UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE CIÊNCIAS APLICADAS

JOÃO VICTOR SANTANA

COMPARAÇÃO DE DESEMPENHO ENTRE AS META-HEURÍSTICAS ALGORITMO GENÉTICO, ENXAME DE PARTÍCULAS E COLÔNIA DE FORMIGAS PARA O AGRUPAMENTO DE BOVINOS

Limeira, SP
2024

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE CIÊNCIAS APLICADAS

JOÃO VICTOR SANTANA

COMPARAÇÃO DE DESEMPENHO ENTRE AS META-HEURÍSTICAS ALGORITMO GENÉTICO, ENXAME DE PARTÍCULAS E COLÔNIA DE FORMIGAS PARA O AGRUPAMENTO DE BOVINOS

Dissertação apresentada à Faculdade de Ciências Aplicadas da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia de Produção e de Manufatura, na área de Pesquisa Operacional e Gestão de Processos.

Orientadora: Profª. Drª. Carla Taviane Lucke da Silva Ghidini

Coorientador: Prof. Dr. Washington Alves de Oliveira

Limeira, SP
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca da Faculdade de Ciências Aplicadas
Ana Luiza Clemente de Abreu Valério - CRB 8/10669

Santana, João Victor, 1999-
Sa59c Comparação de desempenho entre as meta-heurísticas Algoritmo Genético, Enxame de Partículas e Colônia de Formigas para agrupamento de bovinos / João Victor Santana. – Limeira, SP : [s.n.], 2024.

Orientador: Carla Taviane Lucke da Silva Ghidini.
Coorientador: Washington Alves de Oliveira.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Faculdade de Ciências Aplicadas.

1. Análise por agrupamento. 2. Bovino. 3. Meta-heurística. I. Ghidini, Carla Taviane Lucke da Silva, 1976-. II. Oliveira, Washington Alves de, 1977-. III. Universidade Estadual de Campinas (UNICAMP). Faculdade de Ciências Aplicadas. IV. Título.

Informações Complementares

Título em outro idioma: Comparison of performance between meta-heuristics Genetic Algorithm, Particle Swar and Ant Colony for cattle grouping
Palavras-chave em inglês:
Cluster analysis
Cattle
Metaheuristic
Área de concentração: Pesquisa Operacional e Gestão de Processos
Titulação: Mestre em Engenharia de Produção e de Manufatura
Banca examinadora:
Carla Taviane Lucke da Silva Ghidini [Orientador]
Eduardo Machado Silva
Priscila Cristina Berbet Rampazzo
Data de defesa: 19-06-2024
Programa de Pós-Graduação: Engenharia de Produção e de Manufatura

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-8627-4023
- Currículo Lattes do autor: https://lattes.cnpq.br/4583092109395299

BANCA EXAMINADORA – DISSERTAÇÃO DE MESTRADO

Candidato: João Victor Santana
Data da defesa: 19 de junho de 2024

Título da dissertação: Comparação de desempenho entre as meta-heurísticas Algoritmo Genético, Enxame de Partículas e Colônia de Formigas para o agrupamento de bovinos

Prof(a) Dr(a). Carla Taviane Lucke da Silva Ghidini (Presidente, FCA/UNICAMP)

Prof. Dr. Eduardo Machado Silva (UNIFESP/São José dos Campos)

Prof(a). Dr(a). Priscila Cristina Berbert Rampazzo (FCA/UNICAMP)

A ata da defesa, com as respectivas assinaturas dos membros da Banca Examinadora, encontra-se no processo de vida acadêmica do aluno.

Às minhas Tecas: Maria Itelveina Santana e Teresinha Santana.

AGRADECIMENTOS

Agradeço primeiramente a Deus por nunca ter me abandonado, sendo sempre meu ponto de apoio nos momentos difíceis e por me proporcionar inúmeras experiências de crescimento pessoal. Agradeço à minha mãe, Maria Itelvina, por ter sempre me apoiado em todos os sonhos e sonhado junto comigo. Agradeço também à minha tia, Teresinha, por ter me incentivado a correr atrás dos sonhos.

Aos meus irmãos, que hoje dividem este momento de alegria comigo. Aos meus amigos de infância, Ananda, Thiago e outros, que me acompanharam durante toda esta jornada e por todos os momentos vividos com vocês. À minha prima e amiga, Eriane, por um dia ter me acordado para ir fazer o vestibular e ter se tornado minha irmã durante essa caminhada. E também a todos os demais amigos feitos durante esse tempo.

A todos os amigos feitos na sala da pós graduação, em especial ao Gabriel e Jhon Jairo, que tanto me ajudaram e foram parceiros. E também a todos os demais amigos feitos no campus da universidade, que contribuíram para meu crescimento acadêmico, profissional e pessoal.

Gostaria também de agradecer aos vários professores que ministraram aulas durante essa etapa da minha formação e compartilharam um pouco de seu conhecimento comigo. Também agradeço aos professores Carla Ghidini e Washington Oliveira, os quais me orientaram durante esse trabalho e contribuíram para que eu admirasse ainda mais a área de Pesquisa Operacional.

Não menos importantes, agradeço também a toda equipe da FCA, em especial as tias da limpeza e do restaurante universitário, que contribuíram comigo por meio de gestos de carinho e conversas pelos corredores da universidade.

RESUMO

A classificação e o agrupamento de bovinos no pasto com base em características similares é de extrema importância para a gestão eficaz do rebanho, permitindo a identificação de padrões de comportamento, necessidades de cuidados específicos e otimização da produção. Este trabalho se concentra na aplicação de três meta-heurísticas, a saber, Algoritmos Genéticos (AG), Otimização por Enxame de Partículas (PSO) e Otimização por Colônia de Formigas (ACO) para realizar a tarefa de agrupamento de dados relacionados aos bovinos. O objetivo principal consiste em avaliar o desempenho dessas abordagens na formação de agrupamentos, identificando grupos de bovinos com características semelhantes. Para isso, foram realizadas análises comparativas entre os resultados obtidos por meio das meta-heurísticas utilizando-se diferentes métricas de distância, bem como a avaliação da qualidade dos agrupamentos obtidos por meio de métricas de validação de cluster e análise visual dos agrupamentos. Nos testes computacionais, as meta-heurísticas demonstraram sua capacidade de realizar com sucesso o agrupamento dos conjuntos de dados *benchmarks*, tendo sido possível confirmar a qualidade e coerência dos agrupamentos obtidos por meio das métricas de avaliação e a visualização dos resultados, bem como constatar que a distribuição e a densidade dos dados desempenham um papel crucial na eficiência das meta-heurísticas utilizadas, sendo a distribuição o fator de maior influência. Para o agrupamento dos bovinos, foram considerados treze atributos, dos quais três são medidos diretamente no bovino, quatro são por ultrassonografia e seis são atribuidas por meio de avaliação profissional. A definição do melhor número de cluster e da meta-heurística com os melhores resultados ocorreu por meio da análise dos índices de avaliação, sendo dois a melhor quantidade de clusters e identificado que a meta-heurística PSO utilizando a métrica de distância de Manhattan obteve os resultados mais promissores para o particionamento dos dados, tendo conseguido identificar padrões e diferenças significativas entre os clusters, resultando em agrupamentos coesos e interpretáveis.

Palavras-chave: Agrupamento de dados. Bovinos. Meta-heurísticas.

ABSTRACT

Classifying and clustering cattle in pasture based on similar characteristics is of utmost importance for effective herd management, allowing for the identification of behavior patterns, specific care needs, and production optimization. The research in question focuses on the application of three metaheuristic algorithms, namely Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO), using different distance metrics to perform data clustering related to cattle. The main objective of this study is to evaluate the performance of these approaches in forming clusters, identifying groups of cattle with similar characteristics. Throughout the research, comparative analyses were conducted between the results obtained through the metaheuristic algorithms using different distance metrics, as well as the evaluation of the quality of the obtained clusters through cluster validation metrics and visual analysis. In computational tests, the metaheuristic algorithms demonstrated their ability to successfully cluster benchmark datasets, confirming the quality and coherence of the obtained clusters through evaluation metrics and visualization of results, while also noting that data distribution and density play a crucial role in the efficiency of the used metaheuristics, with distribution being the most influential factor. For the grouping of cattle, thirteen attributes were considered, of which three are measured directly on the cattle, four are by ultrasound and six are attributed through professional assessment. The definition of the best cluster number and the meta-heuristic with the best results occurs through the analysis of the evaluation indices, with two being the best number of clusters and and identified as the PSO meta-heuristic using the Manhattan distance metric. obtained the most promising results for data partitioning, having managed to identify patterns and significant differences between the clusters, resulting in cohesive and interpretable groupings.

Keywords: Clustering. Bovines. Metaheuristics.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

AG	Algoritmo Genético
ACO	Colônia de Formigas
AOL	Área de olho de lombo
C	Conformação
CE	Circunferência escrotal
EGP	Espessura de gordura na picanha
EGS	Espessura de gordura subcutânea
GIM	Gordura intramuscular
IA	Inteligência Artificial
IF	Idade
M	Musculosidade
ML	*Machine Learning*
PCA	Análise de componentes principais
PL	Pelagem
PR	Precocidade
PS	Peso
PSO	Enxame de Partículas
R	Característica racial
U	Altura do umbigo

SUMÁRIO

1 INTRODUÇÃO

Os avanços tecnológicos têm transformado radicalmente o mundo em que vivemos, impactando todos os aspectos da nossa sociedade, desde a forma como nos comunicamos até como realizamos tarefas diárias e interagimos com o ambiente ao nosso redor. Esses avanços têm sido impulsionados pela rápida evolução da ciência, resultando em inovações que moldam o presente e definem o futuro em uma velocidade surpreendente. O que era considerado inimaginável há algumas décadas tornou-se realidade hoje, com áreas como Inteligência Artificial, Robótica, Internet das Coisas, Realidade Virtual, Nanotecnologia, Biotecnologia e muitas outras sendo palco de descobertas e invenções revolucionárias (Susskind, 2020; Wang & Siau, 2019).

A inteligência artificial (IA), por exemplo, tem avançado a passos largos, permitindo que máquinas realizem tarefas que antes eram exclusivas dos seres humanos. Algoritmos sofisticados e redes neurais artificiais são capazes de aprender, tomar decisões e resolver problemas complexos com uma eficiência impressionante. Esses avanços tecnológicos têm implicações significativas em vários setores, como saúde, educação, transporte, energia, entretenimento e muitas outras (Ullah et al., 2020).

Aprendizado de Máquina ou Machine Learning (ML) é uma área da IA em que algoritmos computacionais são projetados para emular a inteligência humana, aprendendo com o ambiente ao redor e melhorando gradualmente sua precisão. Em Aprendizagem de Máquina, os modelos desenvolvidos analisam dados, aprendem com eles e depois aplicam o que aprenderam para alcançar o objetivo final ao qual foram implantados, tendo como aspecto importante o processo iterativo de aprendizado, em que os modelos se adaptam de forma independente e são categorizados como supervisionados ou não supervisionados.

A categoria supervisionada é caracterizada principalmente pela utilização de rotulagem nos dados de entrada dos algoritmos, os quais são utilizados para realizar o treinamento dos modelos. Os modelos aprendem através do fornecimento de dois conjuntos de dados, sendo um conjunto de treinamento, que funciona como um supervisor que ensina as máquinas a prever a saída corretamente, e um conjunto de testes. Assim, o modelo é treinado até que possa detectar os padrões e relacionamentos subjacentes entre os dados de entrada e

os rótulos de saída, permitindo que ele produza resultados de rotulagem precisos quando apresentados aos dados do conjunto de testes.

Outra abordagem utilizada em Aprendizado de Máquina é a de aprendizado não supervisionado, em que não há rótulos ou classes conhecidas nos dados de entrada. O aprendizado não supervisionado desempenha um papel importante na análise exploratória de dados, na segmentação de mercado, na detecção de anomalias, entre outras aplicações em que o objetivo é encontrar padrões, estruturas ou agrupamentos intrínsecos nos dados, permitindo uma melhor compreensão do conjunto de dados e a identificação de informações úteis (Solorio-Fernández et al., 2020).

A clusterização ou agrupamento de dados é uma técnica essencial no campo de aprendizado não supervisionado, que visa agrupar objetos ou instâncias similares em conjuntos, chamados de clusters, afim de identificar estruturas e padrões intrínsecos nos dados, permitindo uma melhor compreensão e interpretação dos fenômenos estudados. Assim, é possível segmentar conjuntos de dados em grupos homogêneos, que, neste caso, se refere à proximidade ou similaridade entre os elementos do grupo conforme definido pela métrica escolhida, facilitando a análise e permitindo que essa técnica seja utilizada como uma ferramenta para tomada de decisões (Caruso et al., 2018).

Ainda no âmbito do Aprendizado de Máquina, as meta-heurísticas têm revelado sua eficácia na solução de problemas de otimização complexos e de grande porte, os quais surgem com frequência em diversas áreas. As meta-heurísticas exploram o espaço de soluções em busca de soluções de alta qualidade, permitindo encontrar resultados aproximados ou ótimos em problemas desafiadores. De acordo com Hussain (2019), essas técnicas oferecem uma abordagem poderosa e versátil para a resolução de problemas de otimização em uma ampla gama de domínios, sendo capazes de se adaptar para atender às necessidades específicas do problema em questão.

Nos últimos anos, as meta-heurísticas têm sido amplamente utilizadas para abordar problemas de clusterização de maneira eficaz. Sua aplicação à clusterização surgiu como uma alternativa aos métodos tradicionais, que frequentemente se deparam com desafios de complexidade computacional em que a quantidade de operações realizadas pode resultar em uma carga computacional intratável. Assim, estas técnicas têm sido utilizadas para problemas complexos,

com dados de alta dimensionalidade, ruído ou sobreposição entre os clusters, permitindo explorar diferentes configurações de clusters e encontrar agrupamentos que representem melhor os padrões nos dados (Nanda & Panda, 2014).

Neste trabalho, abordamos os conceitos e técnicas de clusterização por meio de abordagens baseadas em meta-heurísticas, discutindo suas propriedades e limitações. Exploramos o uso das meta-heurísticas Algoritmo Genético (Genetic Algorithm), Otimização por Enxame de Partículas (Particle Swarm Optimization) e Otimização por Colônia de Formigas (Ant Colony Optimization) para problemas de clusterização considerando diferentes métricas de distância e realizando uma comparação de desempenho entre elas.

Com o objetivo de compreender e avaliar o desempenho das meta-heurísticas em diversos cenários, conduzimos testes abordando uma variedade de situações, que incluíram conjuntos de dados benchmark amplamente utilizados pela literatura. Além disso, este estudo envolveu a coleta de dados sobre as características físicas de bovinos e a aplicação das meta-heurísticas supracitadas, com o objetivo de identificar os atributos mais discriminantes nos diferentes grupos e facilitar o agrupamento dos bovinos de forma eficiente e confiável.

1.1 Setor Pecuário

A pecuária desempenha um papel significativo e multifacetado em várias áreas da sociedade. Nos últimos anos, esta atividade avançou em muitas regiões rurais e em desenvolvimento no Brasil com um leque de atividades, desde a criação de bovinos até a produção de alimentos para eles, além do processamento de produtos pecuários e a comercialização, sendo sua interferência percebida no crescimento econômico e no acesso a serviços básicos, como saúde e educação dessas áreas (Vale et al., 2019).

No que tange à produção de alimentos, a pecuária é uma fonte essencial de proteína animal, fornecendo carne, leite, ovos e outros produtos de origem animal que são importantes para a nutrição e saúde humana. Assim, especialmente em regiões onde a produção agrícola é limitada, ela oferece uma fonte adicional de alimentos e ajuda a diversificar a oferta alimentar, reduzindo a dependência exclusiva de culturas agrícolas e contribuindo para a segurança alimentar (Rodrigues Fortes et al., 2020).

Não obstante, é importante destacar que a pecuária também enfrenta desafios, como a pressão pela conservação ambiental e a preocupação com o

bem-estar animal. Essas questões têm impactos tanto na percepção dos consumidores quanto nas regulamentações governamentais (Sanchez-Sabate & Sabaté, 2019; Balogh & Jámbor, 2020), já que os consumidores estão cada vez mais preocupados com a sustentabilidade e o impacto ambiental dos produtos que consomem. Além disso, os governos estão implementando políticas e regulamentações ambientais mais rigorosas, visando mitigar os impactos negativos da pecuária no meio ambiente.

O mercado pecuário é um setor complexo que envolve diversos agentes, como produtores, intermediários, frigoríficos, varejistas e consumidores. No mercado de bovinos vivos, os produtores vendem seus bovinos e os compradores podem ser outros produtores, frigoríficos ou intermediários que atuam na revenda. Há ainda o mercado de produtos pecuários, que são geralmente processados e comercializados por frigoríficos e indústrias alimentícias, e vendidos diretamente para varejistas, restaurantes, supermercados e consumidores finais.

Tratando-se especificamente do mercado de bovinos, o preço de venda desses bovinos é determinado por diversos fatores, como as condições do mercado pecuário e a demanda por carne, que desempenham um papel importante na indicação do preço. Fatores como oferta e demanda, sazonalidade, variações nos preços dos insumos, condições econômicas e preferências dos consumidores podem influenciar os preços no mercado, o que pode ser observado, por exemplo, durante a pandemia de COVID-19 (Martinez et al., 2021). A localização geográfica da propriedade ou do mercado onde o bovino está sendo vendido também pode afetar os preços, já que custos de transporte, proximidade de centros urbanos e disponibilidade de mercados podem influenciar os preços regionais (Asem-Hiablie et al., 2017).

Essa precificação depende ainda de características do próprio animal e dos custos com sua alimentação no processo de crescimento e engorda. Diferentes raças de bovinos têm características distintas em termos de tamanho, conformação, qualidade da carne e adaptabilidade a diferentes condições, com algumas raças sendo conhecidas por produzir carne de alta qualidade e sendo mais valorizadas no mercado (Mc Hugh et al., 2011). A condição corporal do bovino, que se refere ao estado de saúde, nutrição e musculatura do bovino, também pode afetar o preço de venda, já que bovinos em boa condição corporal são mais valorizados, pois indicam uma boa alimentação e cuidados adequados

(Mc Hugh et al., 2010).

Nesse sentido, o agrupamento de bovinos desempenha um papel crucial na gestão eficiente e produtiva da pecuária, uma vez que envolve agrupar os bovinos com base em critérios específicos, como idade, peso, raça, sexo ou estado reprodutivo. Essa prática permite o manejo adequado dos bovinos, o controle da sua alimentação, o manejo sanitário, a melhoria na seleção genética e a otimização da eficiência produtiva, sendo essencial para produtores que buscam maximizar sua produtividade, a saúde e o bem-estar de seu rebanho. Essa abordagem facilita a gestão da venda, uma vez que grupos com diferentes taxas de crescimento podem proporcionar lucros similares, já que custos associados ao crescimento do bovino podem ser otimizados para grupos específicos.

Face ao exposto, neste trabalho buscamos a determinação de um método capaz de agrupar, de maneira eficaz, os bovinos em grupos homogêneos.

1.2 Justificativa

Este trabalho é fundamentado na importância econômica e produtiva da pecuária, bem como na necessidade de otimizar o manejo e a produtividade dos rebanhos. Assim, o agrupamento adequado dos bovinos no pasto, considerando suas características individuais, se torna uma importante ferramenta de apoio à tomada de decisão, sendo fundamental para alcançar altos índices de ganho de peso, saúde e reprodução e, consequentemente, ganhos com a venda dos bovinos.

Através do agrupamento, é possível melhorar o manejo dos bovinos, maximizar a eficiência alimentar, controlar a sanidade do rebanho e promover uma seleção genética mais precisa, isso tudo de acordo com as necessidades nutricionais, idade, sexo ou características de desempenho de cada grupo. Ademais, uma abordagem de agrupamento adequada pode maximizar o uso dos recursos disponíveis, como pastagens e água, e promover o bem-estar e a saúde dos bovinos.

A criação de grupos homogêneos pode permitir a aplicação de técnicas específicas de manejo e melhorar a eficiência reprodutiva. Outro aspecto relevante é o tamanho dos grupos dos bovinos, em que grupos menores podem facilitar a observação e o monitoramento individual dos bovinos, que facilita a detecção precoce de doenças e a implementação de medidas de manejo específicas. Por

outro lado, grupos maiores podem promover interações sociais entre os bovinos, contribuindo para o bem-estar e evitando o isolamento de indivíduos, além de reduzirem os custos de manuseio e de criação de piquetes.

Dentre as diferentes estratégias de agrupamento, a busca pela melhor abordagem envolve a consideração de diversos fatores, como características do rebanho, uso de tecnologias e aspectos econômicos. A efetividade do agrupamento depende de um método de agrupamento preciso e objetivo, que leve em consideração as características do problema. Observa-se, então, a necessidade de estudos aprofundados especificamente nessa área, para fornecer diretrizes e recomendações tanto aos produtores quanto aos métodos de agrupamento.

1.3 Objetivos Geral e Específicos:

O objetivo geral deste trabalho consiste em comparar o desempenho das meta-heurísticas Algoritmo Genético, Enxame de Partícula e Colônia de Formigas para o problema de agrupamento de dados. Para alcançar este objetivo, tem-se como objetivos específicos:

- Realizar a avaliação computacional de agrupamento para conjuntos de dados sintéticos e para grupos de bovinos.
- Compreender e avaliar o impacto da variação dos parâmetros no desempenho de cada meta-heurística.
- Compreender como as características dos dados, tais como, distribuição, densidade e dimensionalidade podem afetar a eficiência das meta-heurísticas.
- Identificar as características mais discriminantes nos diferentes grupos gerados.

