^{1/(2p)}$ (e a igualdade se mantém somente se $(\delta_1 = \ldots = \delta_{2p})$). Assim, temos que $J_k(\mathbf{I}) > J_k(\mathbf{M}_k)$. Logo, conclui-se que $J_k(\mathbf{M}_k))$ alcança um valor estacionário e mínimo.

□

Assumindo que as variáveis intervalares são independentes (ou seja Diag$\boldsymbol{\Sigma}_l \neq$ Diag$\boldsymbol{\Sigma}_k$ $(l, k = 1, \ldots, K)$ para $k = 1, \ldots, K)$, temos o caso especial em que $\mathbf{M}_k$ $(k - 1, \ldots, K)$ é uma matriz diagonal. Então a solução do problema $|P_2|$ é dada por

$$\begin{cases} \mathbf{Q}_k^L = \text{Diag}\left(\sum_{i\in C_k}\left[(\mathbf{x}_{iL} - \mathbf{g}_{kL})\,(\mathbf{x}_{iL} - \mathbf{g}_{kL})^T\right]\right) \\ \mathbf{Q}_k^U = \text{Diag}\left(\sum_{i\in C_k}\left[(\mathbf{x}_{iU} - \mathbf{g}_{kU})\,(\mathbf{x}_{iU} - \mathbf{g}_{kU})^T\right]\right) \end{cases} \tag{3.7}$$

Nesse caso, sejam $\boldsymbol{\lambda}_{kL} = (\lambda_{kL}^1, \ldots, \lambda_{kL}^p)$ e $\boldsymbol{\lambda}_{kU} = (\lambda_{kU}^1, \ldots, \lambda_{kU}^p)$ dois vetores de parâmetros. A distância em Equação 3.1 pode ser rescrita como uma distância Euclideana ponderada do *cluster* C_k dada por

$$d_{M_k}^2(\mathbf{z}_i, \mathbf{r}_k) = \sum_{j=1}^{p}\lambda_{kL}^j(\mathbf{x}_{iL} - \mathbf{g}_{kL})^2 + \lambda_{kU}^j(\mathbf{x}_{iU} - \mathbf{g}_{kU})^2 \tag{3.8}$$

com λ_{kL}^j e λ_{kU}^j definidos como

$$\begin{cases} \lambda_{kL}^j = \frac{\{\prod_{h=1}^{p}[\sum_{i\in C_k}(x_{iL}^h - y_k^h)^2][\sum_{i\in C_k}(x_{iU}^h - w_k^h)^2]\}^{1/2p}}{\sum_{i\in C_k}(x_{iL}^j - y_k^j)^2} \\ \lambda_{kU}^j = \frac{\{\prod_{h=1}^{p}[\sum_{i\in C_k}(x_{iL}^h - y_k^h)^2][\sum_{i\in C_k}(x_{iU}^h - w_k^h)^2]\}^{1/2p}}{\sum_{i\in C_k}(x_{iU}^j - w_k^j)^2} \end{cases} \tag{3.9}$$

A distância em Equação 3.8 é um caso particular da distância introduzida nesse trabalho assumindo independência entre as variáveis intervalares. Além disso, essa distância foi apresentada em Rodríguez e Carvalho (2022).

O problema $[P_3]$ é solucionado por associando o i-ésimo objeto de A ao *cluster* C_k tal que

$$d^2_{M_k}(\mathbf{z}_i, \mathbf{r}_k) \leq d^2_{M_k}(\mathbf{z}_i, \mathbf{r}_t), \text{ para } 1 \leq t \leq K.$$

Demonstração. A prova é fácil de ser compreendida. □

3.3 O ALGORITMO

O algoritmo básico por solucionando a partição P é dado como segue:

1. ***Inicialização***: Escolha K elementos de A para ser o conjunto de K protótipos G^0. Assuma $M^0 = \{\mathbf{I}, \ldots, \mathbf{I}\}$ e associe cada objeto de A a um *cluster* tal que a distância usando Equação 3.1 seja mínima. Obtenha a partição P^0 e coloque $t = 0$.

2. ***Atualizando Protótipos***: Seja $\hat{P} = P^t$ e $\hat{M} = M^t$ obtenha o conjunto dos melhores protótipos G^{t+1} usando Equação 3.5.

3. ***Atualizando Matrizes de Variâncias-Covariâncias***: Seja $\hat{P} = P^t$ e $\hat{G} = G^{t+1}$ obtenha o conjunto de matrizes M^{t+1} usando Equação 3.7 se é adotado independência entre variáveis ou Equação 3.6 se é adotado dependência entre variáveis.

4. ***Atualizando a Partição***: Seja $\hat{G} = G^{t+1}$ e $\hat{M} = M^{t+1}$ obtenha a partição P^{t+1} associe cada objeto de A a um *cluster* tal que a distância usando Equação 3.1 seja mínima.

5. ***Critério de Parada***: Se a partição $P^{t+1} = P^t$ pare o algoritmo, caso contrário faça $t = t + 1$ e vá para o passo 2.

4 EXPERIMENTOS E RESULTADOS

Neste capítulo, comparam-se os resultados do método proposto com os métodos encontrados na literatura. Na primeira seção, serão levantadas as métricas utilizadas para as avaliações, sendo elas IRA e NMI, que serão utilizadas na segunda e terceira seções. Na segunda seção, será detalhado como foram criados e aplicados os experimentos de agrupamento no conjunto de dados simulados. Na terceira seção, os experimentos serão aplicados em um conjunto de dados reais amplamente utilizado em diversos artigos científicos.

Em relação aos métodos avaliados temos, os primeiros métodos, introduzido por Souza, Carvalho e Silva (2004), apresentados na Seção 2.2.2.1. Nestes casos, a distância quadrática é baseada em uma matriz de covariância específica para cada limite dos intervalos. Quando usamos a matriz de covariância completa, conforme a Equação 2.10, o método é chamado de Distância Quadrática Completa Híbrida (DQCH). Quando utilizamos a matriz diagonal, de acordo com a mesma equação, o método é denominado Distância Quadrática Diagonal Híbrida (DQDH), conforme a Equação 2.9.

Outros métodos utilizados para comparação são apresentados, introduzido por Carvalho e Lechevallier (2009b), na Seção 2.2.2.2, utilizam a distância quadrática baseada em uma matriz de covariâncias combinada. Quando utilizamos a matriz de covariância completa, de acordo com a Equação 2.15, denomina-se o método de Distância Quadrática Completa Única (DQCU). Por outro lado, quando utilizamos a matriz diagonal, de acordo com a Equação 2.14, o método é denominado Distância Quadrática Diagonal Única (DQDU).

Já o método proposto é identificado com Distância Mahalanobis Completa Blocada (DMCB), com Equação 3.6, e o caso especial, a versão com a matriz diagonal do método proposto, introduzido por Rodríguez e Carvalho (2022), com Equação 3.7, Distância Mahalanobis Diagonal Blocada (DMDB).

4.1 MÉTRICAS

Nesta seção, serão abordadas as métricas que foram utilizadas para este estudo na área da análise de agrupamento.

Índice Rand Ajustado (IRA)

O índice IRA mede o grau de similaridade entre uma partição a priori e uma partição fornecida pelo algoritmo de agrupamento. Este índice IRA foi escolhido pois ele não é sensível ao número de classes nas partições e as distribuições dos elementos nas classes.

Se $U = \{u_1, \cdots, u_r, \cdots, u_R\}$ é uma partição dada como resultado de um método de *cluster*, e $V = \{v_1, \cdots, v_c, \cdots, v_C\}$ e partição a priori, o índice IRA é definido como:

$$\text{IRA} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{4.1}$$

onde n_{ij} representa o número de objetos que estão nas classes u_i e v_j; $n_{i.}$ indica o n número de objetos que estão na classe u_i; $n_{.j}$ indica o número de objetos que estão na classe v_j ;e n é o número total de objetos.

Os valores do IRA variam no intervalo [-1, 1]. O valor 1 mostra que as partições U e V são as mesmas. Valores próximos a 1 indicam uma forte concordância entre as partições comparadas. Entretanto, valores próximos a 0, ou negativos, indicam a não concordância entre as partições. Assim, o agrupamento apresenta um melhor desempenho à medida que o valor do IRA aumenta.

Informação Mútua Normalizado (IMN)

O Índice de Informação Mútua Normalizado (IMN) é uma medida que avalia o grau de similaridade entre duas partições, considerando a quantidade de informação compartilhada entre elas. O IMN é normalizado, o que permite que a métrica seja menos sensível ao número de *clusters* nas partições e à distribuição dos elementos entre esses *clusters*, em comparação com outras métricas que podem ser influenciadas por essas variações.

Dadas duas partições, $U = \{u_1, \cdots, u_r, \cdots, u_R\}$, resultantes de um algoritmo de agrupamento, e $V = \{v_1, \cdots, v_c, \cdots, v_C\}$, a partição de referência, o IMN é calculado como:

$$\text{IMN} = \frac{I(U;V)}{\sqrt{H(U) \cdot H(V)}} \tag{4.2}$$

Onde $I(U;V)$ é a informação mútua entre as partições U e V, e $H(U)$ e $H(V)$ são as entropias das partições U e V, respectivamente. O IMN varia de 0 a 1, onde 1 indica uma concordância perfeita entre as partições e 0 indica independência entre elas.

4.2 CONJUNTO DE DADOS SINTÉTICOS

Nesta etapa, foram simulados dois conjuntos de dados usuais no plano bidimensional ($\mathbb{R}^2$) com diferentes propriedades estatísticas. Cada conjunto de dados possui 450 pontos espalhados entre quatro classes, de tamanhos e formatos de elipse desiguais.: duas classes de tamanho 150 cada e duas classes de tamanhos 50 e 100. Para avaliar as propriedades do algoritmo desenvolvido foram adotados duas estrategias para definir os pontos inferiores e superiores do intervalo.

Os métodos serão avaliados em quatro conjuntos de dados, cada um apresentando diferentes comportamentos das classes. As principais características desses conjuntos incluem a dispersão das variáveis entre as classes (desigual ou similar) e a presença ou ausência de correlação entre as variáveis. Para avaliar a capacidade do método em lidar com diferenças de comportamento entre os intervalos inferior e superior, serão atribuídos cinco níveis de amplitude. Quanto maior o nível, maior a divergência entre os intervalos inferior e superior. Assim, para cada um dos quatro conjuntos, serão considerados cinco níveis de amplitude, totalizando 20 conjuntos de avaliação.

4.2.1 Os Conjuntos

Para esse experimento, serão desenvolvidos quatro conjuntos com quatro características distintas na distribuição dos agrupamentos e como as variáveis se relacionam entre si, e isso servirá como base para os limites. Posteriormente, para a criação do intervalo propriamente dito, cada um desses conjuntos passará por outros pré-processamentos, que criarão as diferenças entre os limites superior e inferior, resultando em um total de cinco graus de variação de amplitude. Ao final, teremos 20 conjuntos distintos, que terão distinção tanto nas variáveis quanto nos limites.

Para criar os quatro conjuntos iniciais, será desenvolvida criação do limite superior do conjunto de pontos de cada classe, em cada conjunto. As configurações, baseadas no Carvalho e Lechevallier (2009b), foram gerados de acordo com uma distribuição normal bi-variada assumindo independência entre as variáveis e vetor de médias, μ, e matriz de covariâncias, Σ, apresentados como:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ e } \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

Consideraremos quatro configurações de dados diferentes de dados quantitativos.

A configuração de dados 1, em que as matrizes de covariância das classes são diagonais, porém diferentes, foi desenvolvida de acordo com os seguintes parâmetros:

- Classe 1: $\mu_1 = 28, \mu_2 = 23, \sigma_1^2 = 144, \sigma_2^2 = 16,$ e $\rho = 0$;
- Classe 2: $\mu_1 = 62, \mu_2 = 30, \sigma_1^2 = 81, \sigma_2^2 = 49,$ e $\rho = 0$;
- Classe 3: $\mu_1 = 50, \mu_2 = 15, \sigma_1^2 = 49, \sigma_2^2 = 81,$ e $\rho = 0$;
- Classe 4: $\mu_1 = 57, \mu_2 = 48, \sigma_1^2 = 16, \sigma_2^2 = 144,$ e $\rho = 0$;

A configuração de dados 2, em que as matrizes de covariância das classes são diagonais e parecidas, foi desenvolvida de acordo com os seguintes parâmetros:

- Classe 1: $\mu_1 = 28, \mu_2 = 23, \sigma_1^2 = 100, \sigma_2^2 = 9,$ e $\rho = 0$;
- Classe 2: $\mu_1 = 62, \mu_2 = 30, \sigma_1^2 = 81, \sigma_2^2 = 16,$ e $\rho = 0$;
- Classe 3: $\mu_1 = 50, \mu_2 = 15, \sigma_1^2 = 100, \sigma_2^2 = 9,$ e $\rho = 0$;
- Classe 4: $\mu_1 = 57, \mu_2 = 37, \sigma_1^2 = 81, \sigma_2^2 = 16,$ e $\rho = 0$;

A configuração de dados 3, onde as matrizes de covariância das classes não são diagonais e são distintas entre si, foi desenvolvida de acordo com os seguintes parâmetros:

- Classe 1: $\mu_1 = 28, \mu_2 = 23, \sigma_1^2 = 144, \sigma_2^2 = 16,$ e $\rho = 0.8$;
- Classe 2: $\mu_1 = 62, \mu_2 = 30, \sigma_1^2 = 81, \sigma_2^2 = 49,$ e $\rho = 0.7$;
- Classe 3: $\mu_1 = 50, \mu_2 = 15, \sigma_1^2 = 49, \sigma_2^2 = 81,$ e $\rho = 0.6$;
- Classe 4: $\mu_1 = 57, \mu_2 = 48, \sigma_1^2 = 16, \sigma_2^2 = 144,$ e $\rho = 0.9$;

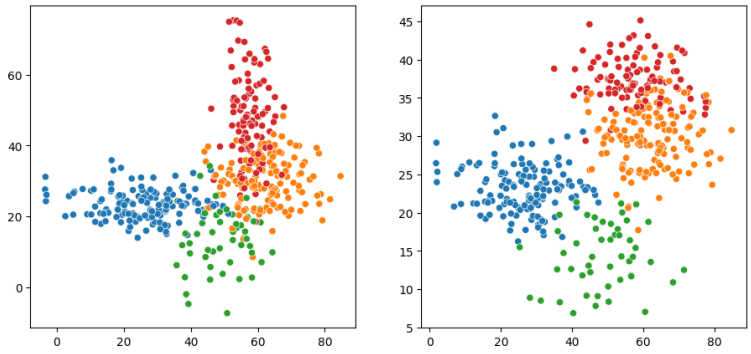
A configuração de dados 4, na qual as matrizes de covariância das classes não são diagonais, mas parecidas, foi desenvolvida de acordo com os seguintes parâmetros:

- Classe 1: $\mu_1 = 28, \mu_2 = 23, \sigma_1^2 = 144, \sigma_2^2 = 16,$ e $\rho = 0.8$;

- Classe 2: $\mu_1 = 62, \mu_2 = 30, \sigma_1^2 = 81, \sigma_2^2 = 49$, e $\rho = 0.7$;
- Classe 3: $\mu_1 = 50, \mu_2 = 15, \sigma_1^2 = 49, \sigma_2^2 = 81$, e $\rho = 0.6$;
- Classe 4: $\mu_1 = 57, \mu_2 = 37, \sigma_1^2 = 16, \sigma_2^2 = 144$, e $\rho = 0.9$;

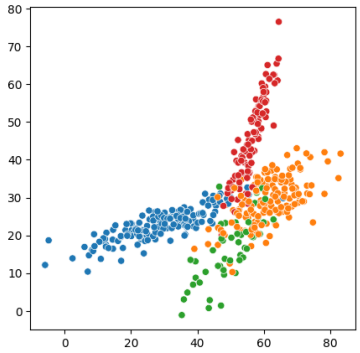
A Figura 3 apresenta como está a representação das classes em cada configuração. Essas configurações serão utilizadas para todas a variações desses conjuntos e a distribuição de faixas será determinado pela criação do limite inferior.

Figura 3 – Configurações Limite Superior para cada Conjunto

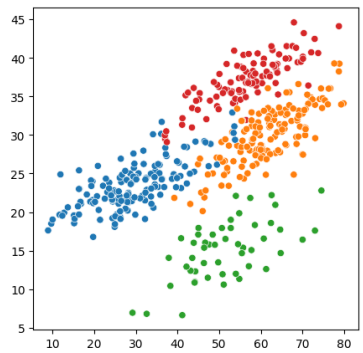


(a) Configuração 1

(b) Configuração 2



(c) Configuração 3



(d) Configuração 4

Fonte: Autor (2024)

No limite inferior será utilizado a distribuição qui-quadrado. Cada um dos 450 pontos serão gerados valores segundo o parâmetro de vetor média λ contendo os valores para cada variável. Para fazer variação de amplitude o fator λ terá cinco variações e aumentando, assim, progressivamente.

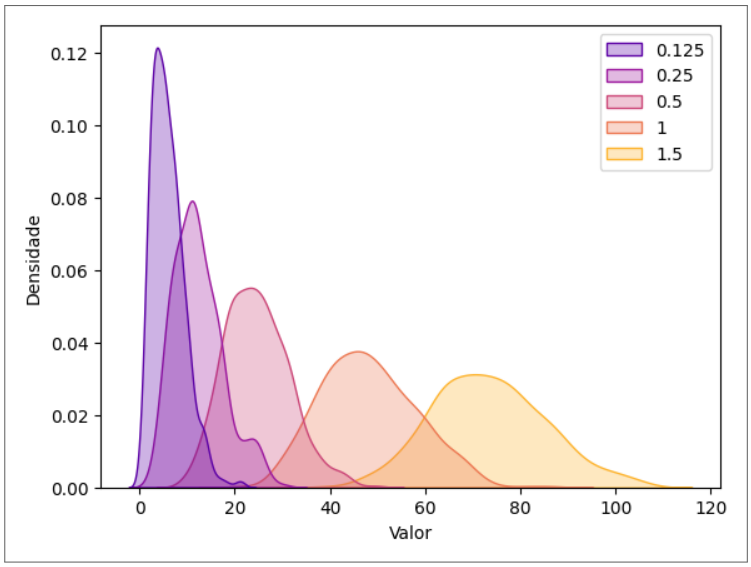
$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

com o cálculo de $x_1^{inf} = x_1^{sup} - \lambda_1$ e $x_2^{inf} = x_2^{sup} - \lambda_2$. Para criar os conjuntos com diferentes amplitudes serão utilizados cinco valores para λ. Essa variação é controlada pelo parâmetro α.

$$\lambda_{p,t} = \alpha_t \left(\max(x_p) - \min(x_p)\right)$$

Onde $\max(x_p)$ representa o valor máximo de x_p e $\min(x_p)$ representa o valor mínimo de x_p na variável p. $\lambda_{p,t}$ é a amplitude da variação para a variável p, controlada pelo parâmetro α_t. O parâmetro α_t pode assumir os valores $\{0.125, 0.25, 0.5, 1, 1.5\}$, permitindo a reprodução de diferentes amplitudes de variação. Na Figura 4, é possível observar como é a distribuição dos valores para um dado $\lambda_{p,t}$.