2 EMBASAMENTO TEÓRICO

Neste capítulo, apresentamos o referencial teórico relacionado à utilização de meta-heurísticas para o agrupamento de dados. Em um primeiro momento, tratamos da teoria e dos fundamentos conceituais que sustentam este estudo, destacando os avanços mais recentes e as áreas que ainda requerem investigação. Abordamos conceitos-chave e modelos teóricos relevantes, estabelecendo uma base sólida para a compreensão do tema em análise. Ademais, exploramos as limitações e desafios encontrados nos estudos existentes, apontando para possíveis lacunas e oportunidades de investigação.

2.1 Aprendizado Não Supervisionado

No aprendizado não supervisionado os modelos são treinados usando um conjunto de dados não rotulados e podem agir sobre esses dados sem qualquer supervisão. Nessa categoria, os modelos funcionam de forma independente dos conjuntos de dados e como não requer uma hipótese para identificar padrões, ele remove o viés da hipótese. Não obstante, como não se sabe quais saídas o algoritmo apresentará, os modelos necessitam que as variáveis de saída sejam validadas para confirmar os resultados apresentados.

O agrupamento de dados, associação e redução de dimensionalidade são as três principais tarefas para as quais se aplicam os modelos não supervisionados. No agrupamento os dados não rotulados são agrupados com base em suas similaridades ou dissimilaridades. Para a associação, os métodos são baseados em regras para encontrar relacionamentos entre variáveis em um determinado conjunto de dados. Na redução de dimensionalidade, dado um determinado conjunto de dados com muitos atributos, reduz-se o número de atributos a um tamanho gerenciável, ao mesmo tempo em que preserva a integridade do conjunto de dados o máximo possível.

2.1.1 Agrupamento de dados

O agrupamento de dados é a principal técnica de aprendizado não supervisionado. Esta técnica foi introduzida inicialmente para a pesquisa de mineração de dados e, com o avanço da tecnologia e do processamento de dados, passou a ser aplicada para a classificação de enormes bancos de dados, tendo se

tornado útil, por exemplo, para a análise exploratória de padrões e segmentação de imagens (Cooper & Folta, 2017).

Uma das definições amplamente adotadas na literatura para agrupamento é a proposta por Jain e Dubes (1988). Segundo eles, os objetos dentro de um mesmo cluster devem ser tão semelhantes quanto possível, enquanto os objetos pertencentes a diferentes clusters devem ser o mais distintos possível. Landau et al. (2011) complementam essa definição, resumindo um cluster como um conjunto de objetos semelhantes, enquanto objetos de diferentes clusters não compartilham a mesma identidade ou um grupo de pontos no espaço onde a distância entre quaisquer dois pontos dentro do cluster é menor do que a distância entre qualquer ponto do cluster e aqueles fora dele.

Gaertler (2005) descreveu dois aspectos principais para o processo de agrupamento: o primeiro envolve questões sobre como encontrar tais decomposições, ou seja, tratabilidade, que se refere ao desenvolvimento ou seleção de algoritmos capazes de lidar eficientemente com conjuntos de dados complexos, grandes ou de alta dimensionalidade. Enquanto isso, o segundo aspecto diz respeito à atribuição de qualidade, ou seja, quão boa é a decomposição computada, que é observada na fase de validação do agrupamento, onde técnicas são aplicadas para avaliar objetivamente a qualidade dos agrupamentos gerados em relação aos dados originais e aos critérios predefinidos de agrupamento.

É importante considerar que diferentes métodos podem produzir resultados distintos para um mesmo conjunto de dados, já que cada método de agrupamento possui suas particularidades, as quais dinamizam o processo de clusterização e tornam seu processamento distinto dos demais. Não obstante, mesmo que haja diferentes métodos de agrupamento, a Figura 1 apresenta procedimentos gerais para o processo de clusterização, os quais são seguidos em praticamente todos os métodos de agrupamento (Xu & Tian, 2015).

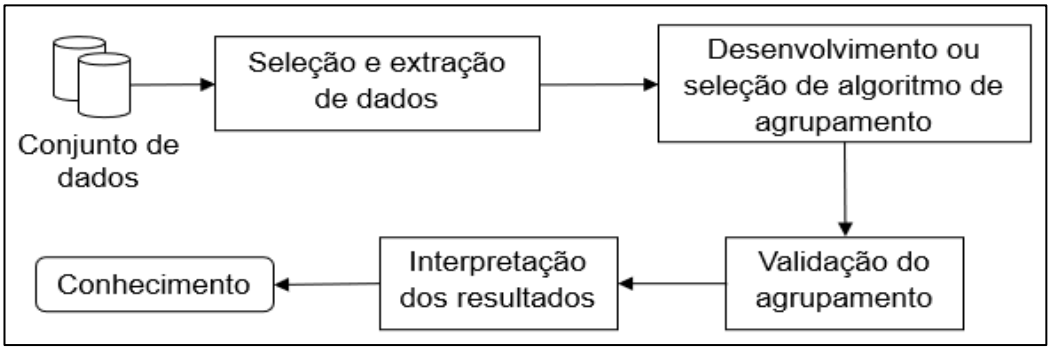


Figura 1: Procedimentos usuais no processo de clusterização de dados.

Fonte: Xu & Tian (2015).

Os procedimentos delineados na Figura 1 podem ser descritos da seguinte maneira:

- **Seleção e extração de dados:** este passo encaixa-se no pré-processamento dos dados, que tem como objetivo selecionar adequadamente os elementos nos quais o agrupamento será executado, de modo a codificar o máximo de informações possível sobre a tarefa de interesse.
- **Desenvolvimento ou seleção do algoritmo de agrupamento:** esta etapa refere-se ao desenvolvimento ou escolha de um algoritmo que resulte em um esquema de agrupamento eficaz para um conjunto de dados.
- **Validação do agrupamento:** nessa fase tem-se o objetivo de verificar a qualidade dos resultados do algoritmo de agrupamento usando técnicas apropriadas.
- **Interpretação dos resultados:** dados os resultados e sua validação, a sua interpretação passa a depender do seu interpretador. Em muitos casos, os especialistas na área de aplicação precisam integrar os resultados do agrupamento com outras evidências e análises experimentais para tirar suas conclusões.

Os métodos de clusterização estão divididos basicamente entre hierárquico e particional. A Figura 2 demonstra uma taxonomia desses métodos onde é possível observar que o método hierárquico está dividido entre as abordagens aglomerativa e divisiva, enquanto o particionado está dividido entre as abordagens baseadas em distância, densidade e modelo. Novas taxonomias para agrupamento de dados foram propostas mais recentemente aprofundando-se nos algoritmos desenvolvidos em cada uma das abordagens e até considerando novas

abordagens (Ezugwu et al., 2021; Oyewole & Thopil, 2023), mas percebe-se que estas são justamente uma atualização, na forma de complementação, da taxonomia apresentada pelos autores supracitados.

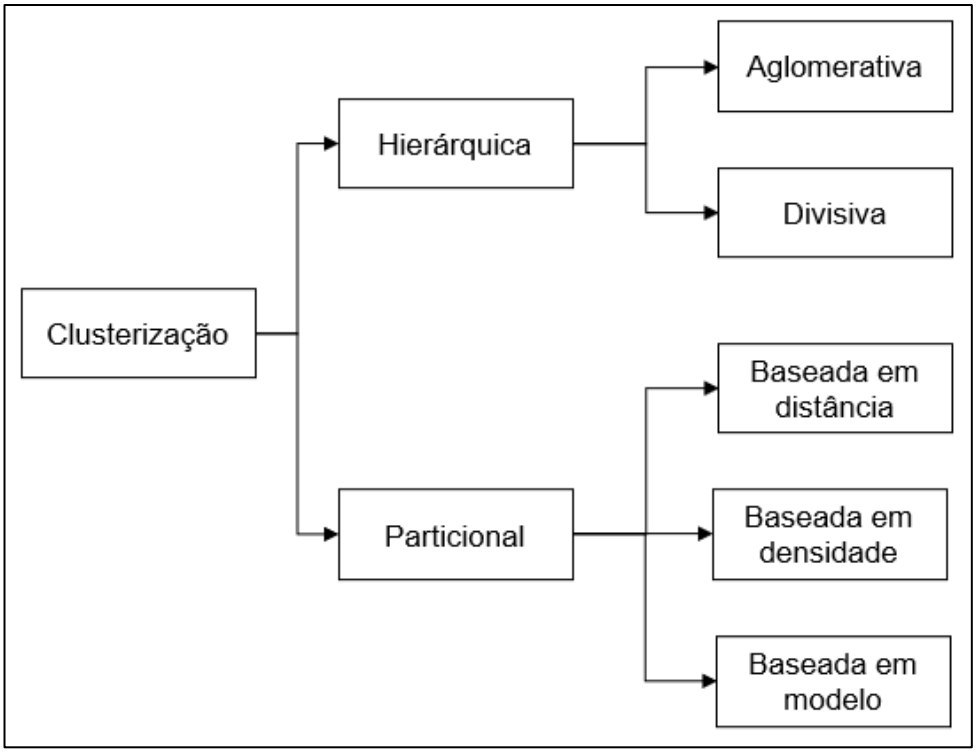


Figura 2: Taxonomia dos métodos e abordagens para agrupamento.
Fonte: Fraley e Raftery (1998).

Reddy e Vinzamuri (2018) mencionaram que as principais diferenças entre os métodos hierárquico e particional estão relacionadas ao tempo de execução, parâmetros de entrada e saída e a forma como os resultados são apresentados. Em geral, o agrupamento particional acaba sendo mais rápido que o hierárquico. O agrupamento hierárquico não requer parâmetros de entrada, enquanto os métodos particionais tradicionais precisam da quantidade de partições para iniciar. Além disso, o hierárquico retorna uma divisão de clusters mais significativa e subjetiva, e o particional resulta em partições estabelecidas de acordo com a quantidade de clusters determinada.

Conforme dito anteriormente, um dos parâmetros de entrada para o agrupamento particional é a definição do número de partições que se pretende obter, ou mesmo a falta desta definição. Quando a quantidade de clusters é definida previamente, denomina-se Problema de K-Clusterização ou, simplesmente, Problema de Clusterização. Se não há essa definição, denomina-

se então como Problema de Clusterização Automática. Chiou e Lan (2001) destacam que há um incremento significativo entre as possíveis diferentes formas de agrupamento para um mesmo grupo de dados entre problema de clusterização e clusterização automática.

Ezugwu et al. (2021) apontaram que, em problemas de análise de agrupamento de dados do mundo real, determinar a quantidade ideal de agrupamentos para um conjunto de dados de alta densidade e dimensionalidade é uma tarefa bastante difícil. Sendo assim, inúmeras pesquisas têm se dedicado a desenvolver métodos para determinar a melhor quantidade de partições para um conjunto de dados, além do desenvolvimento de algoritmos de agrupamento que integram a determinação da quantidade de cluster e o próprio processo de divisão (Zhou, Zu & Liu, 2016; Patil & Baidari, 2019; Shi et al., 2021).

Outra característica do agrupamento particional com abordagem baseada em distância e do método de agrupamento hierárquico é a necessidade de se determinar uma medida de distância, o que também tornou-se objeto de estudos que buscam comparar o desempenho dos algoritmos de agrupamento utilizando diferentes métricas.

Como a distância entre os objetos é a chave para o agrupamento, a proximidade entre eles é mensurada por meio do cálculo da distância determinada, as quais podem ser entendidas como medidas de similaridade e dissimilaridade entre os objetos, dependendo da técnica utilizada. No caso de similaridade, quanto maior o valor observado mais parecidos são os objetos, e na dissimilaridade, quanto maior o valor observado menos parecidos serão os objetos. Assim, cabe que objetos alocados no mesmo clusters sejam similares entre si e dissimilares em relação aos de outros clusters.

Com relação à validação dos clusters produzidos, como os algoritmos de agrupamento definem clusters que não são conhecidos a priori, independentemente dos métodos de agrupamento, a partição final dos dados requer algum tipo de avaliação na maioria das aplicações (Ben-David e Ackerman, 2008). Essa avaliação da qualidade da clusterização é essencial para verificar o desempenho do algoritmo de agrupamento e entender a estrutura dos dados, a fim de garantir que os grupos gerados sejam coerentes com a natureza dos dados.

Liu et al. (2010) apontaram que os critérios de validação podem ser classificados principalmente como externos ou internos, tendo como principal

diferença a utilização de informações externas para a validação. Assim, na validação externa a ideia básica é testar se os pontos do conjunto de dados são estruturados aleatoriamente ou não utilizando-se informações externas, como uma formação de agrupamento esperada ou conhecida. Já na abordagem interna, a validação do agrupamento é avaliada usando dados e recursos contidos no próprio conjunto de dados. Em todos os casos, índices de validade são construídos para avaliar a proximidade entre os objetos em um cluster ou a proximidade entre os clusters resultantes.

2.1.1.1 Clusterização hierárquica

O agrupamento hierárquico é uma técnica de análise de clusters que busca criar uma estrutura hierárquica de agrupamentos. Os métodos de agrupamento hierárquico, com raízes que remontam às décadas de 1960 e 1970, são continuamente reabastecidos com novos desafios e, como uma família de algoritmos, eles são fundamentais para resolver muitos problemas importantes, sendo utilizados em áreas como biociência, química e ciência de dados em geral (Murtagh & Contreras, 2017).

Esse tipo de clusterização é dividida entre os métodos hierárquico, aglomerativo e divisível. No método hierárquico aglomerativo, a decomposição hierárquica ocorre de baixo para cima, onde os clusters iniciais são construídos começando com um único objeto e depois mesclando esses clusters em clusters cada vez maiores, até que todos os objetos estejam finalmente em um único cluster ou até que certas condições de término sejam satisfeitas. Já no método divisível a decomposição ocorre de cima para baixo, onde divide-se o cluster contendo todos os objetos em grupos menores, até que cada objeto forme um grupo por conta própria ou até que satisfaça as condições de parada.

Os Quadros 1 e 2 apresentam o algorimo de agrupamento hierárquico pelo método aglomerativo e divisível, respectivamente, de acordo com a generalização feita por Oyewole e Thopil (2022) para os dois métodos.

Quadro 1: Algoritmo de agrupamento hierárquico pelo método aglomerativo.

Início
Inicialize com N objetos (cada um sendo seu próprio cluster).
Enquanto houver mais de um cluster, faça:
Para cada par de clusters, calcule a dissimilaridade usando a medida de distância determinada.
Identifique o par de clusters com a menor dissimilaridade.
Combine esses dois clusters em um novo cluster.
Recalcule as dissimilaridades entre o novo cluster e os clusters restantes.
Retorna hierarquia final de clusters.
Fim

Quadro 2: Algoritmo de agrupamento hierárquico pelo método divisível.

Início
Inicialize com um único cluster contendo todos os objetos.
Enquanto o número de cluster for menor que o número de objetos, faça:
Identifique o cluster mais heterogêneo com base em uma medida de dissimilaridade.
Separe este cluster em dois novos clusters.
Recalcule a dissimilaridade de todos os clusters.
Retorna conjunto final de clusters.
Fim

A dissimilaridade calculada entre os pares de clusters depende do tipo de ligação utilizada no modelo. Seja na abordagem aglomerativa ou divisiva, a distância entre os clusters no método hierárquico pode ser medida por ligação única (*single linkage*), que calcula a distância mínima entre os clusters antes de mesclá-los, ligação completa (*complete linkage*), em que se calcula a distância máxima entre os clusters antes de mesclá-los e ligação média (*average linkage*), que calcula a distância média entre os clusters antes de mesclá-los.

A forma mais usual de representar o resultado de um algoritmo de agrupamento hierárquico é através de um dendrograma (Figura 3), em que no eixo vertical a altura das uniões indica a dissimilaridade entre os objetos, permitindo identificar grupos próximos ou distantes. Nessa representação, cada um dos níveis intermediários pode ser visto como a combinação de dois clusters do próximo nível inferior ou a divisão de um cluster de nível superior, exibindo graficamente o processo de fusão e os clusters intermediários.

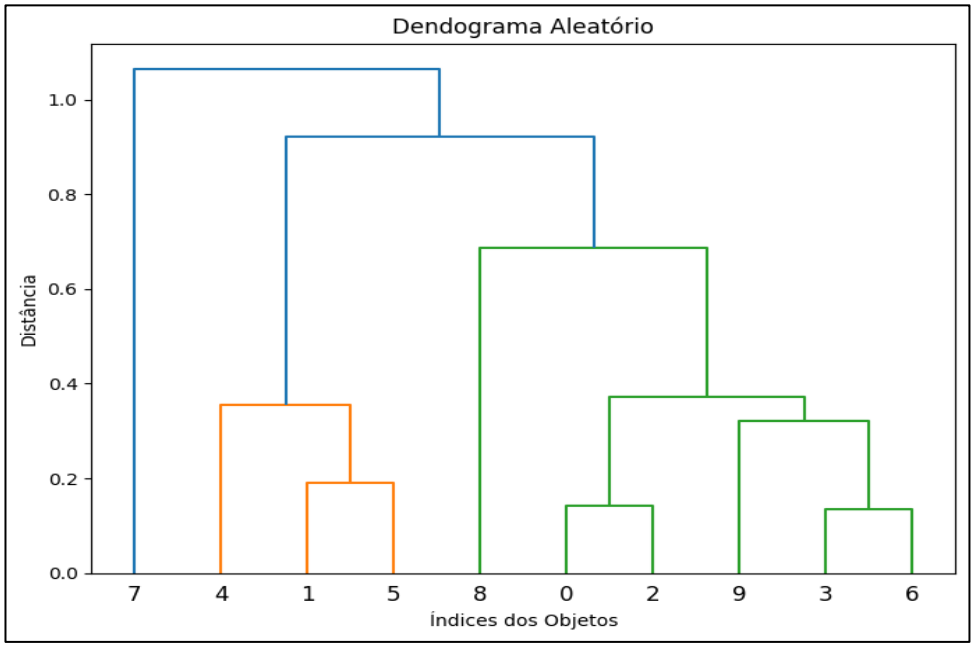


Figura 3: Representação de um dendograma aleatório.
Fonte: Autoria Própria (2024).

2.1.1.2 Clusterização particionada

O método de clusterização particionada tem como objetivo decompor o conjunto de dados em um conjunto de clusters disjuntos. Assim, dado um conjunto de dados com *N* pontos, o método de particionamento constrói as partições dos dados, com cada partição representando um cluster. Ou seja, classifica os dados em *K* grupos, com cada grupo contendo pelo menos um ponto e cada ponto pertencendo exclusivamente a um grupo. Aqui retoma-se a definição de Jain e Dubes (1988) mencionada anteriormente, que diz que os objetos no mesmo cluster devem ser o mais semelhantes possível e objetos em diferentes clusters devem ser o mais diferente possível.

Ao basear-se em distância, os agrupamentos são construídos a partir de vetores centrais (centróides), onde os objetos mais próximos a estes vetores são atribuídos aos respectivos clusters. A definição da quantidade de partições desejadas não é um parâmetro de entrada obrigatório, uma vez que o algoritmo pode ser de agrupamento automático. De qualquer forma, segue-se respeitando a condição de que cada cluster deve conter pelo menos um ponto e cada ponto deve pertencer exclusivamente a um cluster. A Figura 4 ilustra os resultados obtidos por meio de um algoritmo de agrupamento particionado baseado em distância, em que

é possível observar que a partição dos clusters está indicada pela cor dos pontos.

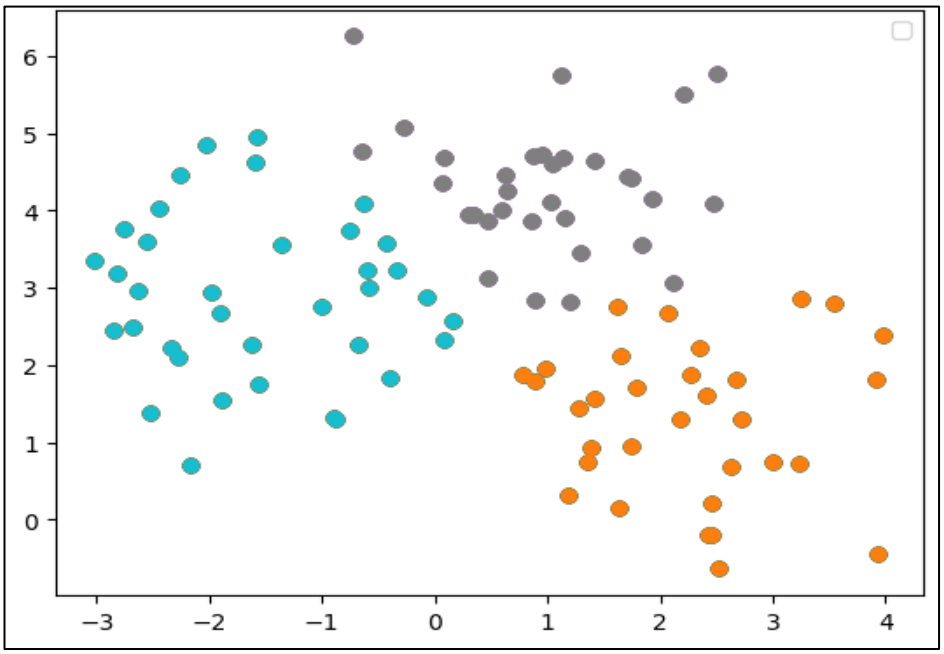


Figura 4: Representação de um particionamento.

Fonte: Autoria Própria (2024).

Assim como no método hierárquico, na clusterização particionada utilizam-se medidas de distância para verificar a similaridade entre os objetos presentes em um mesmo cluster e a dissimilaridade entre objetos em clusters diferentes. Van de Velden et al. (2019) dizem que não há uma medida de distância mais adequada para realizar o agrupamento de dados pelo método particionado, mas pode-se perceber uma prevalência na utilização da Distância Euclidiana nos algoritmos que adotam este método.

Vale destacar o papel do algoritmo de agrupamento particionado K-Means, que aparece como um dos principais métodos utilizados para a clusterização e tornou-se muito popularizado, já que é de simples aplicação e performa a contento com outros algoritmos mais sofisticados (Ahmed et al., 2020). No Quadro 3 é apresentado um pseudocódigo do algoritmo K-Means, contendo todos os passos necessários para sua aplicação.