Figura 4 – Distribuição para cada valor de alfa



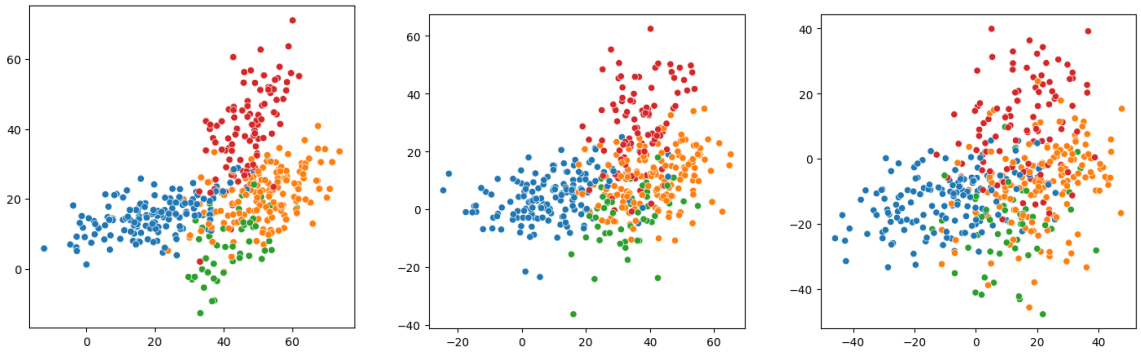
Fonte: Autor (2024)

Com a obtenção dos valores de λ, é possível criar os limites inferiores complementando os limites superiores de cada conjunto de dados simulados. O limite superior permanece constante para cada configuração. Conforme o valor de α aumenta, obtemos uma configuração para o limite inferior cada vez mais irregular e interseccionada, tornando difícil visualizar os limites de cada classe, como podemos observar na Figura 5.

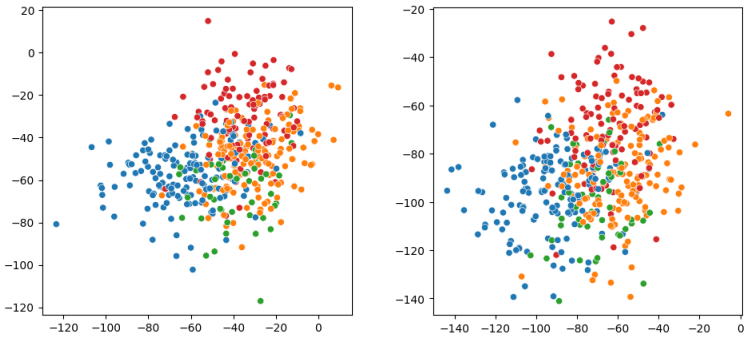
Quando avaliamos os conjuntos unindo os intervalos superior e inferior temos a distribuição dos dados dos conjuntos pronto para execução dos métodos e avaliação, Figura 6.

Na Figura 6a os retângulos intervalares estão relativamente bem agrupados, com sobreposições mais modestas e uma dispersão que ainda permite distinguir diferentes regiões

Figura 5 – Limite Inferior para cada α no Conjunto 3



(a) $\alpha = 0.125$ (b) $\alpha = 0.25$ (c) $\alpha = 0.5$



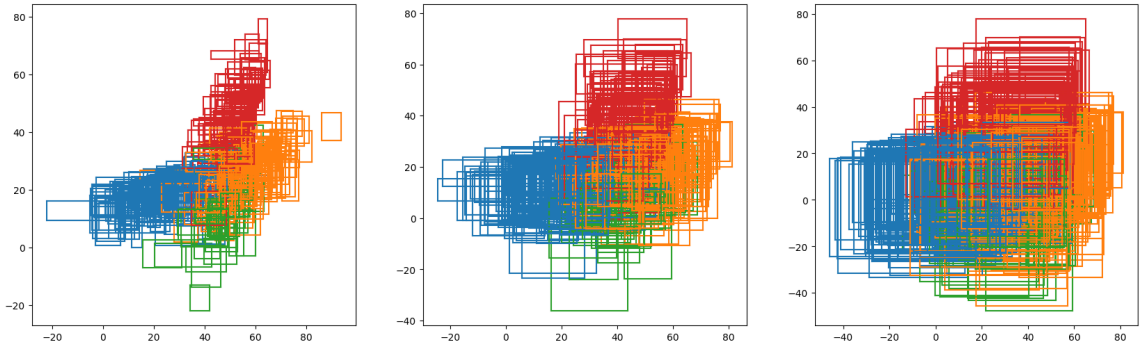
(d) $\alpha = 1$ (e) $\alpha = 1.5$

Fonte: Autor (2024)

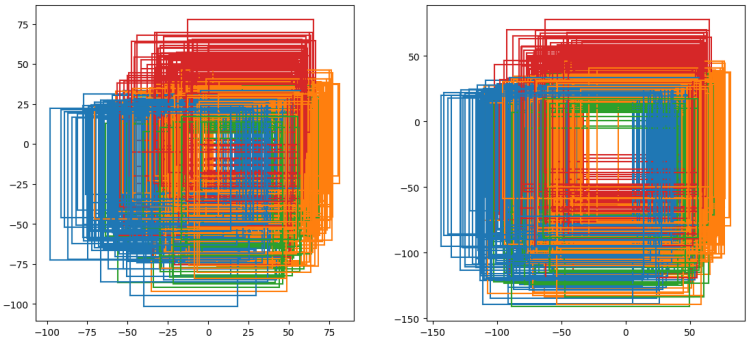
de concentração de pontos. Neste estágio, o conjunto de dados ainda possui uma estrutura menos complexa, onde as distinções entre diferentes grupos são mais claras. Contudo, à medida que o parâmetro aumenta, a sobreposição dos retângulos intervalares se intensifica e a dispersão se expande. Esse aumento de complexidade sugere que os dados começam a se intercalar mais, e as distinções entre diferentes regiões começam a ficar menos evidentes, como exemplificado na Figura 6c Já na última figura, a Figura 6e, o comportamento dos dados atinge um nível ainda maior de complexidade. Os retângulos intervalares estão altamente sobrepostos e espalhados por uma área mais vasta, dificultando a identificação de qualquer estrutura ou padrão claro. A dispersão é máxima, indicando que o conjunto de dados se tornou extremamente intrincado e as distinções entre diferentes grupos são praticamente imperceptíveis.

Além do aumento da complexidade e dispersão, outra característica observável é a extensão das áreas de sobreposição, que cresce significativamente de um conjunto para outro. Isso indica que, conforme o parâmetro aumenta, não apenas a dispersão se amplia, mas também a interação entre diferentes elementos do conjunto de dados se intensifica,

Figura 6 – Configuração 3 para cada alfa



(a) $\alpha = 0.125$ (b) $\alpha = 0.25$ (c) $\alpha = 0.5$



(d) $\alpha = 1$ (e) $\alpha = 1.5$

Fonte: Autor (2024)

levando a um comportamento mais caótico e interligado. O que torna o processo de agrupamento mais desafiador para os métodos.

O que torna os limites inferiores e superiores com comportamento divergentes. Teremos o limite superior na melhor definição possível construídos pelos parâmetros iniciais e o limite inferior vai progressivamente perdendo essa a separabilidade das classes até o ultimo caso.

4.2.2 Os Resultados

Para obter os índices IRA e IMN, foram realizados experimentos de Monte Carlo com 100 replicações. Em cada replicação, os seis métodos de *cluster* são executados até convergirem, conforme o critério de parada de cada método. Essa execução é repetida 50 vezes, e o melhor resultado de cada conjunto de 50 repetições, de acordo com o custo de cada método, é selecionado. Com esses dados, é possível calcular a média, o desvio padrão e outras métricas relevantes para a análise.

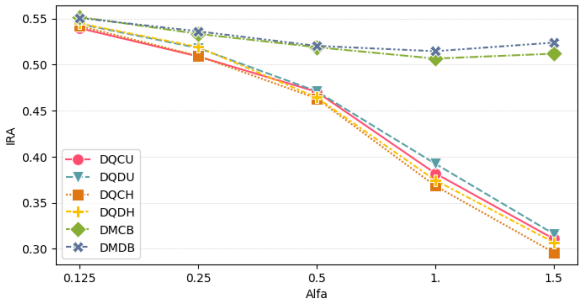
Com as replicações foram obtido os resultados do IRA e IMN através de testes de agrupamento nas quatro configurações iniciais apresentadas anteriormente, utilizando seis métodos diferentes, sendo cinco encontrados na literatura (DQCU, DQDU, DQCH, DQDH, DMDB) e o método proposto DMCB. Esses resultados fornecem uma visão abrangente sobre a eficácia e robustez de cada método em relação à variabilidade do parâmetro α, variando de 0.125 a 1.5.

Após essas execuções foram obtido os seguintes resultados de IRA (Figura 7) e IMN (Figura 8). Para a Configuração 1 tem os resultados IRA representados pela Figura 7a e IMN pela Figura 8a. É possível observar que para α com valor 0.125 todos os métodos estudados apresentam um resultado parecido, próximo a 0.55 tanto de IRA quanto de IMN. A medida que o limite inferior se torna menos definido os métodos baseado em distancia quadrática (DQCU, DQDU, DQCH, DQDH) vão perdendo performance. Até seu pior resultado, com IRA e IMN abaixo de 0.35. Apesar de também apresentar uma perda nos índices de agrupamento essas perdas são muito mais brandas para os métodos com Distância de Mahalanobis (DMCB e DMDB) com a diferença que o método com matriz diagonal (DMDB) apresentando resultado levemente superiores para todos os valores de α. A Configuração 2 apresenta resultado muito próximo aos resultados da configuração 1, representando o IRA pela Figura 7b e o IMN pela Figura 8b.