Quadro 3: Pseudocódigo do Algoritmo K-Means.

Entrada: conjunto de dados com *i* instâncias e *d* dimensões, e o valor de *K* clusters desejados.
Início
    Inicie os *K* centróides com valores aleatórios.
    Enquanto critério de parada não é satisfeito, faça:
        Para cada amostra x_i, **faça**:
            Aloque x_i ao grupo com centróide C_k de menor distância, de acordo com a métrica de distância definida.
        Atualize os centróides C_k de cada cluster como a média dos pontos alocados no cluster.
    Retorna partição final de clusters.
Fim

A abordagem baseada em densidade não requer o número de clusters como parâmetros de entrada e, como não são baseados em medidas de distância, os clusters não são necessariamente grupos de pontos com uma baixa dissimilaridade dentro do cluster e, portanto, não necessariamente têm uma forma definida, mas podem ser arbitrariamente moldados no espaço de dados. Assim, um cluster baseado em densidade passa a ser “um conjunto de pontos espalhados no espaço de dados sobre uma região próxima de alta densidade de objetos, separados de outros clusters baseados em densidade por regiões contíguas de baixa densidade de objetos” (Kriegel et al., 2011).

Por sua vez, o agrupamento baseado em modelo é uma abordagem estatística para o agrupamento. Considera-se que os dados observados foram criados a partir de uma combinação finita de modelos de componentes, em que cada modelo de componente é uma distribuição de probabilidade, sendo geralmente uma distribuição multivariada paramétrica. Apesar da complexidade dos métodos baseados em modelos e de seu alto custo computacional, Bouveyron & Brunet-Saumard (2014) defendem que é melhor utilizá-los ao invés de realizar pré-processamento de dados e redução de dimensões.

2.1.2 Métricas de Distância

Como para o agrupamento particional com abordagem baseada em distância e no método de agrupamento hierárquico a distância entre os objetos é a chave, a proximidade entre eles é mensurada por meio do cálculo dessa distância. Os dados são agrupados em conjuntos por meio do uso de medidas de similaridade ou dissimilaridade entre os objetos, dependendo da técnica utilizada.

No caso de similaridade, quanto maior o valor observado mais parecidos são os objetos, e na dissimilaridade, quanto maior o valor observado menos parecidos serão os objetos. Assim, cabe que objetos alocados no mesmo clusters sejam similares entre si e dissimilares em relação aos de outros clusters.

A escolha da medida de distância também tornou-se objeto de estudos que buscam comparar o desempenho dos algoritmos de agrupamento utilizando-se diferentes métricas. Ao utilizar diferentes métricas de distância no algoritmo K Means, por exemplo, estudos mostram que essas medidas impactam não apenas no desempenho do algoritmo, mas também nos resultados obtidos (Singh, Yadav & Rana, 2013; Kapil & Chawla, 2016; Ghazal et al., 2021).

Distância Euclidiana: é uma medida de distância ou similaridade entre dois pontos em um espaço euclidiano, em que seu resultado é sempre não negativo e é igual a zero apenas quando os pontos são idênticos. Ela calcula a distância entre dois pontos em linha reta, o que reflete a ideia intuitiva de proximidade. Essa distância é dada por:

$$d(x,z) = \sqrt{\sum_{i=1}^{p} (x_i - z_i)^2} \quad (1)$$

Ainda que não seja a métrica mais adequada para todos os tipos de dados, por ser usualmente utilizada em problemas de agrupamento, a distância Euclidiana se torna um bom comparativo para avaliar o desempenho das outras métricas de distância. Sua aplicabilidade generalizada oferece uma base sólida para comparação com métricas menos convencionais, permitindo uma avaliação abrangente do impacto das diferentes abordagens de distância na qualidade dos agrupamentos obtidos (Ghazal et al., 2021).

Distância de Manhattan: também conhecida como métrica do táxi, calcula a soma das diferenças absolutas entre as coordenadas correspondentes dos pontos ao longo de cada dimensão. A mesma pode ser descrita por:

$$d(x,z) = \sum_{i=1}^{n} |x_i - z_i| \quad (2)$$

Testando diferentes distâncias para o problema de agrupamento, Gupta e

Chandra (2020) demonstram que a distância Manhattan obteve melhores resultados do que a distância Euclidiana, por exemplo. Essa não é uma realidade para todos os tipos de conjuntos de dados, mas outros estudos também demonstram o bom desempenho desta métrica para lidar dados normalizados e com alta dimensionalidade (Faisal & Zamzami, 2020; Alam, Muqeem & Ahmad, 2021).

Distância de Chebyshev: esta métrica pode ser descrita pelo valor máximo entre *n* diferenças absolutas, calculando a distância baseada no maior deslocamento entre as coordenadas dos pontos ao longo de cada dimensão. (3)
A mesma pode ser calculada por:

$$d(x, z) = \max_{i=1,\ldots,n} |x_i - z_i|$$

Já ao agrupar um grande banco de dados, Gultom et. al (2018) evidenciam que a distância Chebyshev obteve melhores resultados do que a Euclidiana e Chanberra. Surono e Putri (2021) também descrevem o bom desempenho desta distância ao aliar sua utilização para o agrupamento de dados ao processo de análise de componente principal, que é um dos passos utilizados no pré-processamento deste trabalho.

2.1.3 Definição do número de clusters

O método do cotovelo, também conhecido como *Elbow Method*, é uma técnica utilizada para determinar um número apropriado de clusters em algoritmos de clusterização. O método tem esse nome devido à forma do gráfico gerado quando se plotam os valores da inércia em relação ao número de clusters. O ponto em que a curva começa a nivelar-se, formando um padrão semelhante a um "cotovelo", indica o melhor número de clusters. Na Figura 5, por exemplo, o número de clusters indicado pelo método é 4.

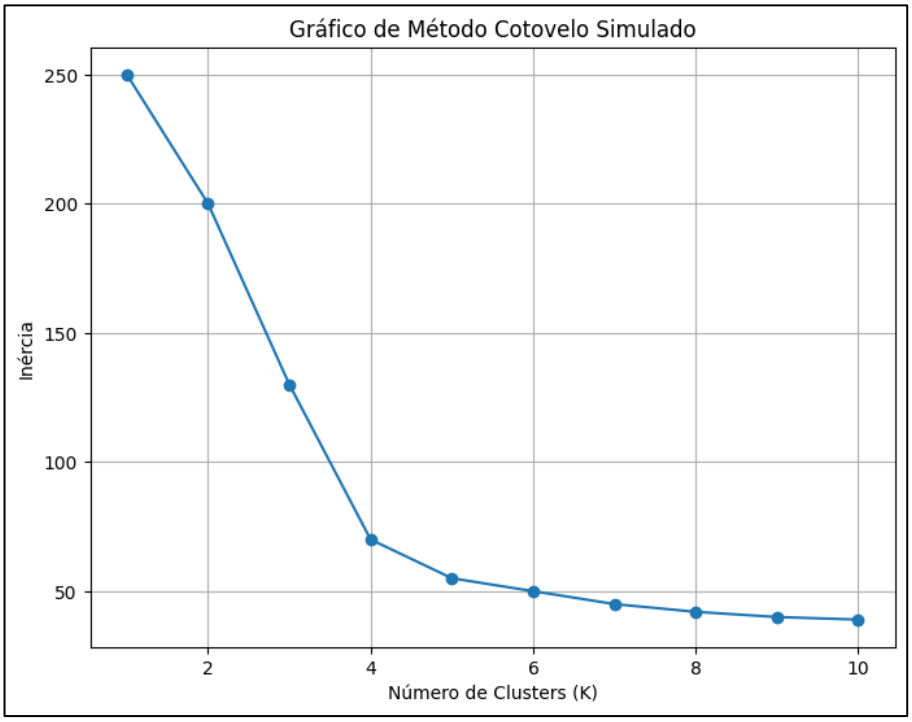


Figura 5: Exemplo de gráfico do método cotovelo.

Fonte: Autoria Própria (2024).

Para cada valor de *K*, calcula-se a inércia somando as distâncias ao quadrado de cada ponto x_{ij} ao centro do seu cluster c_j, como em (4):

$$WSS = \sum_{i=1}^{n} \sum_{j=1}^{k} \left| x_{ij} - c_j \right|^2. \quad (4)$$

A inércia é uma medida que quantifica a variabilidade dentro dos clusters. Quanto menor a inércia, mais compactos e homogêneos são os clusters. O método do cotovelo aproveita o fato de que, à medida que o número de clusters aumenta, a inércia tende a diminuir e, em algum ponto, adicionar mais clusters não levará a uma redução significativa na inércia, sendo este o ponto de inflexão no gráfico. Syakur et al., (2018), descreve esse processo por meio do algoritmo apresentado no Quadro 4.

Quadro 4: Algoritmo do método cotovelo.

Entrada: Conjunto de dados com *n* amostras e o intervalo de valores de clusters *K*.
Início
    Para cada valor de *K,* **faça:**
        Execute um algoritmo de agrupamento (por exemplo, K-Means) para *K* clusters.
        Calcule o valor da inércia do particionamento obtido, de acordo com a equação (2).
    Plote os valores da inércia (eixo Y) em relação ao número de clusters (eixo X).
    Retorna gráfico do cotovelo
Fim

2.1.4 Padronização dos dados

A padronização de dados é um processo que transforma os valores das variáveis para que eles tenham média zero e desvio padrão igual a um. Os valores padronizados são dados pela fórmula *z-score* (5):

$$z = \frac{x - \mu}{\sigma} \quad . \tag{5}$$

em que:

- z: valor padronizado.
- x: valor original da variável.
- μ: média dos valores da variável.
- σ: desvio padrão dos valores da variável.

O resultado da padronização é uma nova distribuição dos dados com média zero e desvio padrão igual a um. Isso significa que os valores transformados terão uma posição relativa em relação à média e à dispersão dos dados originais. A padronização é realizada separadamente para cada variável, ou seja, cada variável é padronizada independentemente das outras. Isso garante que todas as variáveis tenham a mesma escala e contribuam igualmente na análise ou modelo.

2.1.5 Análise de Componente Principal – PCA

A PCA (*Principal Component Analysis*) é uma técnica de análise multivariada que visa reduzir a dimensionalidade dos dados, mantendo a maior parte da informação original. A idéia é identificar os principais componentes que explicam a variabilidade dos dados, transformando-os em um novo conjunto de

variáveis não correlacionadas chamadas de componentes principais. A técnica PCA envolve as seguintes etapas (Chinnamgari, 2019; Jolly, 2018):

- **Centralização dos dados:** subtrair a média de cada variável dos dados para centralizá-los em torno da origem.
- **Cálculo da matriz de covariância:** calcular a matriz de covariância dos dados centralizados para medir as relações lineares entre as variáveis.
- **Cálculo dos autovetores e autovalores:** diagonalizar a matriz de covariância para obter os autovetores (componentes principais) e autovalores associados. Os autovetores representam as direções dos componentes principais, e os autovalores indicam a quantidade de variância explicada por cada componente principal.
- **Seleção dos componentes principais:** ordenar os componentes principais em ordem decrescente de acordo com seus autovalores. Geralmente, apenas os primeiros componentes principais, que explicam a maior parte da variância, são mantidos.
- **Projeção dos dados nos componentes principais:** projetar os dados originais nos componentes principais selecionados, obtendo as novas variáveis não correlacionadas.

2.1.6 Métricas de avaliação

A disposição original dos dados pode ter um impacto significativo nos valores dos índices de validação, já que estes são sensíveis à forma como os dados estão distribuídos e aos padrões de separação entre os clusters (Hancer, Xue & Zhang, 2020). Por exemplo, se os clusters se sobrepõem ou estão muito próximos na disposição original dos dados, se torna mais difícil avaliar a qualidade do agrupamento por meio de uma única métrica de avaliação.

Isso ocorre porque a sobreposição e a proximidade dificultam a distinção clara entre os clusters, o que afeta negativamente a avaliação da coesão interna e/ou a separação entre clusters. Neste sentido, é importante dispor de várias métricas de avaliação, a fim de se obter *insights* complementares sobre a qualidade dos agrupamentos. Entre as métricas de avaliação apresentadas por Liu et al. (2010), quatro métricas largamente utilizadas são:

- **Índice de Silhueta (*Silhouette Score*):** este índice mede a coesão e a separação dos clusters. Para cada objeto *i* em um cluster, a distância média

entre o objeto *i* e todos os outros objetos no mesmo cluster é dado por *a(i),* enquanto a distância média entre o objeto *i* e todos os objetos em cada cluster diferente é dado por *b(i).* Assim, o coeficiente de silhueta para o objeto *i* é dado por:

$$SH(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (6)$$

o qual assume um valor entre -1 e 1 para cada objeto, indicando o quão bem ele se encaixa no seu cluster em comparação com os clusters vizinhos. Quanto mais próximo de 1, melhor. Um valor negativo indica que o objeto pode ter sido atribuído ao cluster errado.

- **Índice de Davies-Bouldin (*Davies-Bouldin Index*):** este índice avalia a separação entre os clusters, levando em consideração a dispersão dentro de cada cluster e a distância entre os centróides dos clusters. Para cada cluster *i*, calcula-se a dispersão média dos objetos dentro do cluster *i*, denotada por S_i. Para cada outro cluster *j* diferente de *i*, calcula-se a distância entre os centróides dos clusters *i* e *j*, denotada por d_{ij}. Assim, calcula-se o valor *R(i, j)* como:

$$R(i,j) = \frac{S_i+S_j}{d_{i,j}} \quad (7)$$

O Índice de Davies-Bouldin para toda a clusterização é dado pela média dos valores *R(i, j)* para todos os clusters *k,* ou seja:

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_{ij} \quad (8)$$

cujo valor varia entre 0 e infinito, sendo que um valor próximo de 0 indica que os clusters estão bem separados e têm uma dispersão interna baixa e quanto maior o valor, pior é a separação entre os clusters.

- **Índice de Calinski-Harabasz:** este índice mede a separação entre os clusters em relação à dispersão dentro dos clusters. Considerando que n é o número de objetos e k é o número de clusters, o cálculo do Índice de Calinski-Harabasz envolve a dispersão interclusters, em que n_i é o número de pontos no cluster i, c_i é o centróide do cluster i e c é o centróide global dos pontos, e a dispersão intracluster, em que x_{ij} é o j-ésimo ponto no

cluster i, c_i é o centróide do cluster i e n_i é o número de pontos no cluster i . Assim, o índice é calculado da seguinte forma:

$$CH = \frac{\sum_{i=1}^{k} n_i.|c_i - c|^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}|x_{ij} - c_i|} \times \frac{n-k}{k-1} \quad (9)$$

Este índice não possui um intervalo fixo, podendo variar dependendo das características dos dados, mas valores mais altos para ele indicam melhor separação entre os clusters.

- **Índice de Rand Ajustado (ARI):** é uma métrica de avaliação de cluster que mede a similaridade entre dois agrupamentos, levando em consideração os rótulos reais das amostras. Considerando que *N* é o número total de amostras, n_{ij} é o número de amostras que são comuns a ambos os agrupamentos (clusters verdadeiros e clusters preditos), a_i é o número total de amostras no cluster i no agrupamento verdadeiro e b_i é o número total de amostras no cluster j no agrupamento predito, o índice ARI é dado por:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_i}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_i}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_i}{2}\right]/\binom{N}{2}} \quad (10)$$

Este índice varia entre -1 a 1, onde 1 indica uma concordância perfeita entre os agrupamentos, 0 significa que os agrupamentos são independentes um do outro e têm uma concordância igual ao acaso e -1 indica total discordância entre os agrupamentos.

3 META-HEURÍSTICAS

Neste capítulo, são apresentadas as bases teóricas para a implementação das três meta-heurísticas utilizadas para o agrupamento de dados, sendo elas a Algorítimo Genético, Enxame de Partículas e Colônia de Formigas. Além disso, são apresentados ainda a função objetivo utilizada pe a forma de codificação das suas soluções.

As meta-heurísticas são utilizadas principalmente para encontrar soluções para problemas com informações incompletas ou quando se tem capacidade de computação limitada. Geralmente, elas são usadas para resolver problemas NP-completos. Alguns exemplos de problemas que podem ser resolvidos por meta-heurísticas são o problema do caixeiro viajante (Osaba et al., 2020), problema da mochila (Cacchiani et al., 2022), problema de empacotamento (Liu et al., 2020) e outros problemas de otimização clássicos da literatura.

Os vários tipos de meta-heurísticas podem ser classificados de acordo com diferentes critérios, como o nível de abstração, a forma de exploração do espaço de soluções, a estrutura de memória, o uso de aleatoriedade. Uma das formas de classificar as meta-heurísticas é de acordo com o número de soluções que elas mantêm e manipulam durante a busca (Boussaid et al., 2013), sendo meta-heurísticas com única solução ou com várias soluções.

Outra forma de classificar as meta-heurísticas é de acordo com a forma como elas se inspiram para modelar o processo de busca. Nesse caso, existem várias classificações na literatura, mas podemos distinguir três tipos principais, que são as meta-heurísticas baseadas em física, em biologia e em comportamento social (Ezugwu et al., 2021).

Considerando o estudo de Milan et al. (2019), que realizou uma análise acerca das meta-heurísticas inspiradas na natureza, é possível perceber que elas podem ser aplicadas a uma variedade de problemas sem precisar de muitas adaptações e podem encontrar soluções boas em um tempo razoável, mesmo para problemas muito difíceis ou complexos. Além disso, podem escapar de ótimos locais e explorar diferentes regiões do espaço de soluções em busca da melhor solução.

Contudo, as meta-heurísticas não garantem encontrar a solução ótima global, nem o quanto a solução encontrada se aproxima do ótimo e podem ser sensíveis aos parâmetros escolhidos, que podem afetar o desempenho e a

qualidade das soluções. Além disso, podem exigir muito esforço computacional, especialmente se o espaço de soluções for muito grande ou complexo.

3.1 Meta-heurísticas para agrupamento de dados

Tendo surgido como uma alternativa aos métodos de agrupamento tradicionais, que muitas vezes enfrentam dificuldades devido à complexidade dos dados, a aplicação de meta-heurísticas tem ganhado destaque na solução de problemas de agrupamento. Sua utilização em problemas complexos envolvendo dados de alta dimensionalidade, ruído ou sobreposição entre clusters vem se tornando cada vez maior, permitindo a exploração de várias configurações e a identificação de agrupamentos que melhor representem os padrões presentes nos dados.

Na sua aplicação para a tarefa de agrupamento, esses algoritmos se destacam por sua capacidade de explorar o espaço de soluções de maneira abrangente e eficiente, escapando de ótimos locais para encontrar soluções de alta qualidade. Embora possam exigir ajustes cuidadosos de parâmetros e serem computacionalmente intensivos, a flexibilidade e eficácia dessas técnicas em adaptar-se a diferentes tipos de problemas de agrupamento fazem delas uma escolha valiosa.

Neste trabalho, as meta-heurísticas AG, PSO e ACO são exploradas individualmente. Das, Abrahan e Konar (2009), ao realizar uma revisão bibliográfica sobre a aplicação de meta-heurísticas para o agrupamento de dados, evidencia que essas três meta-heurísticas podem levar a resultados eficazes e encontrar soluções de alta qualidade para os problemas de agrupamento.

Cada uma dessas meta-heurísticas possui uma abordagem única para resolver problemas de otimização. O algoritmo genético utiliza conceitos de evolução biológica, a enxame de partículas se baseia no comportamento coletivo de um enxame, e a colônia de formigas se baseia no comportamento das formigas em busca de comida, ou seja, algoritmo genético utiliza operadores genéticos como *crossover* e mutação, a enxame de partículas utiliza as interações sociais e individuais das partículas e a colônia de formigas, por sua vez, utiliza trilhas de feromônios.

As três meta-heurísticas oferecem flexibilidade na definição e ajuste de parâmetros. Isso permite adaptá-las às características específicas do problema de

clusterização e otimizar o desempenho do algoritmo. A capacidade de ajustar parâmetros, como tamanho da população, taxa de *crossover* e fatores de influência, é fundamental para adaptar o algoritmo ao problema apresentado e obter resultados mais adequados.

3.1.1 Algoritmo Genético (AG)

O Algoritmo Genético é uma técnica de otimização baseada em princípios inspirados na genética e na evolução natural, que tem mostrado eficácia em diversos problemas complexos, incluindo a clusterização de dados. Uma das principais vantagens do algoritmo genético é sua capacidade de lidar com problemas de otimização global e encontrar soluções aproximadas de alta qualidade. Ele é capaz de explorar o espaço de busca de forma ampla, permitindo a descoberta de diferentes agrupamentos nos dados.

Um dos primeiros trabalhos utilizando o Algoritmo Genético paro agrupamento de dados é o de Maulik & Bandyopadhyay (2000), em que capacidade de busca dos algoritmos genéticos é explorada para encontrar os centros de cluster apropriados no espaço de características, otimizando uma métrica de similaridade dos clusters resultantes e os cromossomos, representados como cadeias de números reais, codificam os centros de um número fixo de clusters. A superioridade do algoritmo AG sobre o amplamente utilizado algoritmo K-means é extensivamente demonstrada pelos autores através da sua aplicação utilizando quatro conjuntos de dados artificiais e três conjuntos de dados do mundo real.

Liu, Wu & Shen (2011) desenvolveram um método de clusterização baseado em algoritmos genéticos chamado Clustering Genético Automático, para valor de K desconhecido. Em seu algoritmo, a seleção de ruído e a mutação de divisão-absorção são projetadas para manter um equilíbrio entre pressão de seleção e diversidade da população. Os resultados experimentais em conjuntos de dados artificiais e do mundo real ilustram a eficácia do método na evolução automática do número de clusters e na geração da partição de clusterização.