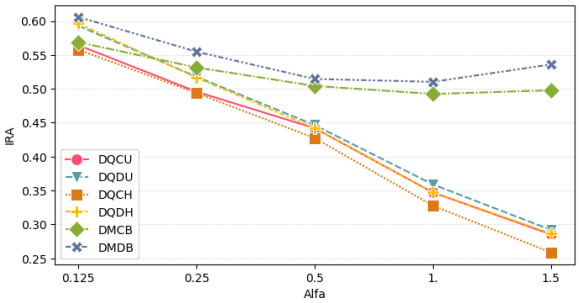
Os resultados da Configuração 3 apresenta o IRA na Figura 7c e o IMN Figura 8c. Para o menor valor de α os métodos que utilizam matriz a covariância no cálculo da distância tiveram melhores resultados (DQCU, DQCH, DMCB). Conforme o grau de dispersão do limite inferior aumenta, representado pelo α, rapidamente os modelo baseado em distância quadrática baixam os seus índices, chegando ao ponto de igualarem os métodos que não utilizam a covariância (DQDU e DQDH). Já o método DMDB inicialmente tem resultados parecidos com os outros método, que assim como ele, mão utilizam a covariância dos dados e fator de dispersão dos limites inferiores não alteram o resultado inicial. O maior destaque positivo vai para o método DMCB que mantêm os melhores resultados ao longo de todas as variações de α demonstrando uma maior desorganização em um dos limites não alteram a sua performance. Configuração 4 tem resultados próximos a configuração 3, como apresentado na Figura 7d para IRA e Figura 8d para o IMN.

Esses resultados permitem observar que o método DMCB tem resultado bons tanto nas configurações sem covariância e com covariância. Já o método DMDB, por não utilizar a covariância nos cálculos de distância, tem sua performance afetada para os conjuntos

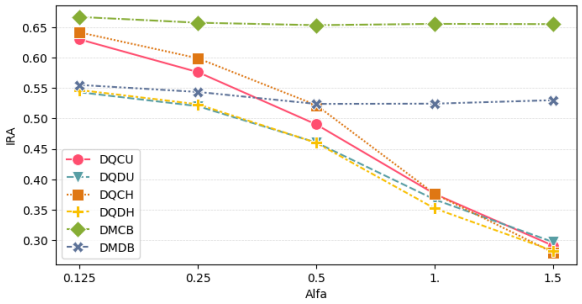
Figura 7 – Resultados IRA



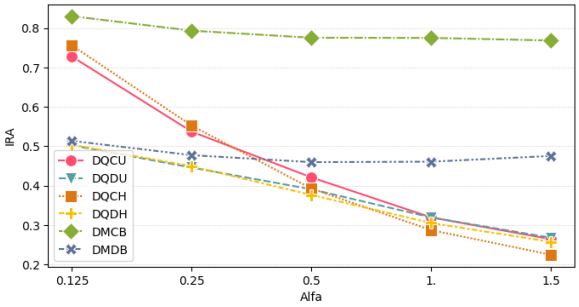
(a) Configuração 1



(b) Configuração 2



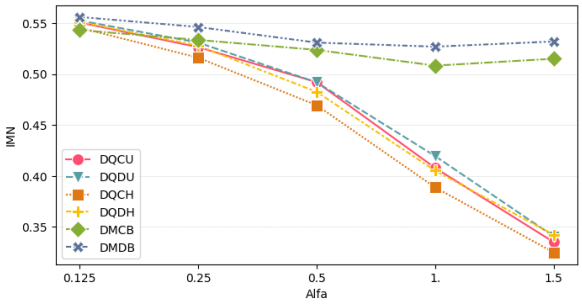
(c) Configuração 3



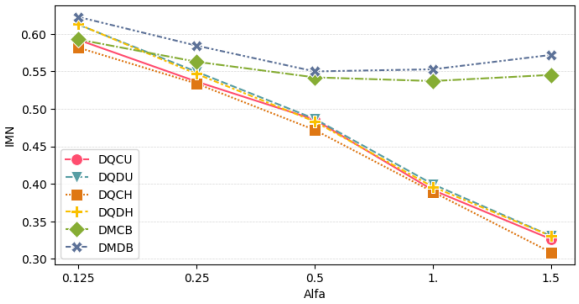
(d) Configuração 4

Fonte: Autor (2024)

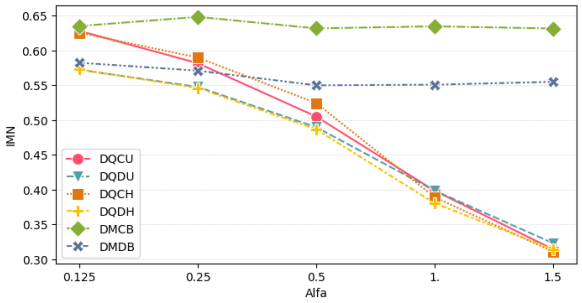
Figura 8 – Resultados IMN



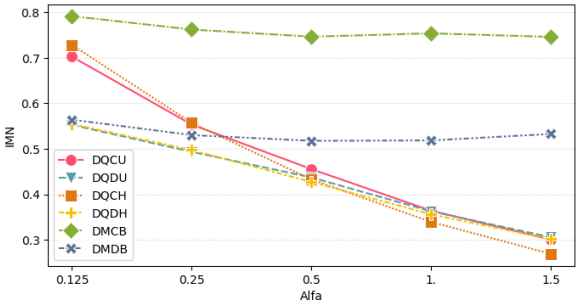
(a) Configuração 1



(b) Configuração 2



(c) Configuração 3



(d) Configuração 4

Fonte: Autor (2024)

com covariância, Conjunto 3 e Conjunto 4. Já demais métodos da literatura, tem grande dificuldade de lidar com situações de turbulência em pelo menos um dos limites.

4.3 CONJUNTO DE DADOS REAIS

Para esta etapa, o método de agrupamento proposto foi aplicado a dados intervalares reais, junto com os demais métodos abordados anteriormente neste trabalho. Para cada conjunto e distância o agrupamento foi repetido 100 vezes. A partição correspondente ao menor critério, segundo cara método, foi selecionada e calculado o ARI e NMI. Os conjuntos selecionados são de dados intervalares e foram aplicados nos artigos abordados ao longo deste trabalho.

4.3.1 Conjunto Temperaturas

O conjunto de dados contém os intervalos de temperatura da cidade apresenta as temperaturas mínimas e máximas mensais das cidades (GURU; KIRANAGI; NAGABHUSHAN, 2004; CARVALHO; LECHEVALLIER, 2009b). Esse conjunto de dados abrange 37 cidades, descritas por 12 variáveis com valores de intervalo. Nesta aplicação, os dados são organizados em quatro grupos distintos, com base em sua latitude e outras características geográfica.

A Tabela 2 mostra os valores do IRA e IMN demonstrando o desempenhos dos métodos sobre o Conjunto Temperatura. Em destaque, apresentam-se as distâncias correspondentes aos melhores desempenhos.

Tabela 2 – Resultado Conjunto Temperatura

Métodos	IRA	IMN
DQCU	0.144	0.090
DQDU	0.462	0.537
DQCH	0.238	0.324
DQDH	0.462	0.537
DMCB	**0.510**	**0.543**
DMDB	0.403	0.511

Fonte: Autor (2024)

Os maus resultados obtidos com os métodos DQCU e DQCH neste conjunto de dados

intervalados ilustram as dificuldades destes métodos em gerenciar a inversão das matrizes que definem as distâncias quadráticas adaptativas nos algoritmos. Assim, é possível notar que os métodos baseados em distâncias quadráticas adaptativas definidas por uma matriz de covariância diagonal (DQDU e DQDH) apresentam resultados melhores. Contudo, o método proposto (DMCB), que apesar de utilizar a matriz de covariância não apresenta as dificuldades que os outros métodos enfrentaram para este conjunto de dados, resultando nos melhores índices para o método.

4.3.2 Conjunto Carros

Este conjunto de dados compreende 33 modelos de automóveis, representados por oito variáveis com valores de intervalo e uma variável nominal (CARVALHO; LECHEVALLIER, 2009b). Para esta aplicação, as oito variáveis com valores de intervalo (Preço, Capacidade do Motor, Velocidade Máxima, Aceleração, Passo, Comprimento, Largura e Altura) foram consideradas para fins de agrupamento. A variável nominal Categoria do Carro foi utilizada como uma variável a priori. Divididos em quatro classes distintas: Utilitário, Berlina, Esportivo e Luxo.

Na Tabela 3 podemos observar os resultados do IRA e IMN demonstrando o desempenhos dos métodos sobre o Conjunto Carro. Em destaque, apresentam-se as distâncias correspondentes aos melhores desempenhos.

Tabela 3 – Resultado Conjunto Carros

Métodos	IRA	IMN
DQCU	0.193	0.350
DQDU	0.589	**0.653**
DQCH	0.262	0.630
DQDH	0.462	0.537
DMCB	**0.608**	0.648
DMDB	0.540	**0.653**

Fonte: Autor (2024)

Assim como no conjunto anterior, o Conjunto Carros também apresentou resultados piores para os métodos baseados na distancia quadrática e que utiliza a matriz de covariância para o calcula da distância (DQCU e DQCH). Ambos os conjuntos tem amostragem pequena de observações favorecendo os modelos que não lidam com covariância (DQCU,

DQDH, DMDB). Contudo, o método proposto (DMCB) consegue lidar com essas situações ainda levando em conta o fator da covariância trazendo atingir o melhor resultado do índice IRA e estar entre os melhores do IMN.

4.3.3 Conjuntos Precipitação

Trata-se de um conjunto constituído por informações climatológicas oficiais e relevantes de diversas cidades ao redor do mundo (FILHO; SOUZA, 2013; SOUZA, 2016). O conjunto abrange um total de 604 cidades, contendo as variáveis de precipitações mínimas e máximas por estação (quatro variáveis no total). Cada cidade possui um clima específico, sendo classificada em uma das seguintes categorias: úmido continental, árido, equatorial, mediterrâneo, monções, oceânico, savana, semiárido, subártico e úmido subtropical. Com essas classificações, formam-se dois grupos para a aplicação de técnicas de agrupamento. O grupo 'Precipitação Europeu' é composto pelas categorias mediterrâneo e oceânico, somando 324 instâncias. Já o conjunto 'Precipitação Global' é a junção do conjunto 'Precipitação Europeu' com outras cidades ao redor do mundo, abrangendo três novas regiões: savana, equatorial e subártica, totalizando 604 observações e cinco classes.