Neste trabalho, o Algoritmo Genético utilizado para agrupamento de dados foi construído com base nos trabalhos de Maulik & Bandyopadhyay (2000) e Liu, Wu & Shen (2011). No Quadro 5 é apresentado o pseudocódigo do Algoritmo Genético, de acordo com os passos descritos por estes autores.

Quadro 5: Pseudocódigo do Algoritmo Genético

Início
Inicialize a população de indivíduos aleatoriamente.
Avalie a aptidão (qualidade dos agrupamentos) de cada indivíduo na população.
Enquanto o critério de parada não for atendido, faça:
Selecione os pais da população para reprodução por meio de técnicas de seleção, levando em consideração a aptidão dos indivíduos.
Aplique operadores genéticos para gerar descendentes:
Cruzamento: Combina informações genéticas dos pais para gerar novos indivíduos (descendentes) por meio de técnicas de cruzamento.
Mutação: Introduz variação genética nos descendentes, alterando aleatoriamente seus genes de centróides.
Avalie a aptidão dos descendentes, calculando a qualidade dos agrupamentos que eles representam.
Atualize a população, substituindo os indivíduos menos aptos pelos descendentes gerados.
Retorna o melhor indivíduo encontrado ao final do algoritmo.
Fim

Um fluxograma do pseudocódigo do Algoritmo Genético aplicado ao agrupamento de dados é apresentado na Figura 6, seguindo os passos apresentados no Quadro 5.

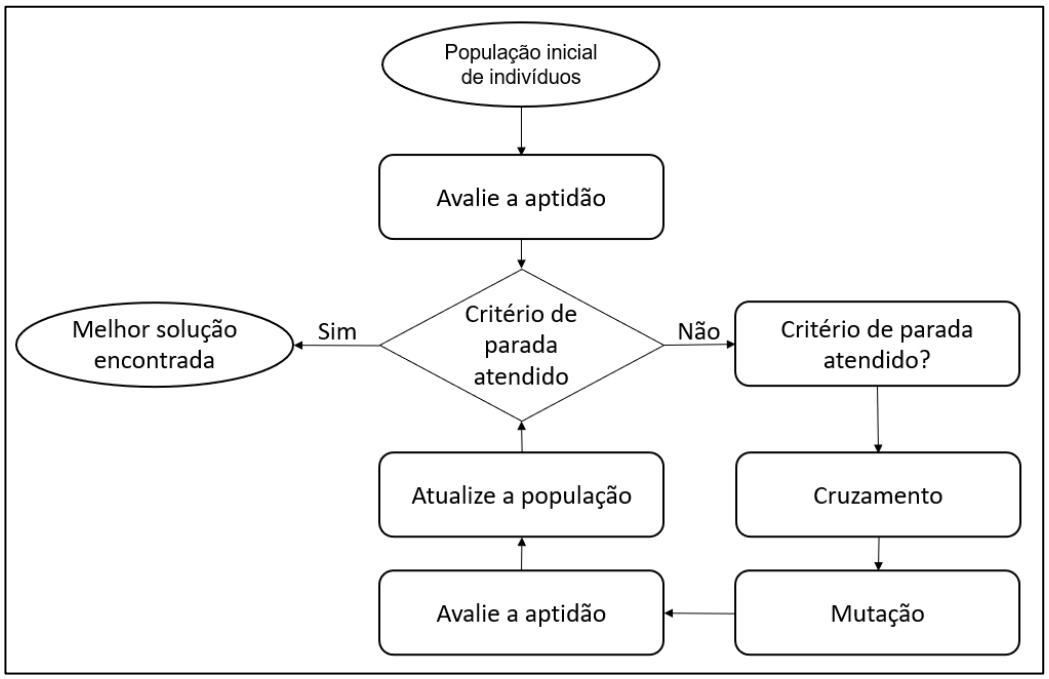


Figura 6: Fluxograma do pseudocódigo do Algoritmo Genético.

Fonte: Autoria Própria (2024).

Inicialmente, gera-se a população de ínidividuos aleatoriamente, onde cada indivíduo representada uma possível solução. Os individuos são caracterizados por seus centróides e a representação dos cromossomos é feita de forma contínua, onde as coordenadas dos centróides são as variáveis que compõem o cromossomo. Por exemplo, se o conjunto de dados possui 2 atributos e o número de clusters é definido como 3 (*K* = 3), então o cromossomo "51.6 72.3 18.3 15.7 29.1 32.2" representa os três centros de cluster (51.6, 72.3), (18.3, 15.7) e (29.1, 32.2).

Conforme o trabalho de Maulik & Bandyopadhhay (2000), o processo de cálculo da aptidão consiste em duas fases. Na primeira fase, os clusters são formados de acordo com os centros codificados no cromossomo. Isso é feito atribuindo cada ponto xi, i = 1,2,…,n, a um dos clusters c_j com centro z_j, tal que:

$$|x_i - z_i| < |x_i - z_p|, p = 1, 2, \dots, K \; e \; p \neq j \tag{11}$$

Após a formação dos clusters, os centros dos clusters codificados no cromossomo são substituídos pelas médias dos pontos nos respectivos clusters. Em outras palavras, para o cluster c_i, o novo centro z_i^* é calculado como a média dos pontos x_i pertencentes a c_i, onde i = 1,2,…,K. Esses $z_i^* s$ substituem os $z_i s$ anteriores no cromossomo. A etapa de avaliação da aptidão de cada solução é realizada conforme descrito na Seção 3.2, utilizando as métricas de distância individualmente.

Para selecionar os pais para reprodução aplicou-se a seleção por roleta. Nesse método, cada cromossomo recebe uma porção na roleta proporcional à sua aptidão. A roleta é imaginada como uma roda dividida em fatias, onde cada fatia representa um cromossomo. Quanto maior a aptidão de um cromossomo, maior é o tamanho da sua fatia na roleta. Em seguida, um sorteio aleatório é realizado para escolher um ponto na roleta, e o cromossomo correspondente à fatia onde o ponto cai é selecionado para fazer parte do grupo de cromossomos que será usado para criar a próxima geração de soluções.

O cruzamento é um processo probabilístico que troca informações entre dois cromossomos parentais para gerar dois cromossomos filhos. Neste trabalho, conforme Maulik & Bandyopadhhay (2000), é utilizado o cruzamento de ponto único (*single-point*). Para cromossomos de comprimento l, um número inteiro aleatório, chamado de ponto de cruzamento, é gerado no intervalo $[1, \; l - 1]$. As

porções dos cromossomos à direita do ponto de cruzamento são trocadas para produzir descendentes.

Nas técnicas de agrupamento baseadas em AG, os operadores de seleção controlam a direção da busca, enquanto os operadores de cruzamento e mutação geram novas regiões para a busca. Consonante à Bandyopadhyay (2011), visando manter um certo nível de controle sobre o processo de otimização, ao mesmo tempo em que reduz a complexidade geral do código e minimiza os riscos associados a variações excessivas na taxa de mutação, o processo de mutação ocorreu com base em uma taxa fixa mutação.

3.1.2 Otimização por Enxame de Partículas (PSO)

O PSO (*Particle Swarm Optimization*) é uma técnica de otimização inspirada em princípios naturais, mais especificamente no comportamento de enxames de pássaros e cardumes de peixes. Assim como o Algoritmo Genético, o PSO é amplamente utilizado em problemas complexos de otimização, incluindo a clusterização de dados. Uma das vantagens do PSO é sua simplicidade conceitual e facilidade de implementação.

Van der Merwe & Engelbrecht (2003) apresentas duas abordagens para o uso da PSO no agrupamento de dados. A primeira abordagem demonstra como o PSO pode ser aplicado para encontrar os centróides de um número pré-determinado de clusters. A segunda abordagem estende o algoritmo, utilizando o método K-means para inicializar o enxame, com o PSO refinando os clusters formados pelo K-means. Os novos algoritmos de PSO foram testados em seis conjuntos de dados e comparados com a performance do K-means. Os resultados indicam que ambas as técnicas de clusterização com PSO apresentam ótimos resultados em comparação com métodos de agrupamento tradicionais.

Ahmadyfard & Modares (2008) propuseram um método de clusterização baseado na combinação da PSO e do algoritmo K-means. Para estes autores, algoritmo PSO mostrou-se eficaz em convergir nas etapas iniciais de uma busca global, mas o processo de busca torna-se muito lento quando próximo do ótimo global. Os resultados experimentais em cinco conjuntos de dados, incluindo dados reais e sintéticos, demonstraram que o algoritmo híbrido supera consistentemente tanto o *K-means* quanto os métodos de clusterização baseados em PSO.

Por sua vez, Chen & Ye (2012) apresentam um algoritmo de análise de

cluster baseado em PSO que utiliza a função objetivo para buscar automaticamente os centros dos clusters em um espaço euclidiano de *n* dimensões. Seus resultados experimentais com quatro conjuntos de dados artificiais demonstram que o algoritmo PSO-clustering apresenta melhor desempenho do que os algoritmos tradicionais de análise de cluster, incluindo o K-means e Fuzzy C-Means. O algorimo utilizado por esses autores é apresentado no Quadro 6.

Quadro 6: Pseudocódigo do Algoritmo de Otimização Enxame de Partículas.

Início
Inicialize um conjunto de partículas aleatoriamente.
Avalie a aptidão de todas as partículas.
Para cada partícula, defina a sua melhor posição pessoal inicial como a sua posição atual.
Defina a melhor posição global como a melhor posição pessoal entre todas as partículas.
Enquanto o critério de parada não for atendido, para cada partícula, faça:
Atualize a velocidade da partícula, levando em consideração sua posição atual e a melhor posição global encontrada.
Atualize a posição da partícula com base na sua velocidade atual.
Avalie a aptidão da partícula na nova posição.
Se a aptidão atual for melhor do que a melhor aptidão pessoal da partícula:
Atualize a melhor posição pessoal da partícula com a sua posição atual.
Se a aptidão atual for melhor do que a melhor aptidão global encontrada:
Atualize a melhor posição global com a posição atual da partícula.
Retorna a melhor posição global encontrada, que representa os centróides que geraram a melhor qualidade de agrupamento.
Fim

A Figura 7 apresenta um fluxograma do pseudocódigo do Algoritmo de Otimização por Enxame de Partículas aplicado ao agrupamento de dados, conforme o Quadro 6.

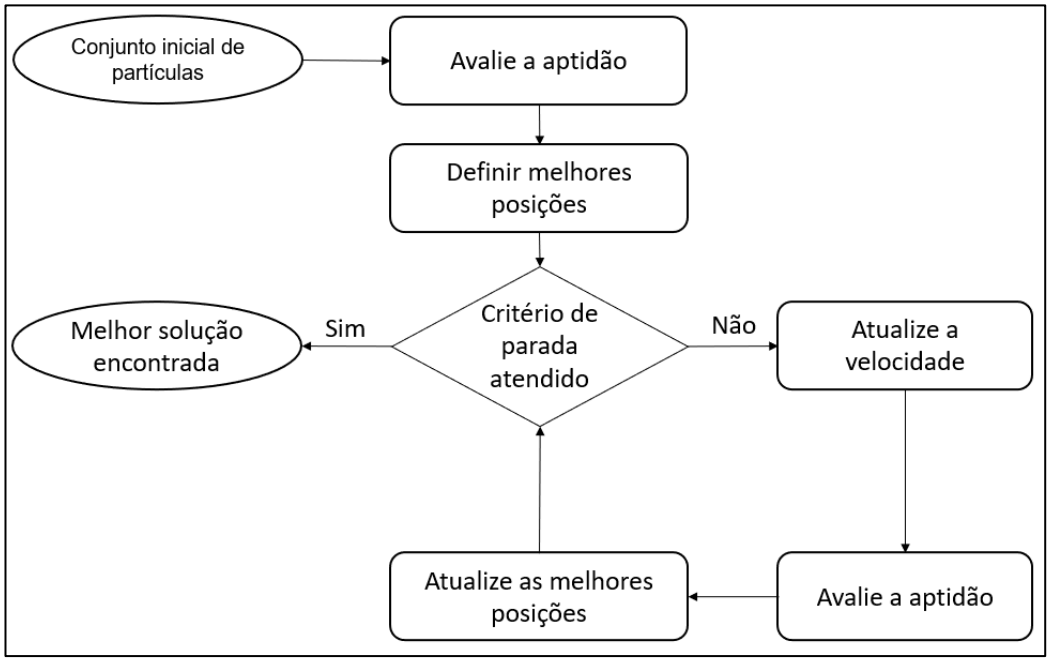


Figura 7: Fluxograma do pseudocódigo do Algoritmo de Otimização por Enxame de Partículas.

Fonte: Autoria Própria (2024).

Em PSO, cada partícula é uma possível solução do problema de agrupamento, onde os centróides dos clusters são as características que definem essa solução, ou seja, as coordenadas dos centróides são as variáveis que compõem cada uma das partículas. Tomando o mesmo exemplo usado na Seção 3.1.1, se o conjunto de dados possui 2 atributos e o número de clusters é definido como 3 (K = 3), então a partícula “51.6 72.3 18.3 15.7 29.1 32.2” representa os três centros de cluster (51.6, 72.3), (18.3, 15.7) e (29.1, 32.2).

A etapa de avaliação da aptidão de cada solução é realizada conforme descrito na Seção 3.2. Na primeira iteração, define-se a posição a melhor posição individual como a posição atual de cada solução e a melhor posição global a partir da melhor posição entre todas as soluções individuais, considerando os valores da avaliação de aptidão.

As fórmulas de atualização de velocidade e posição (Equações 12 e 13) aplicada à clusterização são as mesmas utilizadas no PSO padrão. Essas fórmulas são fundamentais para o funcionamento do PSO, pois permitem que as partículas explorem o espaço procurando soluções melhores e sejam influenciadas pela experiência individual e coletiva do enxame para convergir para uma solução ótima ou próximo dela. Conforme Van de Merwe & Engelbrecht (2003), uma fórmula de

atualização da velocidade v_i aplicado à clusterização é dada por:

$$v_{ik}(t\ +1) = \ w.v_{i,k}(t) + \ c_1.r_1.\left(y_{i,k}\ (t) - \ x_{i,k}(t)\right) + c_2.r_2.\left(\hat{y}_{i,k}\ (t) - \ x_{i,k}(t)\right) \quad (12)$$

em que:

- $v_{\mathrm{ik}}(t\ +1)$: velocidade da partícula i no componente k na iteração $(t\ +\ 1)$.
- w: peso de inércia.
- $v_{\mathrm{i,k}}(t)$: velocidade da partícula i no componente k na iteração (t).
- c_1: fator de aceleração cognitiva.
- r_1: número aleatório gerado no intervalo [0, 1].
- $y_{\mathrm{i,k}}\ (t)$: melhor posição individual do componente k da partícula i na iteração (t).
- $x_{\mathrm{i,k}}(t)$: posição do componente k da partícula i na iteração (t).
- c_2: fator de aceleração social.
- r_2: número aleatório gerado no intervalo [0, 1].
- $\hat{y}_{i,\mathrm{k}}\ (t)$: melhor posição global do componente k na iteração (t).

Após a atualização da velocidade, é possível realizar a atualização da posição atual da partícula por meio de:

$$x_{i,k}(t\ +1) = \ x_{i,k}(t) + \ v_{ik}(t\ +1) \quad (13)$$

em que:

- $x_{i,k}(t\ +1)$: nova posição da partícula i no *k*-ésimo atributo na iteração $(t\ +1)$.
- $x_{i,k}(t)$: posição atual da partícula i no *k*-ésimo atributo na iteração (t).
- $v_{ik}(t\ +1)$: velocidade da partícula i no *k*-ésimo atributo na iteração $(t\ +1)$.

3.1.3 Colônia de Formigas (ACO)

A ACO (*Ant Colony Optimization*) é uma técnica de otimização inspirada no comportamento natural das formigas na busca por caminhos mais curtos entre suas colônias e fontes de alimento. Uma das características distintivas da ACO é sua simulação do processo de comunicação entre as formigas por meio do depósito e evaporação de feromônios em trilhas. Esses feromônios guiam a busca por soluções otimizadas.

Shelokar, Jayaraman & Kulkarni (2004) propuseram uma abordagem de agrupamento baseada em ACO, sendo sua performance rigorosamente testada e comparada com outras meta-heurísticas. Os resultados dos testes mostraram que a abordagem baseada em ACO não só produziu soluções de alta qualidade, mas também obteve bons resultados em termos de número médio de avaliações da função objetivo e do tempo de processamento necessário. Em termos de qualidade de solução, o algoritmo ACO apresentou agrupamentos que melhor capturaram os padrões inerentes aos dados.

Trejos, Murillo & Piza (2004) utilizaram a ACO no problema de particionamento de dados para aprimorar as soluções obtidas pelo método k-means, associando cada formiga a uma partição, que é modificada conforme os princípios da meta-heurística. Comparando-os com outras técnicas, como recozimento simulado, algoritmos genéticos, busca tabu e o próprio k-means, os autores mostraram-se resultados tão bons quanto os dos demais métodos mencionados, evidenciando a eficácia da abordagem baseada em ACO para o particionamento de dados.

No trabalho de Runkler (2005), a ACO foi generalizada mostrar como modelos de agrupamento baseados em funções objetivas, como c-means rígidos e difusos, podem ser otimizados através de extensões específicas da ACO simplificada. Experimentos realizados com conjuntos de dados artificiais e reais demonstram que o agrupamento de formigas produz resultados superiores em comparação com a otimização alternada, principalmente por ser menos sensível a ótimos locais. Essa abordagem permite explorar de maneira mais eficaz o espaço de soluções, evitando armadilhas de subótimos que frequentemente prejudicam outras técnicas de agrupamento.

Considerando os trabalhos supracitados, neste trabalho a implementação da meta-heurística ACO foi baseada em Shelokar, Jayaraman & Kulkarni (2004), Trejos, Murillo & Piza (2004) e Runkler (2005) e os passos de sua implementação podem ser observados no Quadro 7.

Quadro 7: Pseudocódigo do Algoritmo Colônia de Formigas

Início
Inicialize um conjunto de soluções aleatoriamente.
Inicialize a trilha de feromônio.
Avalie a qualidade dos agrupamentos para cada solução na população e defina a melhor solução entre todas as soluções como a melhor solução geral.
Atualize a trilha de feromônio baseado na qualidade das soluções.
Enquanto o critério de parada não for atendido, faça:
Gere um novo conjunto de soluções usando a trilha de feromônio atualizada e a estratégia de exploração.
Realize a busca local no conjunto de soluções geradas.
Avalie a qualidade dos agrupamentos da nova solução.
Se a melhor aptidão das novas soluções for melhor do que a aptidão da melhor solução geral:
Atualize a melhor solução geral com a melhor solução atual.
Atualize a trilha de feromônio baseado na qualidade das soluções.
Retorna a melhor solução geral encontrada.
Fim

Seguindo os passos do Quadro 7, a Figura 8 apresenta um fluxograma do pseudocódigo do Algoritmo Colônia de Formigas aplicado ao agrupamento de dados.

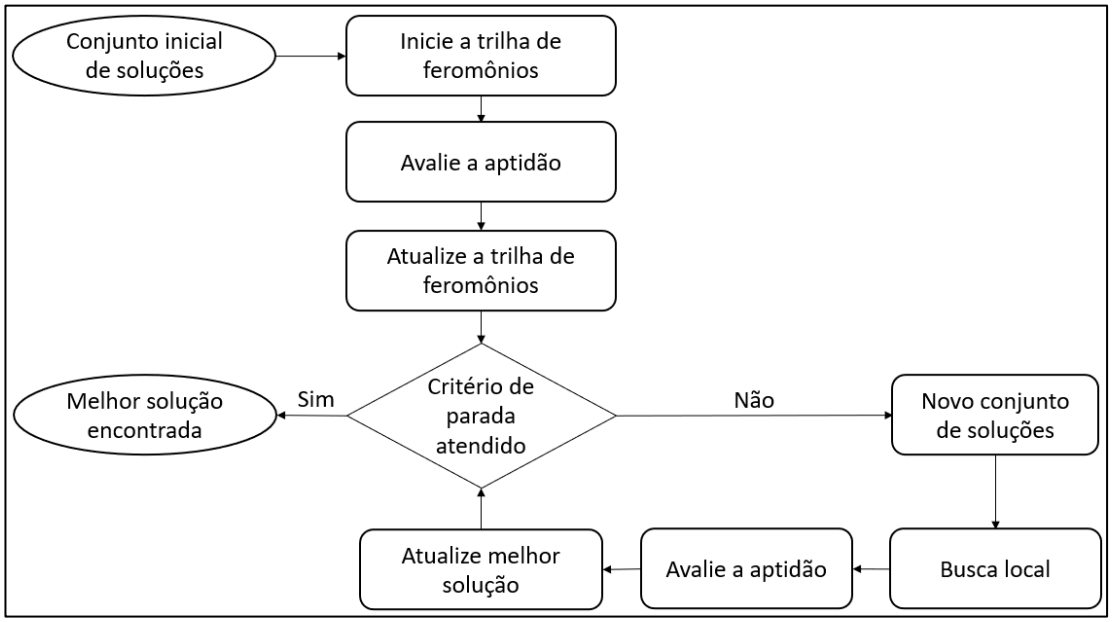


Figura 8: Fluxograma do pseudocódigo do Algoritmo Colônia de Formigas.

Fonte: Autoria Própria (2024).

Em ACO, cada formiga representa uma possível solução, que juntas formam um conjunto de soluções. Nesta abordagem, a codificação de cada solução aloca

diretamente *n* objetos a *k* clusters, de modo que cada solução seja formada com valores inteiros no intervalo [1, *k*]. Por exemplo, para *n*=5 e *k*=3, a solução com a codificação "11322" aloca o primeiro e o segundo objeto ao cluster 1, o terceiro objeto ao cluster 3 e o quarto e quinto objetos ao cluster 2. O centróide de cada cluster da solução é dado pela média de todos os pontos pertencentes a este determinado cluster.

Todos os valores na matriz de feromônios foram definidos com um mesmo valor inicial, a fim de dar a todas as possíveis soluções uma chance igual de serem escolhidas pelos agentes (ou formigas) no início. Essa forma de inicialização promove a diversificação das escolhas feitas pelas formigas nos estágios iniciais do algoritmo, ajudando a evitar que as formigas fiquem presas a uma única solução logo no início e incentivando a exploração de várias alternativas (Li et al., 2022).

Ainda que a codificação de cada solução seja diferente das outras duas meta-heurísticas, isso não interfere no cálculo da função *fitness*, a qual continua sendo realizada conforme descrito na Seção 3.2. A solução que apresenta o melhor resultado para a função de aptidão é definida como a melhor solução geral na iteração inicial.

Seguindo o trabalho de Shelokar, Jayaraman & Kulkarni (2004), no processo de busca local, cada membro da população é ordenado de maneira crescente de acordo com seus valores para a função aptidão e o processo foi realizado com os melhores 20% do total de soluções, representado por *L*. Altera-se o número do cluster de cada amostra na sequência de solução com a probabilidade $p = 0.05$. Sendo *Sk* uma das soluções presente em *L*, onde $k = 1,\ldots, L$, o processo de busca local adotado neste trabalho é apresentado no Quadro 8.

Após realizar a operação de busca local, a matriz de feromônios é atualizada. Esse processo de atualização de feromônios reflete a utilidade das informações dinâmicas fornecidas pelas formigas. Assim, a matriz de feromônios é uma espécie de memória adaptativa que contém informações provenientes das soluções superiores previamente encontradas, e é atualizada no final de cada iteração. O processo de atualização da trilha aplicado neste algoritmo considera as soluções descobertas pelas formigas, de acordo com o critério adotado no nível de iteração *t*. Esses *L* agentes imitam a deposição da trilha de feromônios de formigas reais atribuindo números reais τ_{ij} associados aos atributos da solução. Sendo ρ a taxa de evaporação e $\Delta\tau_{ij}^{l}$ a quantidade de feromônio que a l-ésima

formiga deposita na aresta (*i, j*), as informações da trilha são atualizadas usando a seguinte regra:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \sum_{l=1}^{L} \Delta\tau_{ij}^{l}, i = 1, \dots, n, \; j = 1, \dots, k. \quad (14)$$

Quadro 8: Algoritmo de Busca Local de ACO.

Início
    Inicialize com *k* = 1
    Enquanto k ≤ L faça:
        Inicialize *St* como uma solução temporária com os mesmos valores de *Sk*
        Para cada elemento *i* de *St* faça:
            Gere um número aleatório *r* no intervalo (0, 1)
            Se *r* ≤ *p* então
                Selecione aleatoriamente um número inteiro *j* no intervalo [1, *k*] tal que *Sk(i)* ≠ *j*
                Atribua *St(i) = j*
                Calcule os centros dos clusters e os pesos associados à sequência de solução *St*
                Calcule o valor da função objetivo *Ft*.
        Se *Ft* < *Fk* então
            Atualize *Sk* = *St* e *Fk* = *Ft*
        k = k + 1
    Retorna Sk como a melhor solução encontrada
Fim

Após realizar a operação de busca local, a matriz de feromônios é atualizada. Esse processo de atualização de feromônios reflete a utilidade das informações dinâmicas fornecidas pelas formigas. Assim, a matriz de feromônios é uma espécie de memória adaptativa que contém informações provenientes das soluções superiores previamente encontradas, e é atualizada no final de cada iteração. O processo de atualização da trilha aplicado neste algoritmo considera as soluções descobertas pelas formigas, de acordo com o critério adotado no nível de iteração *t*. Esses *L* agentes imitam a deposição da trilha de feromônios de formigas reais atribuindo números reais τ_{ij} associados aos atributos da solução. Sendo ρ a taxa de evaporação e $\Delta\tau_{ij}^{l}$ a quantidade de feromônio que a l-ésima formiga deposita na aresta (*i, j*), as informações da trilha são atualizadas usando a seguinte regra:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \sum_{l=1}^{L} \Delta\tau_{ij}^{l}, i = 1, \dots, n, \; j = 1, \dots, k. \quad (14)$$

Os agentes constroem suas soluções aplicando as informações fornecidas

pela matriz de feromônios atualizada no final da iteração *t* - 1. Para gerar uma solução *S*, o agente seleciona o número do cluster para cada elemento da solução *S* de uma das seguintes formas:

(i) usando probabilidade q_0 (q_0 sendo um número definido 0 < q_0 < 1), é escolhido o cluster com a maior concentração de feromônios.

(ii) um dos *k* clusters é escolhido usando uma distribuição estocástica com uma probabilidade (1 – q_0), denotada como p_{ij}.

A primeira forma (i) é conhecida como explotação, enquanto a segunda forma (ii) é denominada exploração. Em (ii), a escolha é feita por qualquer um dos três clusters com uma probabilidade normalizada de feromônios (probabilidade de feromônios normalizada para 1), dada por:

$$p_{ij} = \frac{\tau_{ij}}{\sum_{j=1}^{K} \tau_{ij}}, \; j = 1, \ldots, K. \tag{15}$$

3.2 Função *Fitness*

Conforme o trabalho de Ahmadyfard & Modares (2008), para clusterização por meio de meta-heurísticas uma função de *fitness* pode ser dada por:

$$f(i) = M/N_p, \tag{16}$$

$$M = \sum_{j=1}^{k} \sum_{\forall x_p \in C_j} D\,(x_p, z_j)$$

em que:

- $f(i)$: valor da função *fitness* para a partícula i, que representa a qualidade desta solução associada ao centróide.
- M: é a soma da distância entre os centróides c_i e os pontos que pertencem ao cluster específico C_j, onde j varia de 1 a K (número total de clusters).
- D: é a medida de distância (Euclidiana, Manhattan ou Chebyshev) entre os pontos x_p e os respectivos centróides z_j do seu cluster C_j.

3.3 Codificação das soluções

A forma de codificação utilizada em metaheurísticas impacta significativamente sua performance, podendo influenciar tanto a eficiência computacional quanto a qualidade das soluções encontradas. Barceló-Rico, Díez e Bondia (2011) mencionam que a escolha entre essas formas de codificação

depende não somente a meta-heurística em si, mas também da natureza dos dados, da complexidade do problema e, inclusive, dos recursos computacionais disponíveis.

Segundo Das, Abrahan e Konar (2009), no agrupamento de dados por meta-heurísticas, as codificaçações de soluções mais comuns são a por particação, que indica a qual cluster cada ponto pertence, por conjunto, que indica os clusters e seus respectivos pontos, e por centroide, demonstrando as coordenadas dos centróides. Como visto anteriormente, a AG e PSO utilizam a codificação por centróide, enquanto a ACO utiliza a por partição.

A codificação por partição é amplamente utilizada em algoritmos como o K-means, mas é computacionalmente cara para grandes conjuntos de dados, devido ao número de comparações necessárias para atualizar as atribuições dos pontos aos clusters durante as iterações do algoritmo (Menendez, 2021). Essa característica afeta a velocidade de convergência e a escalabilidade da metaheurística, especialmente em problemas com alta dimensionalidade ou complexidade.

Ainda segundo Menendez (2021), a codificação por centróide pode ser mais eficiente em termos de memória e tempo de computação para grandes conjuntos de dados, especialmente quando a dimensionalidade dos dados é alta e a precisão dos centróides é crucial para a definição dos clusters. Ela reduz a quantidade de informações a serem manipuladas, armazenando apenas as coordenadas dos centroides em vez de todos os pontos de dados individuais, o que acelera o processo de busca e permiti o tratamento mais eficiente de conjuntos de dados grandes.

4 METODOLOGIA

Neste capítulo é apresentada a metodologia utilizada para o agrupamento de dados. A metodologia seguiu as etapas usuais do processo de clusterização, conforme apresentado na Seção 2.1.1. Neste caso, a abordagem proposta envolveu pré-processamento dos dados, implementação das meta-heurísticas Algoritmo Genético, Enxame de Partículas e Colônia de Formigas e avaliação dos resultados obtidos, como apresenta a Figura 9.

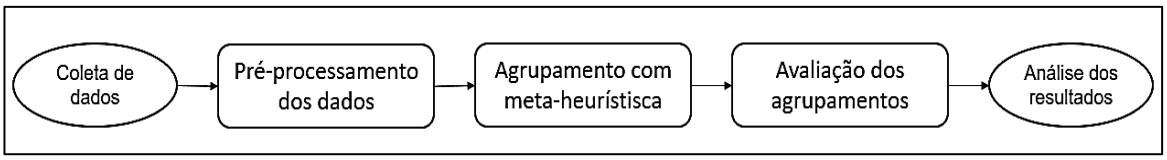


Figura 9: Fluxograma para o agrupamento de dados.
Fonte: Autoria própria (2024).

4.1 Pré-processamento

O pré-processamento dos dados desempenhou um papel crucial no preparo dos dados brutos para a etapa de análise. Foi realizada uma série de tarefas, como limpeza dos dados, tratamento de valores ausentes, padronização dos dados e análise de componentes principais, seguindo o apresentado no Capítulo 2.

A padronização foi realizada por meio da classe “Standard Scaler” da biblioteca Scikit-Learn, que realiza o cálculo do valor de *z* para cada uma das variáveis. Esta biblioteca fornece uma variedade de ferramentas para realizar tarefas relacionadas ao aprendizado de máquina. Já a PCA foi realizada por meio da biblioteca Scikit-Learn e seu passo a passo está apresentado no Quadro 9.

Quadro 9: Algoritmo para PCA em Python.

Entrada: conjunto de dados contendo *n* amostras, com *d* dimensões.
Início
Inicialize a classe PCA
Utilize o método *explained_variance_ratio* para determinar o número ideal de componentes.
Ajuste o modelo utilizando o método *fit*.
Aplique a transformação utilizando o método *transform.*
Retorna conjunto de dados com as dimensões reduzidas.
Fim

4.2 Processamento das meta-heurísticas

As meta-heurísticas AG, PSO e ACO, seguindo os algoritmos apresentados no Capítulo 3, foram processadas individualmente considerando cada umas das métricas de distância apresentadas na Seção 2.1.2. A Figura 10 apresenta um fluxograma das etapas do processamento, que se iniciou com a escolha das três meta-heurísticas supracitadas, definição dos parâmetros e o processamento inicial dos algoritmos e, então, o ajuste dos parâmetros e processamento final dos algoritmos.

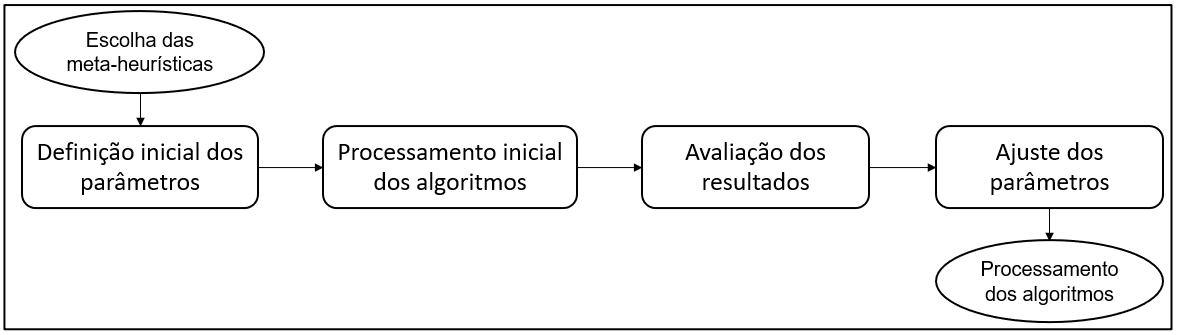


Figura 10: Fluxograma para o processamento das meta-heurísticas.
Fonte: Autoria própria (2024).

Cada meta-heurística possui um conjunto específico de parâmetros que influenciam diretamente a sua eficácia. Assim, com a finalidade de identificar os valores que proporcionam o melhor desempenho para cada meta-heurística, considerando a qualidade das soluções encontradas e o tempo de processamento requerido, foram definidos, testados e avaliados diferentes combinações de valores. As definições finais dos parâmetros de cada meta-heurística são apresentadas no Capítulo 5, referente aos experimentos computacionais.

4.3 Avaliação do agrupamento

A qualidade dos particionamentos obtidos por cada meta-heurísticas foi avaliada considerando os índices Silhueta (SH), Davies-Bouldin (DB), Calinski-Harabasz (CH) e o de Rand Ajustado (ARI), conforme apresentados na Seção 2.1.3. Como os conjuntos de dados utilizados possuem características distintas, eles foram divididos em três grupos e a forma de aplicação dos índices de avaliação foi diferente para como um dos grupos.

Tido como um grupo de conjuntos *benchmarks* e apresentado na Seção 5.1,

o primeiro grupo de dados é formado por um conjunto de dados sobre flores, disponível em *https://www.kaggle.com/datasets/uciml/iris*, e um conjunto sobre vinhos, disponível em *https://archive.ics.uci.edu/dataset/109/wine*. Foram utilizados os índices de avaliação SH, DB e CH, uma vez que as classes dos dois conjuntos deste grupo são desconhecidas.

No caso do segundo grupo de dados, formado por 8 conjuntos também *benchmarks*, disponíveis em *https://github.com/deric/clustering-benchmark*, a avaliação se baseou apenas no índice ARI e na análise visual. São sonhecidas todas as classes de cada um dos conjuntos deste grupo, o que permite a utilização deste índice para avaliação do agrupamento. Na Seção 5.2 são apresentados os resultados detalhados das avaliações para cada um dos 8 conjuntos de dados.

Já para o terceiro grupo, referente a dois conjuntos sobre bovinos, os índices SH, DB e CH também foram utilizados para a avaliação dos agrupamentos gerados, mas dessa vez os seus resultados foram utilizados para determinar o melhor número de clusters e a melhor meta-heurística e métrica de distância para gerar o particionamento final. Nesse caso, os valores dos índices obtidos para diferentes quantidades de clusters foram analisados e a meta-heurística e o respectivo número de clusters que obteve a melhor combinação dos três índices foi elegida como a melhor.

5 EXPERIMENTOS COMPUTACIONAIS

Neste capítulo são apresentados os experimentos computacionais realizados para avaliar o desempenho do Algoritmo Genético, Enxame de Partícula (PSO) e Colônia de Formigas (ACO) para a tarefa de agrupamento de dados. Os testes foram conduzidos utilizando uma variedade de conjuntos de dados, incluindo dois conjuntos de dados *benchmark* largamente utilizados pela comunidade científica, que são um conjunto sobre flores e um sobre vinho, e outros oito gerados para simulação controlada, além de dois conjuntos de dados sobre variabilidade de bovinos.

As meta-heurísticas foram implementadas em linguagem Python, utilizando uma variedade de bibliotecas específicas para análise de dados e aprendizado de máquina, tais como NumPy, Scikit-Learn e Matplotlib, tendo sido processados em um ambiente de computação hospedado no Google Colab. Os testes foram realizados em um computador equipado com um processador Intel Core i5, 7,89GB de RAM e sistema operacional Windows.

5.1 Base de dados *benchmark*

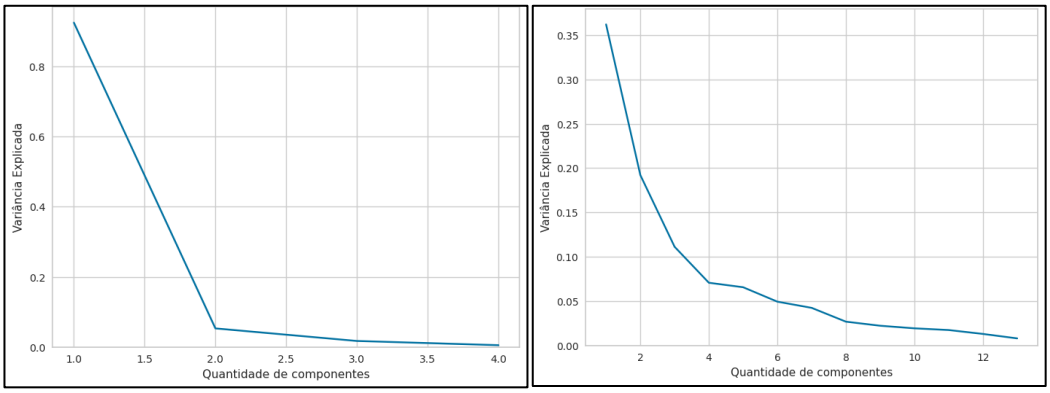
Os testes iniciais foram conduzidos com o objetivo de compreender o comportamento das meta-heurísticas e explorar as configurações dos parâmetros específicos de cada uma delas. Esta fase inicial de experimentação visou obter *insights* sobre como as meta-heurísticas se adaptariam aos conjuntos de dados, bem como determinar as configurações ótimas dos parâmetros para otimizar o desempenho. Considerando que o resultado obtido varia diretamente em função dos parâmetros definidos, essa análise preliminar foi fundamental para o refinamento posterior dos algoritmos.

Inicialmente, foram testados dois conjuntos de dados distintos: um contendo informações sobre espécies de flores e outro com dados sobre variedades de vinhos. O primeiro conjunto de dados, denominado “Conjunto A”, é referente a espécies de flores e possui 150 amostras e 4 atributos, que são: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. O segundo conjunto de dados, referente aos vinhos, tem 178 amostras e 13 atributos sobre informações como teor alcoólico, acidez, pH, dentre outras características químicas, sendo denominado como “Conjunto B”.

O primeiro passo dos experimentos computacionais foi realizar a padronização de dados nos Conjuntos A e B, que é etapa essencial de pré-processamento. A padronização consistiu em aplicar a transformação nos dados numéricos presentes nos conjuntos de dados por meio da Equação (5), de forma a colocá-los em uma mesma escala, conforme apresentado na Seção 2.1.4.

Outra etapa essencial de pré-processamento envolveu a aplicação da técnica PCA aos conjuntos de dados, para reduzir a complexidade dos atributos numéricos, mantendo as informações críticas para a formação dos clusters, conforme explicado na Seção 2.1.5. Para os Conjuntos A e B, a utilização do PCA permitiu a redução da dimensionalidade dos atributos numéricos, o que, por sua vez, viabilizou a visualização dos dados em um espaço com menor quantidade de dimensões.

As Figuras 11(a) e 11(b) mostram a variância explicada em relação à quantidade de componentes de cada conjunto de dados. Levando em consideração o ponto de inflexão da variância, foi determinado que o número ideal de componentes para o Conjunto A é igual a 2, enquanto para o Conjunto B é igual a 4. Cabe destacar que o Conjunto A possui 4 atributos, o que significa que a sua dimensionalidade original já é relativamente baixa. Apesar disso, a aplicação do PCA foi realizada para fins de análise e exploração, com o intuito de verificar se a redução de dimensionalidade poderia trazer benefícios adicionais no agrupamento deste conjunto.



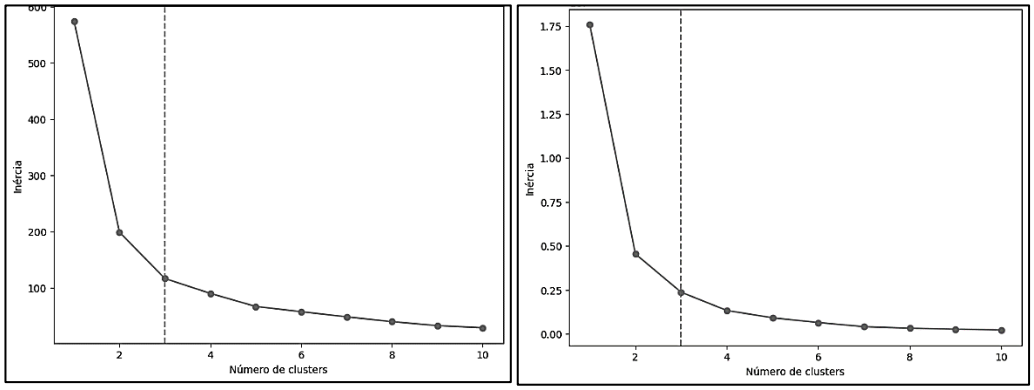
(a) Conjunto A (b) Conjunto B

Figura 11: Gráfico de variância explicada por componentes principais para os conjuntos de dados A e B.

Fonte: Autoria Própria (2024).

A Figura 12 exibe o gráfico do método cotovelo, que demonstra a variação

da inércia em relação ao número de clusters (*k*) considerando os dados após a aplicação do PCA. Lembrando que o cotovelo representa o ponto ideal de clusters, onde a inércia começa a diminuir mais lentamente, indicando o número ótimo de clusters para os seus dados, como descrito na Seção 2.1.3. Nesse caso, o número ideal de clusters para os Conjuntos A e B é igual a 3.



(a) Conjunto A (b) Conjunto B

Figura 12: Gráfico do método cotovelo para os conjuntos de dados A e B.
Fonte: Autoria Própria (2024).

Com a padronização dos dados e a definição do melhor número de clusters para cada um dos conjuntos de dados, foi possível comparar o desempenho das meta-heurísticas em termos de agrupamento e qualidade das soluções considerando diferentes contextos e cenários de aplicação.

Os parâmetros escolhidos foram guiados por uma combinação das definições adotadas nas próprias referências de cada meta-heurística, aliada ao conhecimento prévio do problema e experimentação iterativa. Foi difinido 20 como o número de soluções para as três meta-heurísticas. Para AG, foi definido a roleta como método de selesção, single-point como crossover e 0.1 como taxa de mutação. A probabilidade de busca local, utilizada no processo de busca local realizado em ACO, foi definida como 0.05, foi estabelecido 0.98 como probabilidade de se utilizar o valor das trilhas de feromônios e a taxa de evaporação foi fixada em 0.01. Para PSO, o valor da inércia foi 1.0, o coeficiente cognitivo (C_1) e social (C_2) foram definidos como 1.5. Todos os valores definidos dos parâmetros para cada meta-heurística são apresentados na Tabela 1.