Na Tabela 15 é possível observar os valores de IRA e IMN dos métodos sobre os três conjuntos Precipitação. Em destaque, apresentam-se as distâncias correspondentes aos melhores desempenhos.

Tabela 4 – Resultado Conjuntos Precipitação

Método	**Precipitação Europeu**		**Precipitação Global**	
	IRA	**IMN**	**IRA**	**IMN**
DQCU	0.610	0.641	0.285	0.399
DQDU	0.639	0.676	0.103	0.267
DQCH	0.623	0.661	0.291	0.417
DQDH	**0.643**	**0.680**	0.102	0.430
DMCB	0.623	0.661	**0.350**	**0.468**
DMDB	0.641	0.671	0.107	0.261

Fonte: Autor (2024)

No conjunto 'Precipitação Europeu', todos os métodos apresentaram resultados próximos entre si, com maior relevância para aqueles que não utilizam a covariação para calcular a distância. No entanto, no conjunto 'Precipitação Global', essa realidade muda,

destacando-se os ganhos obtidos pelos métodos que utilizam a matriz de covariância completa, principalmente o método proposto (DMCB).

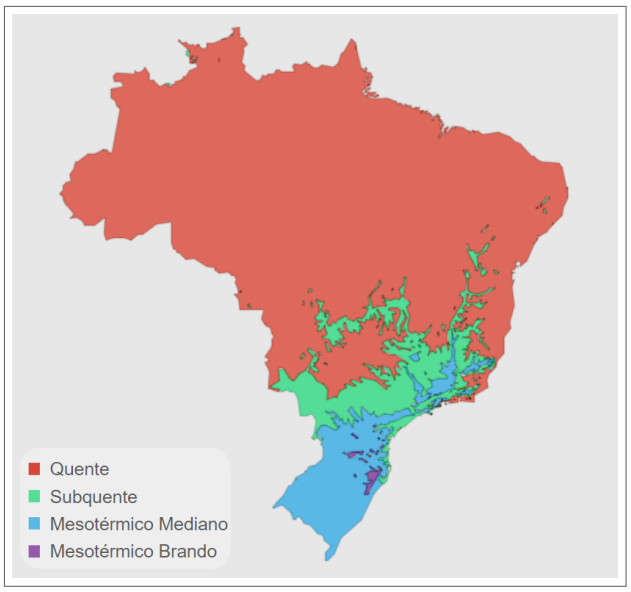
Esses resultados, em conjuntos de dados reais, demostrou que o método proposto obteve progresso nos índices avaliados, em comparação com os métodos da literatura.

5 APLICAÇÃO EM DADOS CLIMÁTICOS

Esta seção tem como objetivo avaliar o agrupamento obtido pelo método proposto (DMCB) da temperatura das estações meteorológicas automáticas. O conjunto de dados será pre-processado para criar o conjunto de dados intervalares das temperaturas de cada estação. Para a avaliação aplicada 100 vezes para usando o que minimiza a função custo.

O Instituto Brasileiro de Geografia e Estatística (IBGE) (2002) classifica as zonas climáticas brasileiras em quatro classes de temperatura: Quente - média $> 18°C$ em todos os meses; Subquente - média entre $15°C$ e $18°C$ em pelo menos 1 mês; Mesotérmico mediano - média $> 10°C$; Mesotérmico brando - média entre $10°C$ e $15°C$ (Figura 9).

Figura 9 – Zonas Temperatura Brasileira



Fonte: IBGE adaptado (2002)

Apesar da classificação vista anteriormente classificar o território brasileiro em quatro conjuntos distintas, pelo fato da área classificada como Mesotérmico Brando ser bastante limitada e baixa representação, os agrupamentos serão aplicados para busca de três *clusters*.

5.1 ÍNDICES DE INTERPRETAÇÃO

Ferramentas de partição e interpretação de *clusters* são essenciais para que os usuários possam entender os resultados de algoritmos de agrupamento. Elas permitem avaliar a heterogeneidade dos dados em diferentes níveis, incluindo a heterogeneidade geral, a

heterogeneidade dentro de cada *cluster* (intra-*cluster*) e entre os *clusters* (inter-*cluster*). Além disso, essas ferramentas ajudam a identificar a contribuição individual de cada variável na formação dos *clusters*, proporcionando uma visão mais detalhada e precisa dos padrões presentes nos dados.

Celeux et al. (1989) foram pioneiros ao introduzir uma família de índices baseados na soma dos quadrados para a interpretação de *clusters* e partições. Carvalho, Brito e Bock (2006) adaptaram esses índices para dados intervalares particionados pelo algoritmo de *d*inâmico, mantendo a distância euclidiana quadrada não adaptativa. Carvalho e Lechevallier (2009b) aplicou essas métricas em seu algoritmo de distancia intervalar baseada em distancia quadrática. Por outro lado, Pimentel e Souza (2016) expandiram esses índices para algoritmos multivariados fuzzy c-means utilizando distâncias euclidianas quadradas adaptativas.

Nesta seção, foi adaptado esses índices para dados com valor de intervalo particionados pelos algoritmos de agrupamento dinâmico com distâncias de Mahalanobis adaptativas apresentados na Seção 3.

Medidas de Dispersão Geral

Seja uma partição $P = \{C_1, \cdots, C_K\}$ de Ω em K *clusters* de tamanho n_k $(k = 1, \cdots, K)$, que foi obtido a partir de um dos algoritmos de agrupamento dinâmico adaptativo apresentados na Seção 3. Seja um elemento do *cluster* $\mathbf{x}_i = (x^1, \cdots, x^p)$, um integrante do protótipo $\mathbf{y}_k = (y_k^1, \cdots, y_k^p)$ e o protótipo geral $\mathbf{y} = (y^1, \cdots, y^p)$.

Dispersão Geral Global: Os dados com valor de intervalo é medida de acordo com a função de distância usada. Podemos medir heterogeneidade geral como a junção da heterogeneidade de cada *cluster*.

$$T = \sum_{k=1}^{K} T_k \quad \text{onde} \quad T_k = \sum_{i \in C_k} d_{M_k}^2 (\mathbf{x}_i, \mathbf{y}) \tag{5.1}$$

Dispersão Geral do cluster: Consideramos a heterogeneidade geral dentro dos *clusters* C_k e a medimos. Ao juntar a dispersão de cada agrupamento temos a dispersão do geral do *cluster*.

$$J = \sum_{k=1}^{K} J_k \quad \text{onde} \quad J_k = \sum_{i \in C_k} d_{M_k}^2 (\mathbf{x}_i, \mathbf{y}_k) \tag{5.2}$$

Dispersão Geral entre clusters: Mede a dispersão dos representantes do textitcluster, ou seja, a distinção de cada *cluster* com o o conjunto, dado por

$$B = \sum_{k=1}^{K} B_k \quad \text{onde} \quad B_k = n_k d^2_{M_k}(\mathbf{y}_k, \mathbf{y}) \tag{5.3}$$

Pode-se facilmente mostrar que, quaisquer que sejam as funções de distância, as seguintes relações são válidas

$$T = J + B$$

Índices de Interpretação

Com as Medidas de Dispersão Geral calculado é possível construir os Índices de Interpretação do agrupamento obtido. Os Índices de Interpretação de Partição interpretam a qualidade geral de uma partição após ter aplicado um algoritmo de agrupamento aos dados é um problema importante na análise de agrupamento.

Índice de Heterogeneidade Geral: A proporção da soma dos quadrados globais explicada pela partição $P = (C_1, \cdots, C_K)$ é definida como

$$R = \frac{B}{T} = 1 - \frac{J}{T} \tag{5.4}$$

Um valor maior de R significa *clusters* mais homogêneos e uma melhor representação do *cluster* pelo seu protótipo.

Índices de Interpretação de *cluster*

Outro problema importante na análise de *cluster* é avaliar a homogeneidade e excentricidade dos *clusters* individuais de uma partição após ter aplicado um algoritmo de agrupamento aos dados.

A proporção dos Dispersão Geral Global relativo a cada *cluster* C_k é dada por

$$T(k) = \frac{T_k}{T} \tag{5.5}$$

A contribuição do *cluster* C_k para as Medidas de Dispersão Geral dentro do *cluster* é dada por:

$$J(k) = \frac{J_k}{J} \tag{5.6}$$

Um valor relativamente grande de $J(k)$ indica que o *cluster* C_k é relativamente heterogêneo em comparação com os outros *clusters*.

A contribuição de um *cluster* C_k para as Medidas de Dispersão Geral entre *clusters* é medida por

$$B\left(k\right) = \frac{B_k}{B} \tag{5.7}$$

Um valor alto de $B\left(k\right)$ indica que o *cluster* C_k está bastante distante do centro global em comparação com a totalidade de todos os *clusters*. Observe que $\sum_{k=1}^{K} T_k = \sum_{k=1}^{K} J_k = \sum_{k=1}^{K} B_k = 1$

5.2 CONJUNTO DE DADOS

O conjunto de dados utilizados para esta aplicação são formados por dados de estações meteorológicas do brasileiros disponibilizadas pelo Instituto Nacional de Meteorologia (portal.inmet.gov.br). Para isso, foram utilizadas o banco de dados histórico das estações automáticas distribuídas ao longo do território brasileiro. Essas estações obtêm dados meteorológicos a cada hora com dados de insolação, umidade, precipitação, temperatura e vento. Para esta avaliação foram utilizados os valores da variável Temperatura de compensada media de bulbo seco, em cada estação dos anos de 2019 a 2023.

Para cada dia do período de 5 anos, calculamos a temperatura média diária utilizando a média das 24 temperaturas horárias registradas ao longo do dia. Seja T_j^h a temperatura horaria na hora j, temos a temperatura média diária de um dia i no mês M, $T_{i,M}^d$ dada por:

$$T_{i,M}^{\mathrm{d}} = \frac{1}{24} \sum_{j=1}^{24} T_j^h$$

Para cada mês M, coletamos todas as temperaturas médias diárias em uma lista T_M^d, definida como:

$$T_M^d = \{T_{i,M}^d \mid i \in \text{dias do mês } M\}$$

Com essa lista, determinamos a temperatura média diária máxima e mínima do mês M. A temperatura média diária máxima do mês M, $T_{M,\text{máx}}$, é obtida como:

$$T_{M,\text{máx}} = \max(T_M^d)$$

Analogamente, a temperatura média diária mínima do mês M, $T_{M,\text{mín}}$, é obtida como:

$$T_{M,\text{mín}} = \min(T_M^d)$$

Os valores obtidos são organizados nos dados como mostra a Tabela 5. Cada estação contém as variáveis intervalares contendo as temperaturas médias diárias máximas e mínimas para cada mês ao longo do período de estudo.

Tabela 5 – Temperaturas médias diárias máximas e máximas mensal para uma estação

Mês	$T_{M,\text{mín}}$	$T_{M,\text{máx}}$
Jan	$T_{1,\text{mín}}$	$T_{1,\text{máx}}$
Fev	$T_{2,\text{mín}}$	$T_{2,\text{máx}}$
Mar	$T_{3,\text{mín}}$	$T_{3,\text{máx}}$
...	...	...
Dez	$T_{12,\text{mín}}$	$T_{12,\text{máx}}$

Fonte: Autor (2024)

Aplicando esses passos para todas as estações, é possível reorganizar os dados da temperatura coletadas pelas estações automáticas em dados intervalares mensais, como apresentado na Tabela 6

Tabela 6 – Conjunto de Temperaturas Médias Diárias Mensais Brasileiras

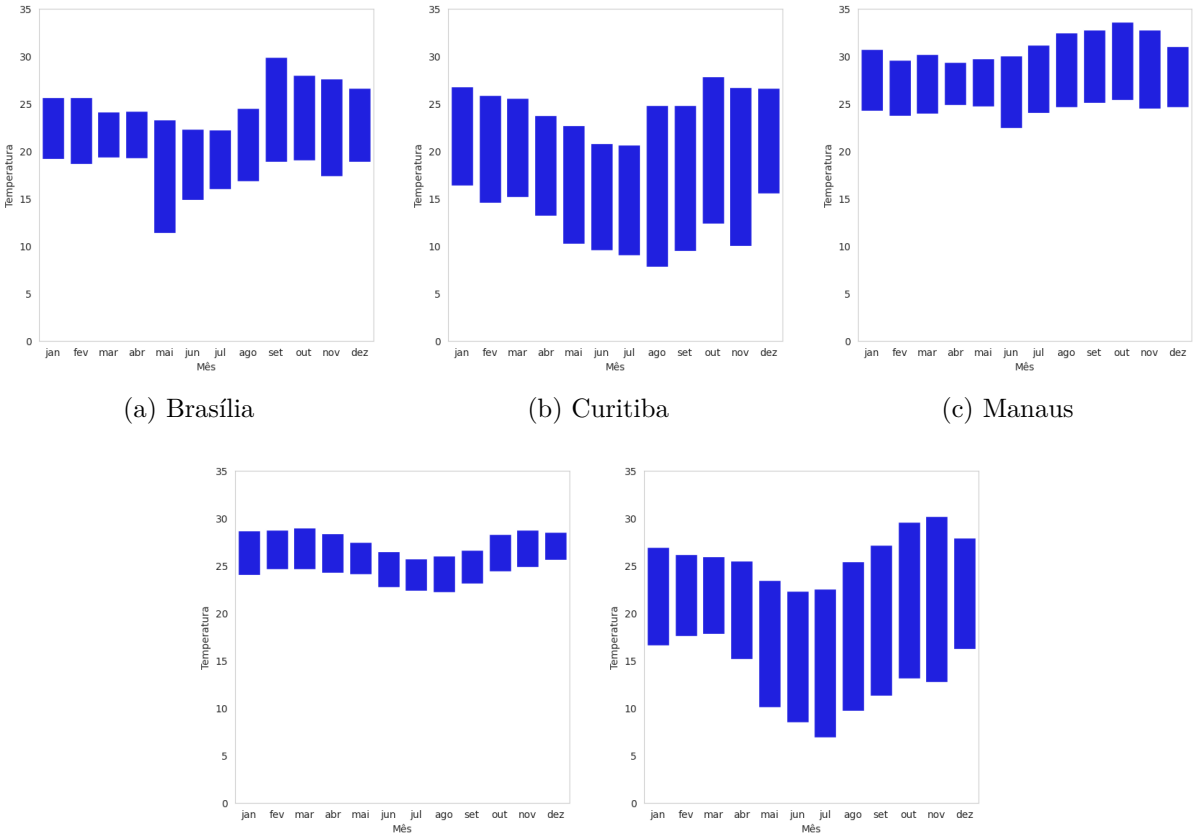
Estação	**Jan**	**Fev**	**...**	**Jun**	**Jul**	**...**	**Dez**
	$[T_{\text{mín}}, T_{\text{máx}}]$	$[T_{\text{mín}}, T_{\text{máx}}]$	...	$[T_{\text{mín}}, T_{\text{máx}}]$	$[T_{\text{mín}}, T_{\text{máx}}]$	...	$[T_{\text{mín}}, T_{\text{máx}}]$
Brasília	[19.1, 25.7]	[18.6, 25.7]	...	[14.8, 22.3]	[16.0, 22.3]	...	[18.9, 26.7]
Curitiba	[16.3, 26.8]	[14.5, 25.9]	...	[9.5, 20.8]	[9.0, 20.7]	...	[15.5, 26.7]
...	...	...	...	...	...	...	...
Manaus	[24.2, 30.8]	[23.7, 29.6]	...	[22.4, 30.0]	[24.0, 31.2]	...	[24.6, 31.0]
...	...	...	...	...	...	...	...
Recife	[24.0, 28.7]	[24.6, 28.8]	...	[22.7, 26.5]	[22.4, 25.7]	...	[25.6, 28.6]
São Paulo	[16.6, 26.9]	[17.6, 26.2]	...	[8.4, 22.4]	[6.9, 22.6]	...	[16.2, 28.0]

Fonte: Autor (2024)

Utilizando o gráfico radar para demostrar todas as variáveis obtidas é possível observar graficamente o comportamento de cada estação, Figura 10. Esses comportamentos permite o estudos e avaliação de padrões climáticos pelo o Brasil.

Ao final, o conjunto conta com 590 estações meteorológicas distribuídas ao longo do território brasileiro. Cruzando as informações geoespaciais das estações é possível delimitar a localização de cada estação utilizada no conjunto, Figura 11. Esses dados geoespaciais

Figura 10 – Temperaturas



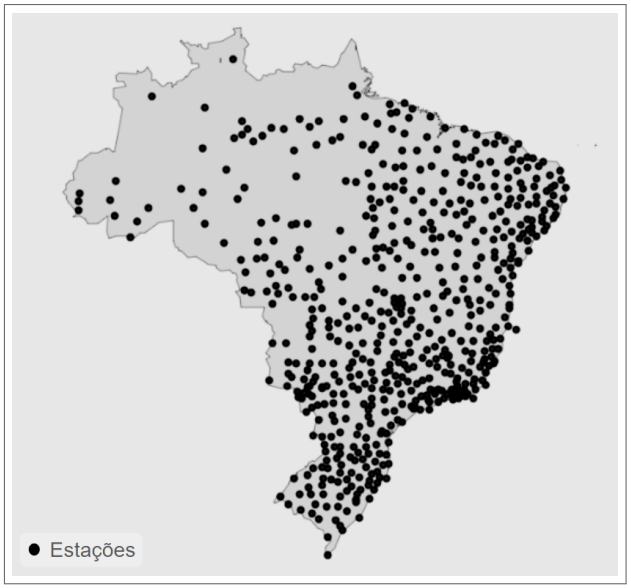
(a) Brasília (b) Curitiba (c) Manaus

(d) Recife (e) São Paulo

Fonte: Autor (2024)

não fazem parte do conjunto de treinamento e são um complemento a visualização do resultado final.

Figura 11 – Distribuição das Estações Meteorológicas Automáticas



Fonte: Autor (2024)

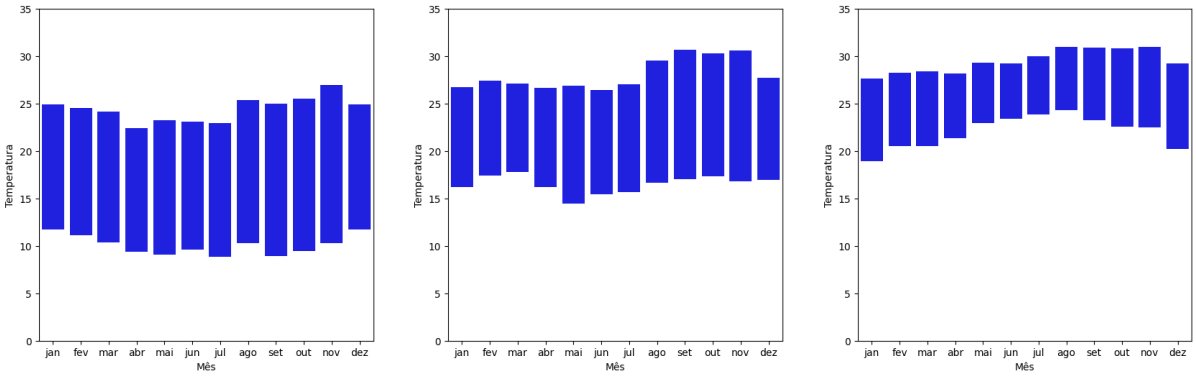
5.3 RESULTADOS

Após aplicação do método proposto no conjunto desenvolvidos das temperaturas máximas e minímas mensais é possível obter as partições e centroides dos representantes para cada agrupamento obtido. *Cluster* 1 com 118 representantes, *Cluster* 2 com 246 representantes e *Cluster* 3 com 226 integrantes.