Tabela 1: Valores dos parâmetros para cada meta-heurística.

AG		ACO		PSO	
Parâmetro	Valor	Parâmetro	Valor	Parâmetro	Valor
População	20	Formigas	20	Partículas	20
Seleção	Roleta	Probabilidade de busca local (*P*)	0.05	Inércia (w)	1.0
Crossover	*Single-point*	Probabilidade para utilizar o valor das trilhas (q_0)	0.98	Coeficiente cognitivo (C_1)	1.5
Taxa de mutação	0.1	Taxa de evaporação (ρ)	0.01	Coeficiente social (C_2)	1.5
Nº máximo de iterações	1000	Nº máximo de iterações	1000	Nº máximo de iterações	1000

O critério de parada adotado em todas as meta-heurísticas foi um número máximo 1000 de iterações, considerando que os testes preliminares mostraram que a maioria das melhorias significativas na qualidade da solução ocorria dentro desse intervalo. Optar por um número máximo de iterações é uma maneira de garantir que o algoritmo de otimização não seja processado indefinidamente. Ao adotar esse critério de parada, uma análise visual pode oferecer *insights* importantes em relação à evolução da função objetivo ao longo das iterações, sendo possível verificar se o algoritmo está convergindo em direção a um valor mínimo ou máximo de maneira estável ou se está demonstrando muitas flutuações.

As Figuras de 13 a 15 e de 16 a 18 demonstram o desempenho da função objetivo para cada uma das meta-heurísticas considerando as diferentes métricas de distâncias para os conjuntos A e B, respectivamente. Essas imagens ilustram como cada algoritmo evoluiu ao longo das iterações em busca de soluções otimizadas, possibilitando observar como a função objetivo se modifica (eixo y) em resposta às iterações executadas (eixo x). Os gráficos demonstram a melhor (azul) e a pior (laranja) solução, bem como a média (verde) entre todas as soluções geradas em cada iteração.

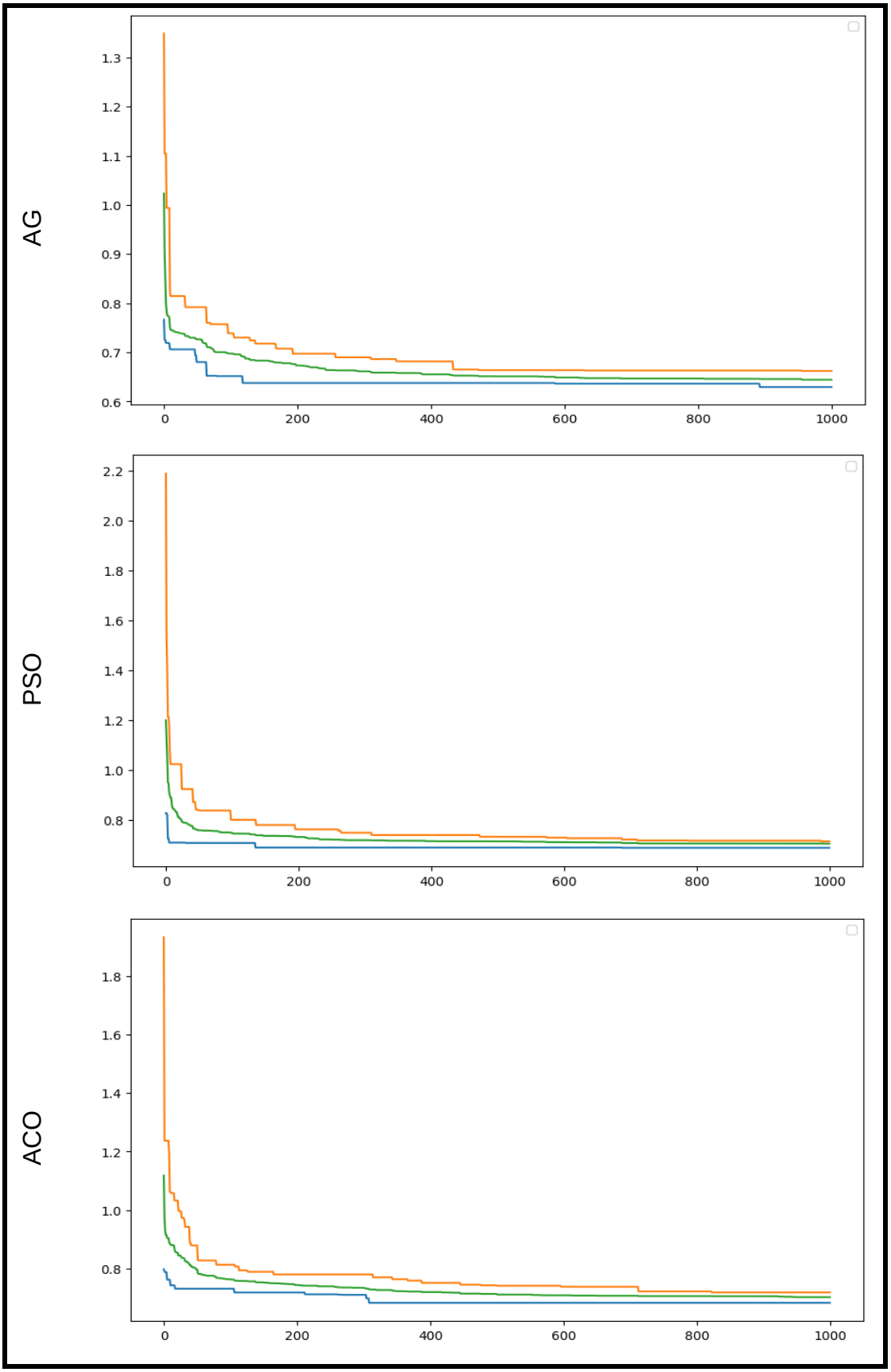


Figura 13: Evolução da função objetivo de cada meta-heurística considerando a métrica Distância Euclidiana para o Conjunto A.

Fonte: Autoria Própria (2024).

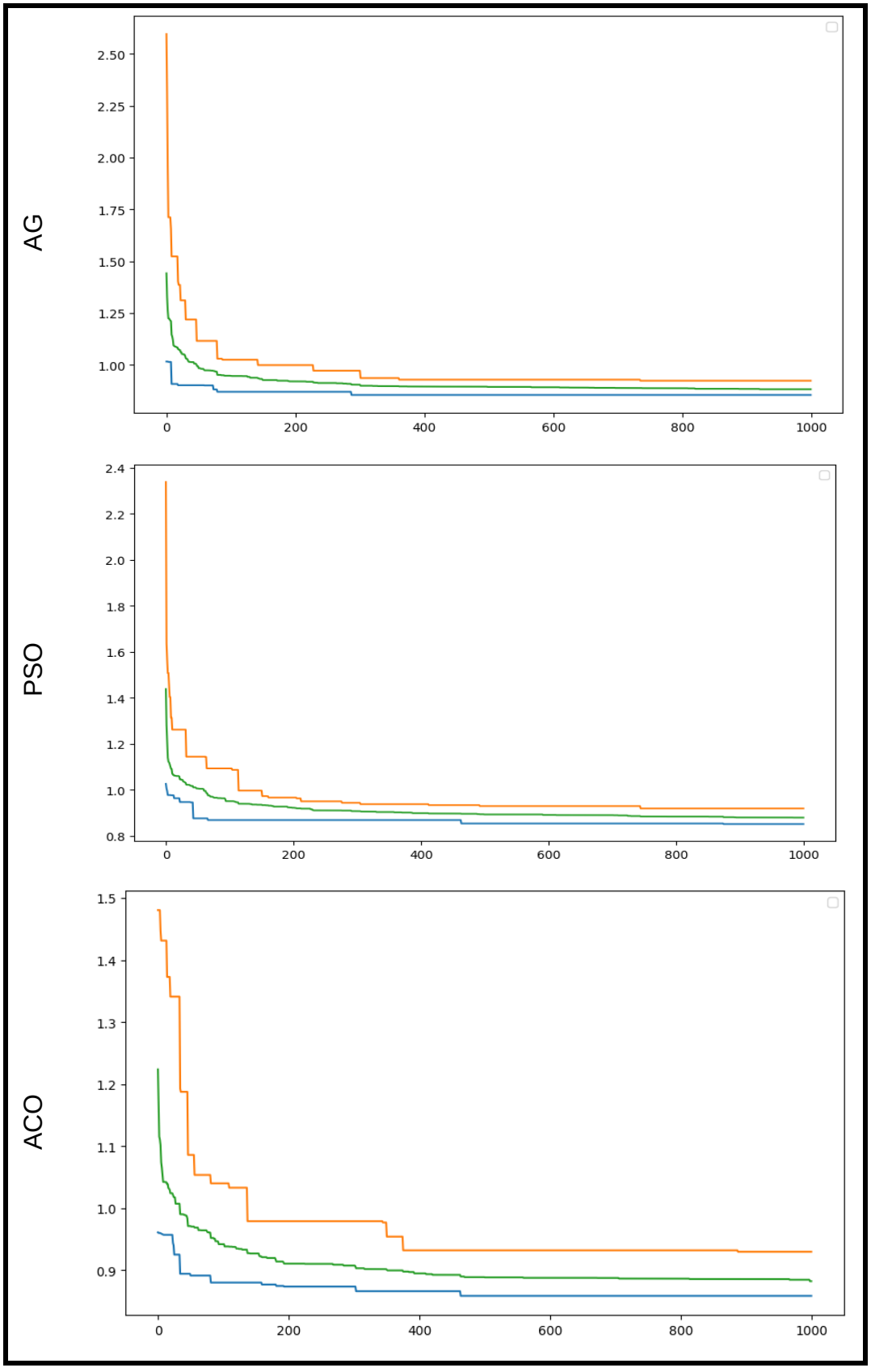


Figura 14: Evolução da função objetivo de cada meta-heurística considerando a métrica Distância Manhathan para o Conjunto A.

Fonte: Autoria Própria (2024).

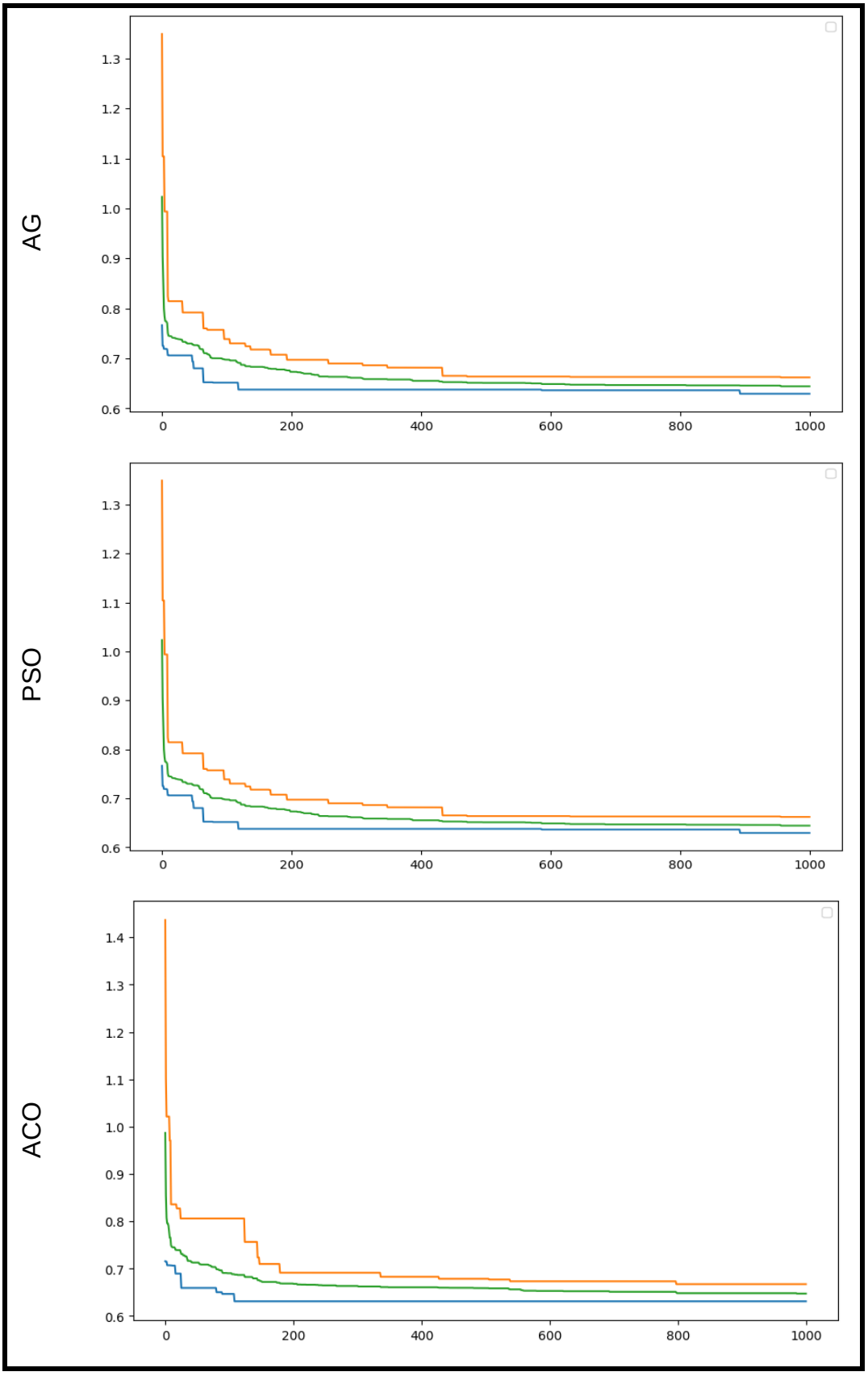


Figura 15: Evolução da função objetivo de cada meta-heurística considerando a métrica Distância Chebyshev para o Conjunto A.

Fonte: Autoria Própria (2024).

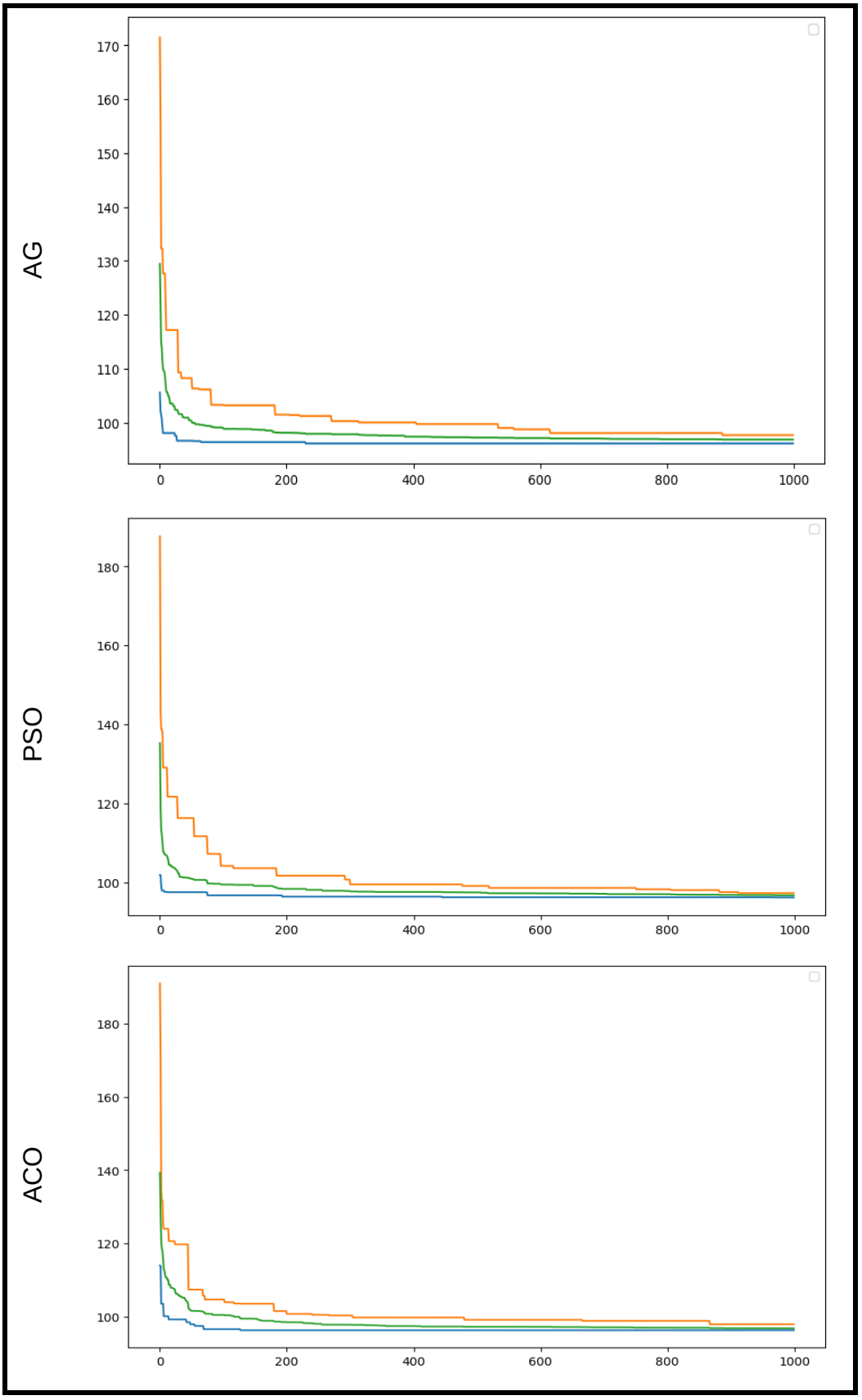


Figura 16: Evolução da função objetivo de cada meta-heurística considerando a métrica Distância Euclidiana para o Conjunto B.

Fonte: Autoria Própria (2024).

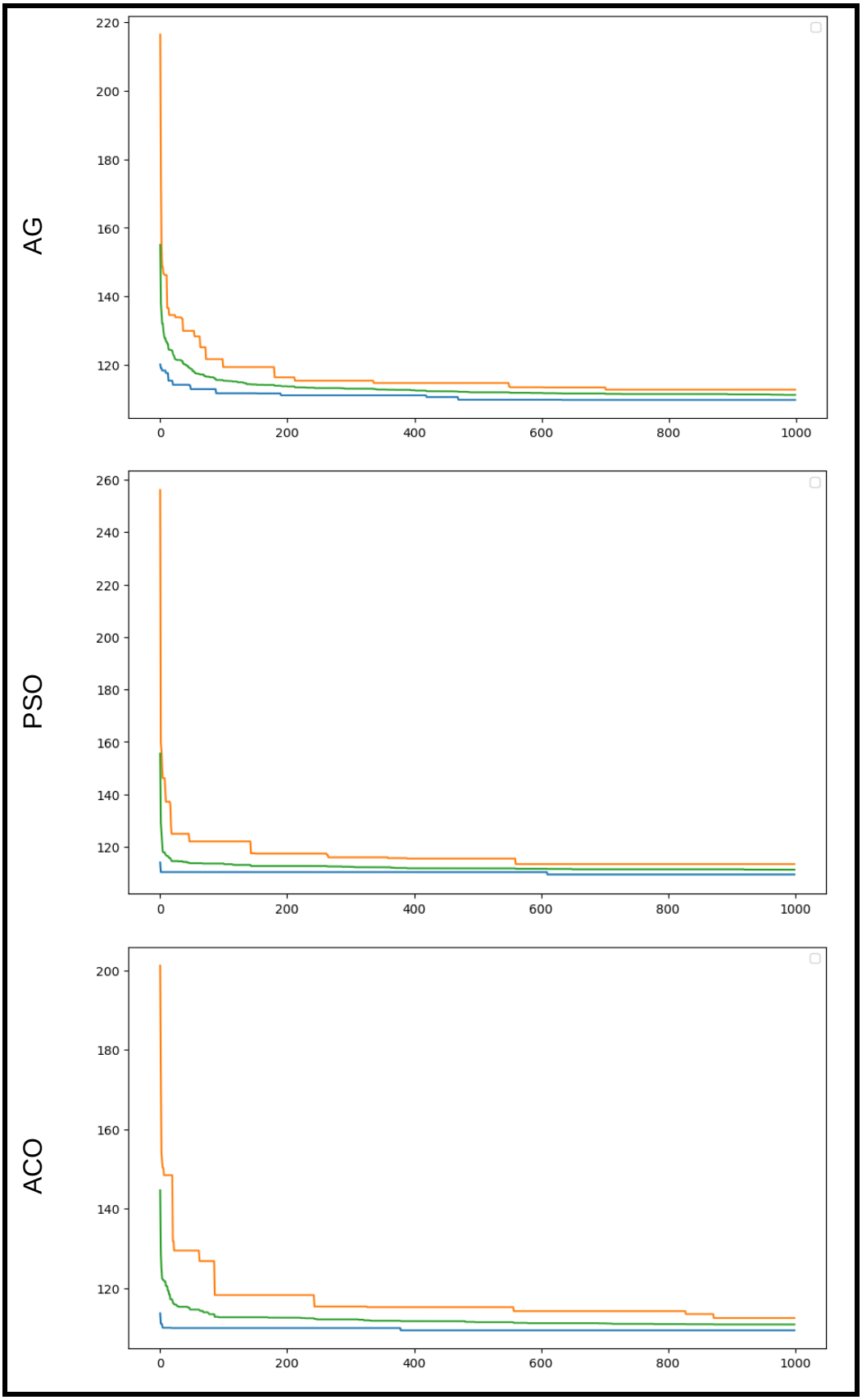


Figura 17: Evolução da função objetivo de cada meta-heurística considerando a métrica Distância Manhathan para o Conjunto B.

Fonte: Autoria Própria (2024).