Os protótipos resultantes da aplicação (Figura 12) permitem observar o comportamento médio de cada partição. No *Cluster* 1 (Figura 12a) o maior diferencial são os menores valores de temperatura de todo o agrupamento, com mínimas próximas a 10°C e máximas em torno de 25°C. No *Cluster* 2 (Figura 12b) os valores intermediários aumentam tanto a temperatura mínima, para 15°C, quanto a temperatura máxima, para 25°C. Já no *Cluster* 3 (Figura 12c) estão localizados os maiores valores de temperatura, com mínimas de 25°C e máximas de 30°C, sendo essa a menor amplitude de valores entre as três partições. Esses agrupamentos correspondem ao esperado para a realidade das cidades brasileiras, onde grande parte possui essas características.

Ao comparamos as estações (Figura 10) da subseção anterior com os centroides da sua partição (Figura 12) nota-se um comportamento similar entre as estações e seus centroides correspondentes. Especificamente, Curitiba (Figura 10b) e São Paulo (Figura 10e) estão no *Cluster* 1 (Figura 12a); Brasília (Figura 10a) está no *Cluster* 2 (Figura 12b); Manaus (Figura 10c) e Recife (Figura 10d) estão no *Cluster* 3 (Figura 12c). Quando

Figura 12 – Centroides

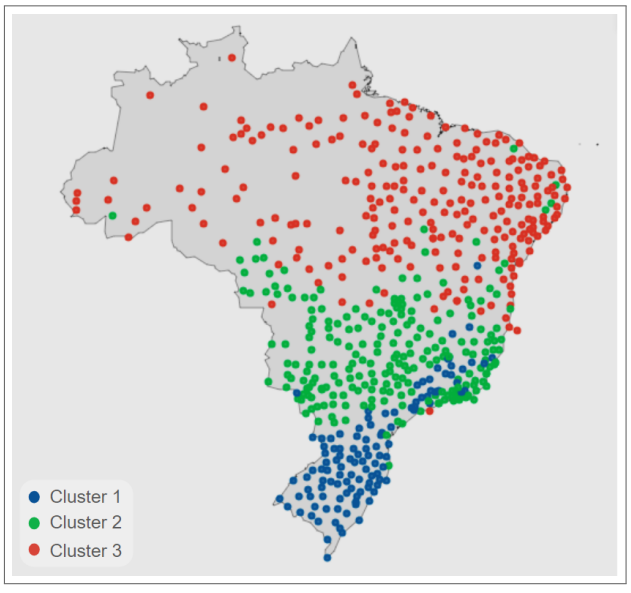


(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fonte: Autor (2024)

cruzamos as partições obtidas com sua geolocalização podemos avaliar a distribuição de cada agrupamento no mapa brasileiro (Figura 13). Nota-se a concentração dos *clusters* em regiões especificas: *Cluster* 1 combindo grande parte da região sul; *Cluster* 2 com maior parte da região sudeste e centro-oeste; *Cluster* 3 com o norte e nordeste.

Figura 13 – Estações Meteorológicas Automáticas Agrupadas



Fonte: Autor (2024)

Partição tem como Índice de Heterogeneidade Geral $R = 0.574$, mostrando que a maior parte da variabilidade do agrupamento é devido à separação dos clusters. Isso indica que a separação tende a ser de boa qualidade. Para avaliarmos a nível de cada agrupamento, podemos ver os Índices de Interpretação de *Cluster* na Tabela 7 Quando avaliamos cada

índice, é possível observar que o *Cluster* 1 é o mais homogêneo, apresentando a menor variabilidade dentro do cluster, $J(1) = 25.3\%$ seguido do *Cluster* 3 com $J(3) = 33.1\%$, e o *Cluster* 2 com maior variabilidade interna, sendo responsável por $J(2) = 41.6\%$. Quando avaliamos o índice relativo aos valores entre clusters, é possível dizer que o *Cluster* 2 está mais próximo ao centro global, com $B(2) = 3.0\%$, enquanto os *clusters* 2 e 3 estão bem distantes e mais bem definidos.

Tabela 7 – Resultado Índices

Cluster	Tamanho	$T(k)$	$J(k)$	$B(k)$
1	118	37.4	25.3	46.3
2	226	19.4	41.6	3.0
3	246	43.2	33.1	50.7

Fonte: Autor (2024)

Quando cruzamos os dados dos índices com a avaliação dos centroides e a distribuição geoespacial das estações podemos perceber que o *Cluster* 2 sofre da situração de ser um *cluster* com valores intermediários entre os já bem delimitados *Cluster* 1 e *Cluster* 3.

6 CONCLUSÃO

Neste trabalho, foi apresentado um novo método para agrupamento de dados intervalares que utiliza a distância de Mahalanobis. Esse método se diferencia dos já existentes no estado da arte por ser especialmente eficaz em lidar com dados que apresentam assimetria de comportamento entre os limites inferior e superior. Ao tratar as variâncias e covariâncias de ambos os limites de forma independente, o método otimiza os resultados dos agrupamentos, proporcionando maior precisão, mesmo em cenários com grande disparidade entre os limites.

Para avaliar essas características, foram desenvolvidos diferentes conjuntos de dados intervalares com graus de dispersão progressivos. O método proposto (DMCB), em comparação com cinco algoritmos da literatura que têm abordagem similares, obteve os melhores resultados tanto em conjuntos com dependência quanto em independência entre variáveis.

O método foi aplicado em um conjunto de dados reais de conjuntos climáticos. Foi possível avaliar que o método proposto (DMCB) foi capaz de encontrar padrões e agrupar os dados climáticos compatíveis com a classificação desenvolvida pelo IBGE.

6.1 TRABALHOS FUTUROS

Durante o desenvolvimento deste trabalho, foram identificadas várias oportunidades para pesquisas futuras. Entre elas, está a possibilidade de aplicar o método proposto a diferentes conjuntos de dados, explorando outras distribuições de probabilidade, a fim de avaliar o comportamento do modelo em uma variedade de cenários. Além disso, o método pode ser testado em áreas como controle de qualidade, análise de risco e finanças, onde dados intervalares são frequentemente encontrados. Outra linha de investigação relevante seria comparar o desempenho do agrupamento utilizando dados intervalares e dados pontuais clássicos, para identificar possíveis ganhos de precisão, robustez e eficiência.

Outros possíveis estudos é adaptar o método a outros tipos de dados simbólicos com formato *boxplot*, histograma ou dados poligonais.

BIBLIOGRAFIA

ABONYI, J.; FEIL, B. *Cluster analysis for data mining and system identification*. [S.l.]: Springer Science & Business Media, 2007.

BERKHIN, P. A survey of clustering data mining techniques. In: *Grouping multidimensional data: Recent advances in clustering*. [S.l.]: Springer, 2006. p. 25–71.

BILLARD, L. Symbolic data analysis: what is it? In: SPRINGER. *Compstat 2006-Proceedings in Computational Statistics: 17th Symposium Held in Rome, Italy, 2006*. [S.l.], 2006. p. 261–269.

BOCK, H.-H.; DIDAY, E. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. [S.l.]: Springer Science & Business Media, 2012.

CARVALHO, F. d. A. D.; BRITO, P.; BOCK, H.-H. Dynamic clustering for interval data based on l 2 distance. *Computational Statistics*, Springer, v. 21, p. 231–250, 2006.

CARVALHO, F. d. A. D.; LECHEVALLIER, Y. Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition*, Elsevier, v. 42, n. 7, p. 1223–1236, 2009.

CARVALHO, F. d. A. de; LECHEVALLIER, Y. Dynamic clustering of interval-valued data based on adaptive quadratic distances. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, IEEE, v. 39, n. 6, p. 1295–1306, 2009.

CELEUX, G.; DIDAY, E.; GOVAERT, G.; LECHEVALLIER, Y.; RALAMBONDRAINY, H. *Classification automatique des données*. [S.l.]: Dunod Paris, 1989.

CHAVENT, M.; LECHEVALLIER, Y. Dynamical clustering of interval data: optimization of an adequacy criterion based on hausdorff distance. In: *Classification, Clustering, and Data Analysis: Recent Advances and Applications*. [S.l.]: Springer, 2002. p. 53–60.

DIDAY, E. Introduction à l'approche symbolique en analyse des données. *RAIRO-Operations Research*, EDP Sciences, v. 23, n. 2, p. 193–236, 1989.

DIDAY, E. Thinking by classes in data science: the symbolic data analysis paradigm. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 8, n. 5, p. 172–205, 2016.

DIDAY, E.; NOIRHOMME-FRAITURE, M. *Symbolic data analysis and the SODAS software*. [S.l.]: John Wiley & Sons, 2008.

EVERITT LANDAU, L.; STAHL. *Cluster Analysis*. [S.l.]: Wiley, 2011.

FILHO, T. d. M. e S.; SOUZA, R. M. Fuzzy learning vector quantization approaches for interval data. In: IEEE. *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.], 2013. p. 1–8.

GOVAERT, G. Classification automatique et distances adaptatives. *These de*, 1975.

GURU, D.; KIRANAGI, B. B.; NAGABHUSHAN, P. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. *Pattern Recognition Letters*, Elsevier, v. 25, n. 10, p. 1203–1213, 2004.

Instituto Brasileiro de Geografia e Estatística (IBGE). *Clima*. 2002. Acesso em: 2 mai. 2024. Disponível em: <https://www.ibge.gov.br/geociencias/informacoes-ambientais/climatologia/15817-clima.html>.

IRPINO, A.; VERDE, R. Dynamic clustering of interval data using a wasserstein-based distance. *Pattern Recognition Letters*, Elsevier, v. 29, n. 11, p. 1648–1658, 2008.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, Acm New York, NY, USA, v. 31, n. 3, p. 264–323, 1999.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 2009.

KUMAR, V. *Chapman & Hall/CRC data mining and knowledge discovery series*. [S.l.]: Chapman & Hall/CRC, 2007.

MAO, J.; JAIN, A. K. A self-organizing network for hyperellipsoidal clustering (hec). *Ieee transactions on neural networks*, IEEE, v. 7, n. 1, p. 16–29, 1996.

PENG, W.; LI, T. Interval data clustering with applications. In: IEEE. *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*. [S.l.], 2006. p. 355–362.

PIMENTEL, B. A.; SOUZA, R. M. de. Multivariate fuzzy c-means algorithms with weighting. *Neurocomputing*, Elsevier, v. 174, p. 946–965, 2016.

RODRÍGUEZ, S. I. R.; CARVALHO, F. d. A. T. de. Clustering interval-valued data with adaptive euclidean and city-block distances. *Expert Systems with Applications*, Elsevier, v. 198, p. 116774, 2022.

SOUZA, L. C. d. Agrupamento e regressão linear de dados simbólicos intervalares baseados em novas representações. Universidade Federal de Pernambuco, 2016.

SOUZA, L. C. de; SOUZA, R. M. C. R. D.; AMARAL, G. J. A. do. Dynamic clustering of interval data based on hybrid l q distance. *Knowledge and Information Systems*, Springer, v. 62, n. 2, p. 687–718, 2020.

SOUZA, R. M. de; CARVALHO, F. A. de; TENORIO, C. P. Two partitional methods for interval-valued data using mahalanobis distances. In: SPRINGER. *Advances in Artificial Intelligence–IBERAMIA 2004: 9th Ibero-American Conference on AI, Puebla, Mexico, November 22-26, 2004. Proceedings 9*. [S.l.], 2004. p. 454–463.

SOUZA, R. M. de; CARVALHO, F. de AT de; SILVA, F. C. Clustering of interval-valued data using adaptive squared euclidean distances. In: SPRINGER. *Neural Information Processing: 11th International Conference, ICONIP 2004, Calcutta, India, November 22-25, 2004. Proceedings 11*. [S.l.], 2004. p. 775–780.

SUN, L.; ZHU, L.; LI, W.; ZHANG, C.; BALEZENTIS, T. Interval-valued functional clustering based on the wasserstein distance with application to stock data. *Information Sciences*, Elsevier, v. 606, p. 910–926, 2022.

XU, W. *Symbolic data analysis: interval-valued data regression*. Tese (Doutorado) — University of Georgia Athens, GA, 2010.

APÊNDICE A – RESULTADOS DADOS SINTÉTICOS - IRA

Tabela 8 – Resultado IRA Conjunto 1

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.540 (0.047)	0.509 (0.040)	0.471 (0.035)	0.382 (0.042)	0.311 (0.039)
DQDU	0.544 (0.043)	0.518 (0.038)	0.471 (0.039)	0.393 (0.041)	0.316 (0.037)
DQCH	0.542 (0.043)	0.509 (0.034)	0.463 (0.039)	0.368 (0.039)	0.296 (0.030)
DQDH	0.545 (0.043)	0.519 (0.035)	0.464 (0.038)	0.375 (0.042)	0.307 (0.036)
DMCB	0.552 (0.040)	0.533 (0.038)	0.519 (0.035)	0.507 (0.040)	0.512 (0.043)
DMDB	0.551 (0.044)	0.536 (0.039)	0.520 (0.038)	0.515 (0.040)	0.524 (0.044)

Fonte: Autor (2024)

Tabela 9 – Resultado IRA Conjunto 2

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.564 (0.079)	0.496 (0.074)	0.442 (0.049)	0.347 (0.038)	0.285 (0.034)
DQDU	0.593 (0.067)	0.518 (0.065)	0.446 (0.048)	0.359 (0.041)	0.292 (0.030)
DQCH	0.558 (0.081)	0.494 (0.070)	0.427 (0.045)	0.328 (0.040)	0.259 (0.039)
DQDH	0.596 (0.064)	0.516 (0.064)	0.442 (0.044)	0.347 (0.041)	0.287 (0.037)
DMCB	0.569 (0.081)	0.531 (0.075)	0.504 (0.062)	0.492 (0.056)	0.498 (0.067)
DMDB	0.606 (0.067)	0.555 (0.079)	0.515 (0.058)	0.510 (0.067)	0.536 (0.076)

Fonte: Autor (2024)

Tabela 10 – Resultado IRA Conjunto 3

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.630 (0.044)	0.576 (0.043)	0.490 (0.050)	0.376 (0.050)	0.291 (0.039)
DQDU	0.543 (0.044)	0.520 (0.044)	0.460 (0.041)	0.367 (0.046)	0.297 (0.043)
DQCH	0.641 (0.043)	0.599 (0.042)	0.522 (0.048)	0.376 (0.051)	0.280 (0.044)
DQDH	0.547 (0.044)	0.523 (0.043)	0.460 (0.041)	0.352 (0.042)	0.283 (0.038)
DMCB	0.667 (0.040)	0.657 (0.048)	0.653 (0.045)	0.655 (0.044)	0.655 (0.037)
DMDB	0.555 (0.042)	0.544 (0.042)	0.524 (0.038)	0.524 (0.042)	0.530 (0.046)

Fonte: Autor (2024)

Tabela 11 – Resultado IRA Conjunto 4

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.727 (0.084)	0.537 (0.099)	0.421 (0.051)	0.319 (0.041)	0.264 (0.032)
DQDU	0.502 (0.075)	0.446 (0.073)	0.391 (0.049)	0.319 (0.038)	0.268 (0.031)
DQCH	0.756 (0.079)	0.552 (0.095)	0.392 (0.061)	0.287 (0.036)	0.224 (0.033)
DQDH	0.504 (0.077)	0.449 (0.073)	0.376 (0.053)	0.305 (0.043)	0.257 (0.031)
DMCB	0.830 (0.039)	0.793 (0.088)	0.775 (0.114)	0.775 (0.105)	0.768 (0.109)
DMDB	0.514 (0.071)	0.477 (0.072)	0.459 (0.064)	0.461 (0.067)	0.475 (0.068)

Fonte: Autor (2024)

APÊNDICE B – RESULTADOS DADOS SINTÉTICOS - IMN

Tabela 12 – Resultado IMN Conjunto 1

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.550 (0.035)	0.526 (0.026)	0.492 (0.028)	0.408 (0.031)	0.335 (0.031)
DQDU	0.552 (0.032)	0.531 (0.027)	0.492 (0.029)	0.420 (0.030)	0.341 (0.030)
DQCH	0.545 (0.039)	0.516 (0.036)	0.469 (0.042)	0.389 (0.037)	0.325 (0.033)
DQDH	0.551 (0.033)	0.528 (0.026)	0.482 (0.028)	0.405 (0.028)	0.342 (0.030)
DMCB	0.543 (0.049)	0.533 (0.042)	0.524 (0.037)	0.508 (0.045)	0.515 (0.038)
DMDB	0.556 (0.033)	0.546 (0.029)	0.531 (0.029)	0.527 (0.029)	0.532 (0.033)

Fonte: Autor (2024)

Tabela 13 – Resultado IMN Conjunto 2

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.592 (0.057)	0.536 (0.051)	0.486 (0.035)	0.392 (0.030)	0.326 (0.033)
DQDU	0.612 (0.049)	0.549 (0.046)	0.487 (0.035)	0.399 (0.029)	0.331 (0.026)
DQCH	0.582 (0.061)	0.534 (0.051)	0.472 (0.042)	0.389 (0.034)	0.309 (0.038)
DQDH	0.613 (0.048)	0.547 (0.046)	0.483 (0.032)	0.396 (0.030)	0.330 (0.038)
DMCB	0.592 (0.060)	0.563 (0.056)	0.542 (0.050)	0.537 (0.047)	0.545 (0.049)
DMDB	0.622 (0.050)	0.584 (0.056)	0.550 (0.043)	0.553 (0.050)	0.572 (0.057)

Fonte: Autor (2024)

Tabela 14 – Resultado IMN Conjunto 3

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.628 (0.031)	0.582 (0.029)	0.505 (0.038)	0.399 (0.039)	0.315 (0.033)
DQDU	0.572 (0.034)	0.548 (0.033)	0.490 (0.031)	0.399 (0.036)	0.323 (0.039)
DQCH	0.626 (0.041)	0.590 (0.040)	0.525 (0.039)	0.389 (0.044)	0.311 (0.038)
DQDH	0.573 (0.034)	0.546 (0.032)	0.487 (0.031)	0.381 (0.034)	0.314 (0.032)
DMCB	0.635 (0.057)	0.648 (0.036)	0.632 (0.047)	0.635 (0.042)	0.632 (0.042)
DMDB	0.583 (0.031)	0.571 (0.031)	0.550 (0.031)	0.551 (0.034)	0.555 (0.036)

Fonte: Autor (2024)

Tabela 15 – Resultado IMN Conjunto 4

Método	0.125	0.250	0.500	1.000	1.500
	Média (Desvio Padrão)				
DQCU	0.703 (0.063)	0.553 (0.070)	0.455 (0.037)	0.364 (0.030)	0.301 (0.030)
DQDU	0.553 (0.061)	0.493 (0.063)	0.437 (0.038)	0.363 (0.030)	0.306 (0.028)
DQCH	0.728 (0.061)	0.557 (0.074)	0.433 (0.048)	0.339 (0.036)	0.269 (0.037)
DQDH	0.555 (0.064)	0.497 (0.063)	0.427 (0.042)	0.356 (0.033)	0.301 (0.030)
DMCB	0.792 (0.043)	0.762 (0.074)	0.747 (0.095)	0.754 (0.087)	0.746 (0.088)
DMDB	0.564 (0.060)	0.530 (0.062)	0.518 (0.059)	0.518 (0.060)	0.533 (0.062)

Fonte: Autor (2024)