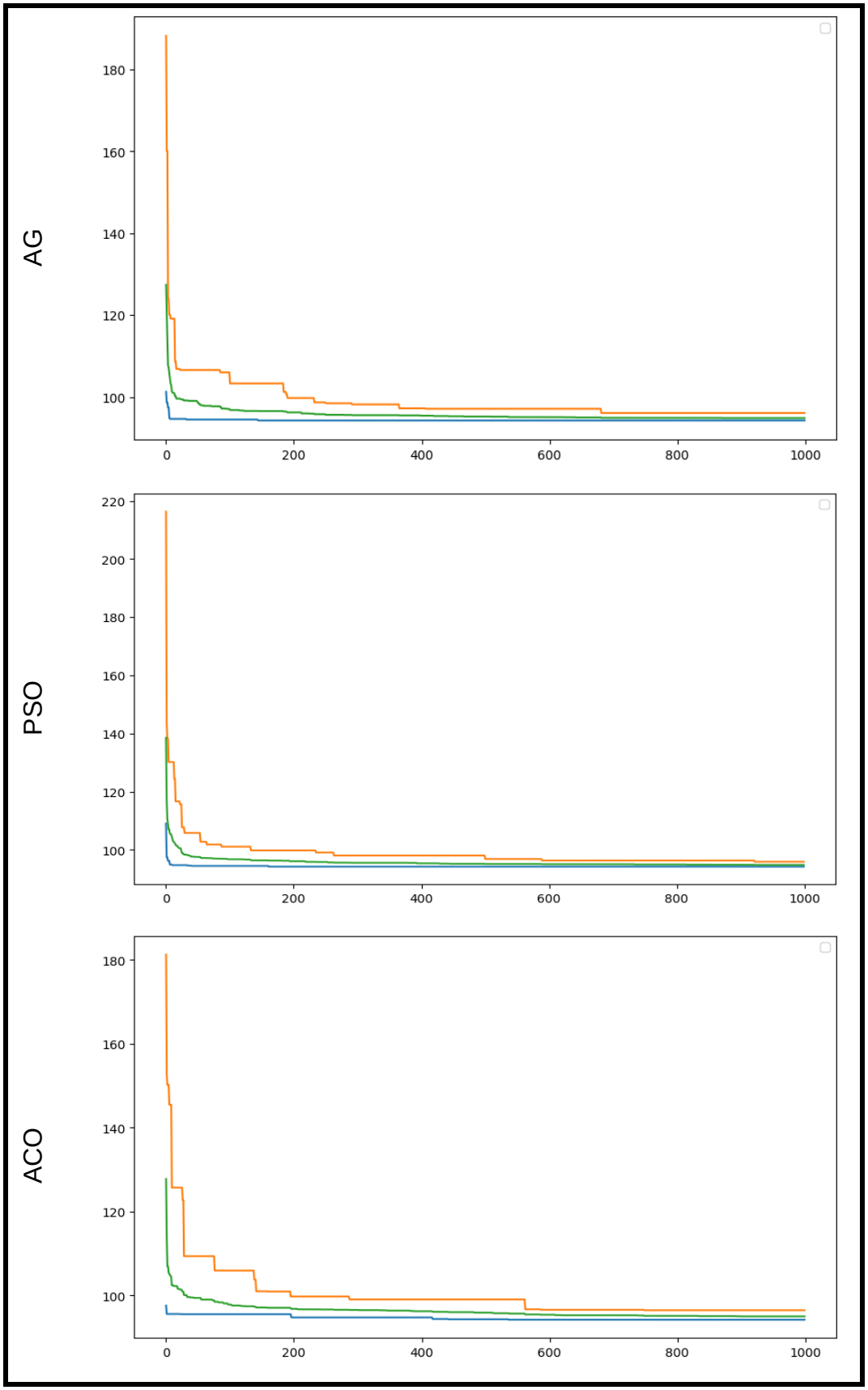


Figura 18: Evolução da função objetivo de cada meta-heurística considerando a métrica Distância Chebyshev para o Conjunto B.

Fonte: Autoria Própria (2024).

Ao analisar os gráficos referente ao agrupamento do Conjunto A e B, observa-se que as três meta-heurísticas foram capazes de identificar suas melhores soluções de forma rápida, requerendo um número reduzido de iterações em relação ao máximo permitido. Isso é evidenciado pelo fato de que as linhas que representam as melhores soluções atingem uma estabilização notável em um curto espaço de tempo.

O gráfico de evolução destaca ainda uma importante dinâmica das meta-heurísticas, que é a exploração e explotação. Enquanto a exploração permite que o algoritmo busque novas soluções potencialmente ótimas, a explotação garante que ele aproveite ao máximo as soluções já encontradas. A distância considerável entre a melhor solução e a pior solução, bem como a média das soluções de cada iteração, sugere uma exploração ativa e diversificada do espaço de busca. Essa diversidade é essencial para a exploração, permitindo que o algoritmo investigue diferentes regiões em busca de soluções ainda não descobertas. Por outro lado, a evolução da melhor solução indica que o algoritmo também está se concentrando em refinar e aprimorá-la, caracterizando seu processo de explotação.

Uma análise comparativa dos gráficos referente aos Conjuntos A e B revela que as três meta-heurísticas apresentaram um comportamento semelhante para ambos os conjuntos de dados, independente da métrica de distância utilizada. Os gráficos exibem padrões similares de convergência e flutuações nos dois contextos analisados. Isso sugere estabilidade nas três meta-heurísticas em lidar com as diferentes características dos dados, resultando em um comportamento de otimização mais uniforme.

Definidos os parâmetros, os algoritmos foram processados 10 vezes individualmente para seus próprios testes de eficiência, cada vez com soluções iniciais geradas aleatoriamente. A avaliação dos resultados para cada conjunto de dados é baseada na média dos índices Silhueta (SH), Davies-Bouldin (DB), e Calinski-Harabasz (CH) das 10 soluções encontradas e no tempo médio de processamento para o número máximo de 1000 iterações.

A Figura 19 apresenta os particionamentos que obtiveram os melhores resultados para os índices analisados para cada meta-heurística, sendo possível observar que todos são idênticos. A concordância dos particionamentos gerados mostra que as meta-heurísticas convergiram para a mesma solução ótima. Isso é reafirmado pelo fato de que os resultados apresentados na Tabela 2 apontam para

similaridades significativas entre os agrupamentos, independentemente da meta-heurística ou da métrica de distância utilizada.

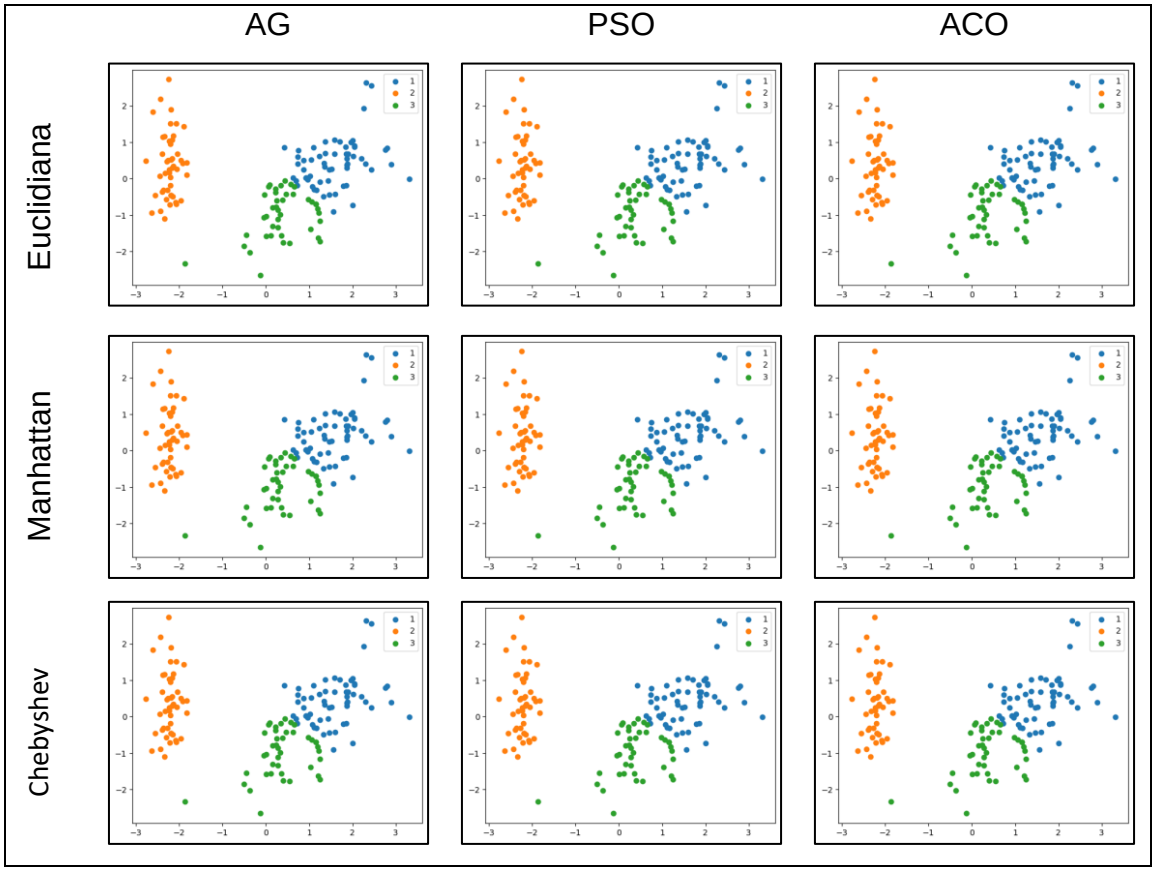


Figura 19: Particionamento obtido por cada meta-heurística considerando diferentes métricas de distância para o Conjunto A.
Fonte: Autoria Própria (2024).

A Tabela 2 apresenta a média dos resultados obtidos a partir de 10 processamentos de cada uma das três meta-heurísticas para o conjunto A. Como métrica de avaliação da qualidade dos agrupamentos gerados, tem-se como melhores agrupamentos os que alcançaram valores mais próximos de 1 para o índice SH, mais próximo de 0 para DB e mais alto para CH. Os resultados para as três meta-heurísticas utilizando as três medidas de distância foram os mesmos, com valores de 0,613 para o índice SH, 0,548 para o índice DB, e 280,12 para o índice CH.

Considerando os resultados obtidos, o valor de 0,613 para o SH sugere que os objetos nos clusters estão relativamente próximos aos objetos de seus próprios clusters e distantes dos objetos de outros clusters, sendo uma indicação positiva

de que os clusters têm uma boa coesão interna. O valor de 0,548 para o DB indica que a dispersão entre os clusters é um pouco maior em relação à dispersão dentro dos clusters, mas ainda é relativamente baixa, o que significa que os clusters têm uma boa separação. O CH de 280,12 corrobora como indicativo de clusters bem definidos e separados.

Tabela 2: Resultados obtidos pelas três meta-heurísticas para o Conjunto A.

Método	Índice de avaliação			CPU time (s)
	SH	DB	CH	
AG – Euclidiana	0,613	0,548	280,12	2,43
AG – Manhattan	0,613	0,548	280,12	2,44
AG – Chebyshev	0,613	0,548	280,12	2,51
PSO – Euclidiana	0,613	0,548	280,12	2,17
PSO – Manhattan	0,613	0,548	280,12	2,34
PSO - Chebyshev	0,613	0,548	280,12	2,3
ACO – Euclidiana	0,613	0,548	280,12	32,13
ACO – Manhattan	0,613	0,548	280,12	34,02
ACO – Chebyshev	0,613	0,548	280,12	31,24

A similaridade entre as soluções geradas indica que esta solução é, de fato, um ótimo global, sendo a melhor solução possível para o Conjunto A. Assim, ainda que os valores obtidos para os índices SH, DB e CH estejam distantes do ideal teórico, os resultados para estes índices parecem ser os melhores possíveis dadas as características deste conjunto de dados. Nesse caso, mesmo que os valores dos índices de qualidade dos agrupamentos não atinjam os ideais teóricos, a solução obtida ainda pode ser considerada relevante e representativa para este conjunto.

Considerando que as três meta-heurísticas encontraram a mesma solução utilizando as três medidas de distância, pode-se afirmar que há uma consistência notável nos resultados de cada uma, independente da medida de distância escolhida. A convergência para a mesma solução sugere que as meta-heurísticas foram capazes de explorar efetivamente o espaço de busca com suas diferentes maneiras de exploração, adaptando-se de forma consistente às características de cada medida de distância e conseguindo capturar efetivamente a estrutura

subjacente dos dados.

Observa-se que a única diferença entre os resultados das meta-heurísticas foi o seu tempo de processamento. Para o conjunto de dados A, as meta-heurísticas AG e PSO necessitaram de tempos entre 2 e 3 segundos para cada uma das medidas de distância utilizadas. Por sua vez, a ACO necessitou de mais tempo para realizar as 1000 iterações, ficando na casa dos 30 segundos para completar seu processamento para cada uma das três medidas de distância.

Semelhante ao apresentado para o Conjunto A, a Figura 20 traz o particionamento para o conjunto de dados B que obteve o melhor resultado para os índices considerados em análise, entre os 10 processamentos de cada meta-heurística utilizando cada uma das diferentes métricas de distância. É possível perceber, novamente, a semelhança entre todos os particionamentos apresentados, o que evidencia que as meta-heurísticas, ao menos uma vez, alcançaram a mesma solução.

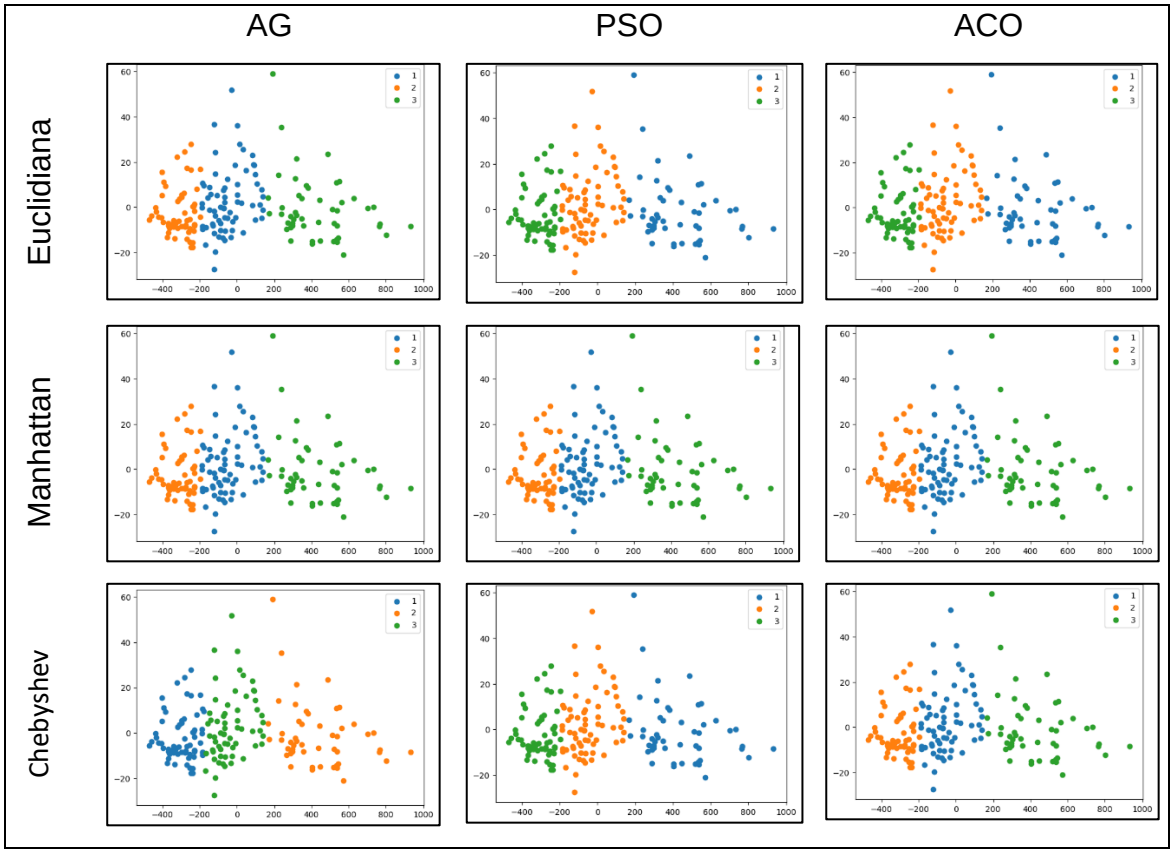


Figura 20: Particionamento obtido por cada meta-heurística considerando diferentes métricas de distância para o Conjunto B.

Fonte: Autoria Própria (2024)

A Tabela 3 apresenta os resultados da média dos índices obtidos para o

Conjunto B. A melhor média de soluções encontrada foi 0,563 para SH, 0,529 para DB e 533,10 para CH, alcançada pelas meta-heurísticas AG e PSO utilizando as três medidas de distância. Ainda considerando a média das soluções, a ACO apresentou resultados menos satisfatórios, sendo seu o melhor valor encontrado utilizando a distância Chebyshev, com valores de 0,561 para SH, 0,535 para DB e 526,10 para CH.

Os valores do índice SH para todas as meta-heurísticas sugerem que os agrupamentos para o conjunto de dados B têm uma coesão interna razoável, mas com alguma sobreposição ou falta de clareza em relação aos limites de cada cluster. Isso significa que, embora os pontos dentro de cada cluster estejam relativamente próximos uns dos outros, pode haver alguma interseção ou ambiguidade nas áreas onde os clusters se sobrepõem.

Por outro lado, a boa separação entre os clusters, conforme indicada pelo valor do índice DB, sugere que os clusters estão distintamente separados uns dos outros. Isso é um indicativo de que, apesar da possível sobreposição indicada pelo SH, os clusters são, em sua maioria, bem definidos e separados. Além disso, é importante mencionar que o índice CH também sugere uma boa separação entre os clusters, com um valor que reflete uma estrutura global bem definida nos dados.

Tabela 3: Resultados obtidos pelas três meta-heurísticas para o Conjunto B.

Método	Índice de avaliação			CPU time (s)
	SH	DB	CH	
AG – Euclidiana	0,563	0,529	533,10	4,23
AG – Manhattan	0,563	0,529	533,10	3,56
AG – Chebyshev	0,563	0,529	533,10	4,02
PSO – Euclidiana	0,563	0,529	533,10	3,58
PSO – Manhattan	0,563	0,529	533,10	3,49
PSO - Chebyshev	0,563	0,529	533,10	4,00
ACO – Euclidiana	0,555	0,540	509,08	45,21
ACO – Manhattan	0,556	0,538	511,17	45,24
ACO – Chebyshev	0,561	0,535	526,10	45,54

Cabe destacar que a ACO alcançou resultados iguais aos da AG e PSO em alguns dos processamentos, mas também apresentou particionamento sutilmente

distintos em outros, com pontos alocados em clusters diferentes das demais meta-heurísticas, o que gerou a diferença na sua média final. Essa diferença na atribuição de pontos não é incomum em análises de clusterização e pode ocorrer devido a nuances na maneira como os algoritmos operam, uma vez que as meta-heurísticas não são determinísticas e pequenas variações na inicialização ou nos parâmetros podem levar a resultados ligeiramente diferentes.

Tendo em vista que a AG e PSO atingiram os melhores resultados utilizando as distâncias Euclidiana, Manhattan e Chebyshev e a ACO não alcançou os mesmos resultados utilizando as mesmas medidas de distância, evidencia-se que as métricas não foram o que impactaram os resultados obtidos, mas sim as características de busca das próprias meta-heurísticas. Entende-se, neste caso, que a determinação do mesmo critério de parada para as três foi o que levou a resultados distintos, já que 1000 iterações não foi o suficiente para a ACO.

Embora sejam diferentes, as três métricas de distância foram eficazes na análise dos mesmos dados, tanto para o Conjunto A quanto para o Conjunto B. Cada métrica conseguiu capturar os aspectos relevantes da estrutura dos dados e se adaptar às características específicas de cada conjunto. Isso se deve à flexibilidade dessas métricas e à sua capacidade de representar diferentes aspectos de similaridade entre os pontos, proporcionando uma representação adequada das relações entre eles.

Novamente, houve diferença no tempo de processamento associado a cada uma das meta-heurísticas. Para o conjunto B, a AG e PSO tiveram como média de tempo de processamento valores variando entre 3 e 4 segundos, sendo um tempo pequeno, relativamente. A ACO, por sua vez, ficou com seu tempo de processamento médio em torno de 45 segundos, sendo também um tempo relativamente pequeno, mas maior que o das outras duas.

Com relação ao tempo de processamento da ACO, vale mencionar que esta meta-heurística, como apresentado na Seção 3.1.3, incorpora um mecanismo de busca local, o qual envolve avaliar e ajustar minuciosamente as soluções parciais em busca de melhorias. Como resultado, a ACO pode levar mais tempo para concluir uma iteração em comparação com métodos que não realizam essa busca local intensiva. Além disso, a ACO utiliza a codificação por partição e, a depender do tamanho do conjuntos de dados, essa forma de codificar as soluções se torna computacionalmente cara, já que o número de soluções cresce significativamente

e requer a atualização contínua das posições dos pontos em relação aos clusters.

Em termos gerais, resultados onde as diferentes meta-heurísticas geraram os índices e particionamento idênticos (Conjunto A) e muito semelhantes (Conjunto B), reforçam a robustez dessas abordagens e a sua confiabilidade na identificação de agrupamentos. Esses resultados também evidenciam a eficácia de cada meta-heurística para a tarefa de agrupamento, em que, entre as três avaliadas, a AG e PSO demonstraram ser as mais eficazes, uma vez que obtiveram desempenho semelhante em termos de qualidade dos índices de avaliação e um tempo de processamento menor em comparação com a ACO para ambos os conjuntos de dados.

5.2 Segunda base de dados *benchmark*

Como apresentado na Seção 3.2, o agrupamento realizado pelas meta-heurísticas é baseado na função da distância entre os pontos. Essa característica torna a avaliação do desempenho em relação a diferentes tipos e distribuições de dados uma tarefa essencial, já que a compreensão dos limites e das capacidades das meta-heurísticas em situações que exigem a consideração de distâncias entre pontos é vital para a aplicação bem-sucedida dessas técnicas em uma ampla gama de domínios.

Nesse sentido, dando continuidade aos experimentos computacionais, foram conduzidos testes utilizando dados *benchmark* como parte da avaliação das meta-heurísticas aplicadas, com o objetivo de aprofundar a compreensão sobre o desempenho dessas técnicas em cenários controlados e bem definidos.

Para tal, foram considerados oito conjuntos de dados distintos com classes conhecidas, cada um caracterizado por estruturas e propriedades diversas. A seleção destes conjuntos de dados foi realizada de forma criteriosa, considerando que cada conjunto representa cenários específicos, incorporando variações dimensionais, densidades diversas, morfologias de agrupamento distintas e distribuições peculiares. A Tabela 4 contém informações sobre os 8 conjuntos de dados, sendo elas: o número de itens, de atributos e de classes.

Tabela 4: Descrição dos oito conjuntos *benchmark.*

Conjunto	Nº de itens	Nº de atributos	Nº de classes
1	312	2	3
2	800	2	2
3	1016	2	2
4	2000	2	3
5	211	3	7
6	800	3	2
7	1000	2	2
8	1000	2	2

Dado que as classes dos dados são conhecidas, foi utilizado o índice de avaliação Adjusted Rand Index (ARI), conforme apresentado na Seção 2.1.6, para avaliar a qualidade dos agrupamentos produzidos pelas diferentes meta-heurísticas. A Tabela 5 apresenta os valores obtidos para o índice ARI para os 8 conjuntos de dados, estando ela dividada por cada uma das três meta-heurísticas e a respectiva métrica de distância utilizada.

Tabela 5: Resultados do índice ARI de cada meta-heurística para os oito conjuntos.

Meta-heurística – Distância	Conjunto							
	1	2	3	4	5	6	7	8
AG – Euclidiana	-0,002	1,0	0,767	1,0	0,938	0,072	1,0	0,000
AG – Manhattan	-0,004	1,0	0,763	1,0	0,929	0,072	1,0	0,000
AG – Chebyshev	-0,004	1,0	0,766	1,0	0,935	0,072	1,0	0,001
PSO – Euclidiana	0,007	1,0	0,861	1,0	0,942	0,123	1,0	0,018
PSO – Manhattan	0,012	1,0	0,837	1,0	0945	0,130	1,0	0,130
PSO - Chebyshev	0,020	1,0	0,819	1,0	0,961	0,131	1,0	0,131
ACO – Euclidiana	-0,003	1,0	0,722	0,995	0,922	0,002	1,0	0,002
ACO – Manhattan	-0,002	1,0	0,741	0,984	0,901	0,000	1,0	0,000
ACO – Chebyshev	0,000	1,0	0,758	0,992	0,911	0,002	1,0	0,167

As Figuras de 21 a 28 mostram os particionamentos dos conjuntos de 1 a 8, respectivamente, que obtiveram os melhores índice ARI entre todas as meta-heurísticas com as diferentes métricas de distância, conforme valores

apresentados na Tabela 5. Com o intuito de otimizar a apresentação dos resultados e evitar redundâncias, optou-se por exibir apenas o melhor particionamento de cada conjunto de dados, a fim de dar enfoque aos aspectos mais relevantes das análises.

O Conjunto 1 possui forma de espiral, que é um exemplo clássico de uma estrutura de dados complexa e não linear, em que a distância entre pontos próximos pode variar consideravelmente e os clusters podem ser mais difíceis de se definir. Para esse conjunto, todas as meta-heurísticas obtiveram valores próximos de 0, indicando uma concordância ao acaso. A Figura 21 demonstra o particionamento obtido através da meta-heurística PSO utilizando a distância Chebyshev, que foi a que obteve o melhor índice entre todas.

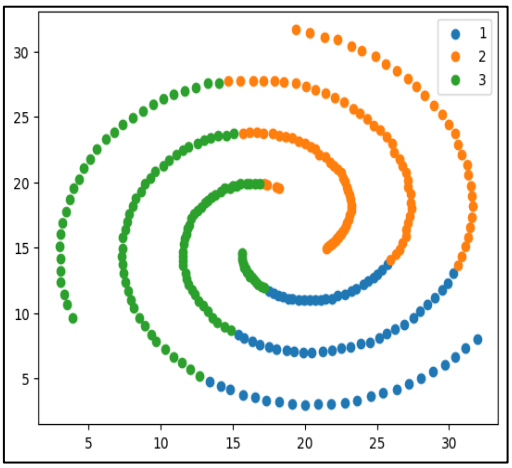


Figura 21: Particionamento obtido para conjunto de dados em forma de espiral.
Fonte: Autoria Própria (2024).

Neste caso, evidencia-se a limitação das três meta-heurísticas em lidar com esse tipo de conjunto de dados. Essa limitação pode ser atribuída às métricas de distância, que favorecem naturalmente a formação de clusters circulares, uma vez que as distâncias são calculadas em relação a um ponto central, o que pode levar a uma representação imprecisa da realidade subjacente dos dados, onde os clusters não capturam completamente a forma espiral.

A Figura 22 ilustra o particionamento obtido para o Conjunto 2, que se caracteriza por uma distribuição uniforme, com pontos dispostos de tal maneira que formam dois losangos distintos no espaço. Para esse caso, todas as meta-heurísticas alcançaram 1 para o índice ARI, tendo sido capazes de identificar e separar os agrupamentos de maneira precisa, criando clusters que refletem

fielmente a estrutura real dos dados.

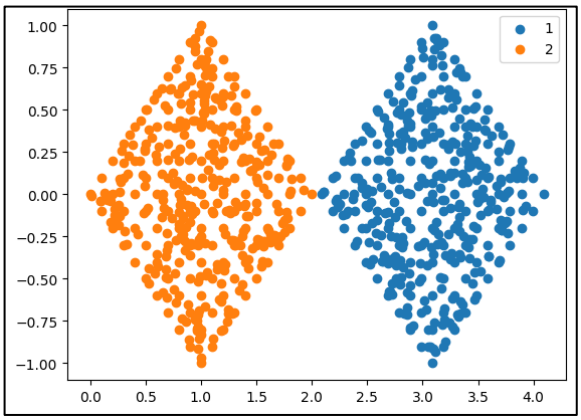


Figura 22: Particionamento obtido para conjunto de dados em forma de losango.
Fonte: Autoria Própria (2024).

A distribuição uniforme e a clara separação entre os padrões facilitaram a tarefa de agrupamento, permitindo que as meta-heurísticas capturassem com precisão a estrutura subjacente dos dados. A identificação precisa dos clusters, refletida pelo índice ARI máximo, indica que as meta-heurísticas foram capazes de adaptar-se eficientemente à natureza específica deste conjunto, mesmo ele apresentando um padrão não convencional.

O Conjunto 3 possui uma estrutura retangular, mas com uma distribuição não uniforme entre cada retângulo, resultando em áreas mais e menos densas dentro de cada um deles. A Figura 23 apresenta o particionamento obtido por meio da meta-heurística PSO utilizando a distância Euclidiana, que obteve o melhor valor para o índice ARI. Importante destacar que os valores de todas as meta-heurísticas para este conjunto ficaram mais próximas de 1 do que de 0, indicando uma boa concordância entre os agrupamentos obtidos e os verdadeiros para as três.

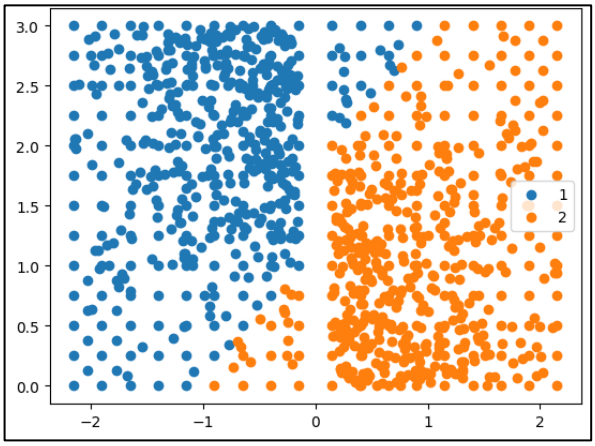


Figura 23: Particionamento obtido para conjunto de dados em forma de retângulo com distribuição não uniforme.
Fonte: Autoria Própria (2024).

Ainda que todas as meta-heurísticas tenham apresentado uma boa concordância com o agrupamento verdadeiro, a falta de regularidade na distribuição representou um obstáculo para o seu desempenho, já que os centros dos clusters foram direcionados para as áreas mais densas, resultando em clusters que não capturaram completamente a estrutura dos dados. Isso é compreensível, dado que os agrupamentos são formados considerando a menor distância entre os centróides e os pontos do seu respectivo cluster e, naturalmente, os centróides são direcionados para as regiões mais densas a fim de diminuir essa distância.

Em muitos casos reais, os dados não seguem uma distribuição uniforme, com certas regiões do espaço, contendo mais pontos do que outras e resultando em densidades distintas. Nesses casos, os clusters em regiões densas podem ser super representados, enquanto clusters em regiões menos densas podem ser negligenciados. Essa heterogeneidade representa mais uma dificuldade com a qual os algoritmos de agrupamento baseados em distância têm que lidar (Thrun, 2021).

A Figura 24 demonstra o particionamento obtido por meio das meta-heurísticas AG e PSO para o Conjunto 4, o qual apresenta formas esféricas com densidades diferentes. Neste caso, mesmo com a diferença na densidade de pontos de cada grupo do conjunto, as duas meta-heurísticas obtiveram o valor 1 para ARI, tendo sido capazes de identificar e separar os agrupamentos de maneira precisa, criando clusters que refletem de maneira adequada a estrutura real dos dados.

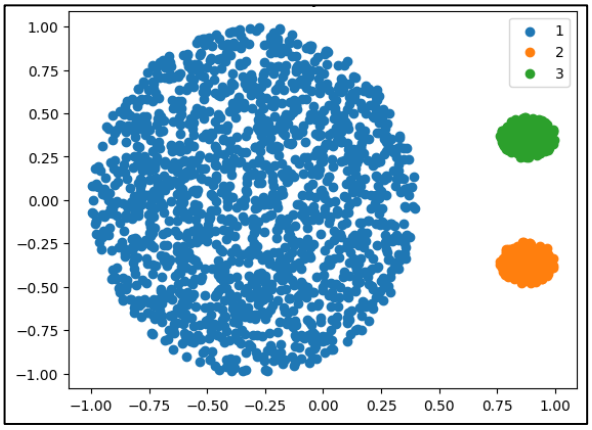


Figura 24: Particionamento obtido para conjunto de dados em forma esférica com densidades distintas.

Fonte: Autoria Própria (2024).

Diferente do Conjunto 3, em que a densidade varia dentro de um mesmo cluster, neste conjunto a densidade dos clusters menores é maior em comparação com ao maior. É importante observar que existe uma distância significativa entre os clusters, indicando uma separação clara e definida entre os grupos, o que torna mais simples para os algoritmos de agrupamento identificar e atribuir corretamente os pontos de dados aos clusters apropriados, já que essa separação espacial entre os grupos impede que os centróides de um cluster influenciem indevidamente o agrupamento de dados a outros clusters.

Nota-se que as meta-heurísticas, de modo geral, conseguiram distinguir essa diferença de densidade entre as áreas e não foram afetadas pela diferença na quantidade de pontos em cada cluster. Vale ressaltar que a meta-heurística ACO obteve valores próximos de 1 para este conjunto com as três métricas de distâncias. Nesse caso, entende-se que a quantidade de iterações determinada não foi suficiente para alcançar a melhor solução, já que as soluções desta meta-heurística são formadas alocando diretamente um ponto ao cluster e não em relação ao centróide. Assim, quanto maior o conjunto de dados, mais processamento e recursos computacionais são exigidos para alocar cada um dos pontos.

Na Figura 25, apresenta-se o particionamento obtido por meio da PSO com distância Chebychev para o Conjunto 5, que possui sete grupos no espaço tridimensional, com alguns desses grupos apresentando maior densidade de pontos em relação aos demais. Conforme a Tabela 5, todas as meta-heurísticas obtiveram valores do índice ARI próximos a 1, indicando que obtiveram um bom

particionamento, ainda que não haja uma concordância perfeita com os dados reais.

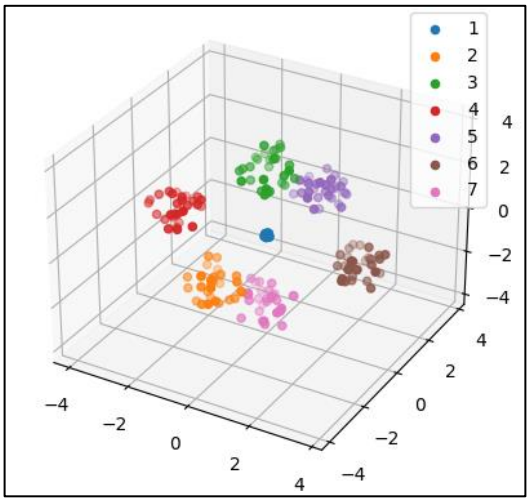


Figura 25: Particionamento obtido para conjunto de dados tridimensional com grupos de densidade variada.
Fonte: Autoria Própria (2024).

Para esse conjunto de dados, observa-se sobreposição entre os clusters 3 e 5 e entre os clusters 2 e 7, os quais estavam mais próximos uns dos outros. Essa sobreposição indica que a fronteira entre os grupos não está claramente definida, o que atribui mais complexidade ao processo de definição do particionamento para os pontos dessa região. Essa falta de clareza na separação entre os clusters traz ambiguidades na atribuição de pontos de dados a um cluster específico, especialmente para aqueles localizados nas áreas de sobreposição, o que tornou o particionamento das meta-heurísticas impreciso para essa região.

A Figura 26 refere-se ao particionamento da PSO com a distância Chebyshev para o Conjunto 6, em que a densidade varia significativamente no espaço tridimensional, com um agrupamento central de alta densidade e diversos pontos ao seu redor. Todas as meta-heurísticas, independentemente da distância utilizada, obtiveram valores próximos a 0 para o índice ARI, indicando uma baixa concordância entre os agrupamentos gerados e os agrupamentos verdadeiros do conjunto de dados.

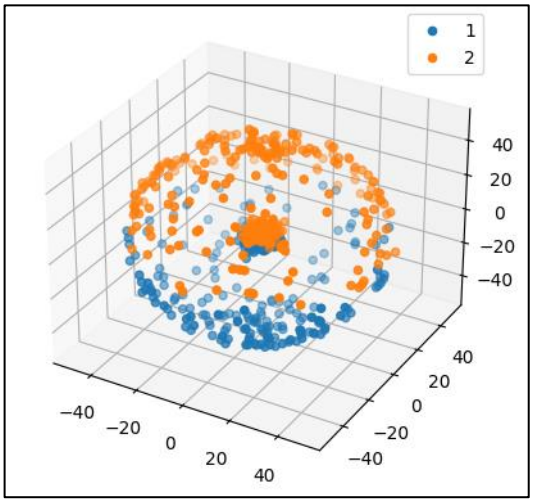


Figura 26: Particionamento obtido para conjunto de dados tridimensional com alta densidade central.

Fonte: Autoria Própria (2024).

Com os pontos distribuídos nesta configuração, pode ser difícil para algoritmos baseados em distância discernir a verdadeira estrutura dos dados, já que acabam sendo influenciados pela distribuição espacial dos pontos e da densidade local. Neste caso, as meta-heurísticas conseguiram agrupar os pontos do grupo central em um mesmo cluster, mas não conseguiram separá-lo dos pontos ao seu redor. Sabendo-se que estas meta-heurísticas usam a distância como critério principal para o agrupamento, é importante reconhecer suas limitações em situações como esta, em que a estrutura dos dados é mais complexa e não pode ser totalmente capturada apenas pela proximidade dos pontos.

As Figuras 27 e 28 fornecem uma ilustração importante sobre como a proximidade e a distância entre os dados influenciam significativamente o processo de agrupamento. Nelas, é apresentado o particionamento obtido para dois conjuntos de dados que exibem a mesma forma subjacente, mas um em que as formas estão mais próximas e outro em que elas estão mais distantes. Mesmo quando os dados compartilham a mesma forma subjacente, a proximidade relativa entre as formas resultou em agrupamentos distintos.

Na Figura 27, é apresentado o particionamento obtido para o Conjunto 7, em que os pontos estão distribuídos de maneira que as duas formas subjacentes estão distantes uma da outra. Para esse conjunto, todas as meta-heurísticas obtiveram 1 para o índice ARI, indicando uma clusterização que reflete com precisão a estrutura inerente dos dados, independente da medida de distância

utilizada.

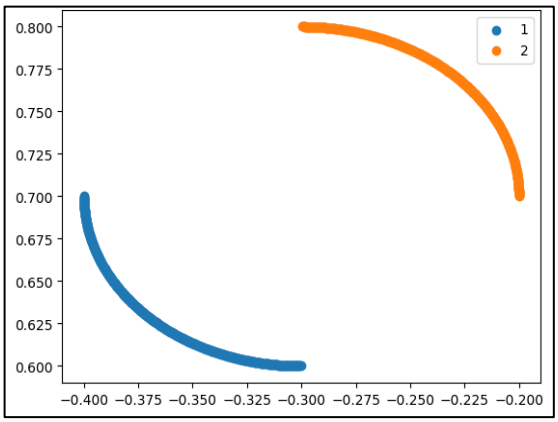


Figura 27: Particionamento obtido para conjunto de dados em forma de arcos distantes.

Fonte: Autoria Própria (2024).

Nesse caso, a distância considerável entre os clusters minimiza a influência de fatores como forma de distribuição e a densidade local entre os grupos, simplificando o processo de agrupamento e permitindo que as meta-heurísticas se concentrem principalmente na separação espacial dos clusters. Como os dados dos dois grupos estão distantes, os centróides de cada agrupamento final também se tornam suficientemente distantes um do outro para que não haja ambiguidade no particionamento.

Por sua vez, a Figura 28 refere-se ao particionamento do Conjunto 8, em que, apesar de sua semelhança com o Conjunto 7, os pontos estão dispostos de maneira que as duas formas estão mais próximas. Como mostra a Tabela 5, para esse conjunto todas as meta-heurísticas apresentaram valores próximos à 0 para o índice ARI, havendo uma baixa concordância entre os agrupamentos produzidos e os agrupamentos esperados.

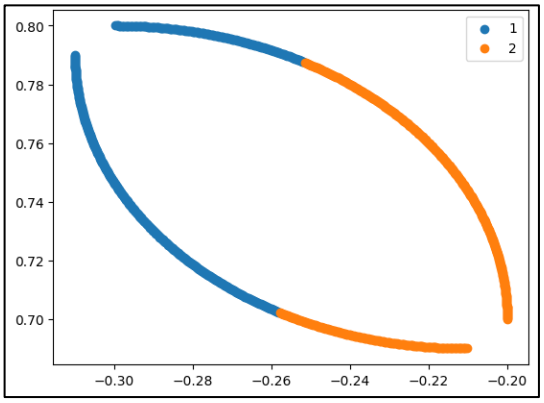


Figura 28: Particionamento obtido para conjunto de dados em forma de arcos próximos.

Fonte: Autoria Própria (2024).

A proximidade entre as formas levou a um agrupamento diferente do esperado, já que a proximidade entre as formas redirecionou os centróides para outra região e os pontos foram agrupados de acordo com sua proximidade imediata ao centróide, resultando em agrupamentos que não refletem a estrutura subjacente dos dados e parecem menos coerentes. Isso ocorre justamente porque os centróides são influenciados pela distribuição espacial dos pontos, movendo-se em direção a áreas com maior densidade ou onde a proximidade entre os pontos é maior.

Considerando os resultados obtidos nestes testes, reitera-se um fator importante para a tarefa de agrupamento, que é a interação entre densidade e distribuição dos dados, os quais são dois conceitos intrinsecamente ligados, porém distintos (Gungor & Ozmen, 2017). Esses resultados enfatizam que, tanto no cenário bidimensional quanto no tridimensional, a distribuição dos pontos de dados é um fator crítico na identificação e compreensão dos clusters. A densidade, por sua vez, embora ainda essencial, é intrinsecamente entrelaçada com a distribuição.

De modo geral, os resultados revelaram a influência preponderante da distribuição espacial dos dados sobre a densidade na resposta das meta-heurísticas utilizando critérios de distância. A distribuição espacial dos dados, que reflete como os pontos estão organizados e dispostos no espaço, emergiu como um fator decisivo na clusterização e formação de agrupamentos. Nos cenários onde os pontos de dados estão mais próximos uns dos outros ou exibem padrões

específicos de distribuição levaram a resultados distintos na clusterização.

Enquanto a densidade dos dados, que se refere à concentração de pontos em áreas específicas, continua sendo importante, os testes demonstraram que a distribuição pode sobrepujar a densidade em sua influência sobre os resultados das meta-heurísticas. Assim, a maneira como os pontos de dados estão organizados e distribuídos em áreas específicas do espaço teve um impacto mais significativo na formação de clusters.

Os resultados obtidos revelaram ainda que, independentemente da métrica de distância empregada, seja ela Euclidiana, Manhattan ou Chebyshev, todas apresentaram resultados semelhantes em relação à qualidade dos agrupamentos produzidos. Essa constatação sugere que, para estes conjuntos de dados, a escolha da métrica de distância não teve um impacto significativo nos resultados finais do agrupamento. Tal semelhança pode ser atribuída à natureza dos dados e à distribuição dos pontos no espaço, que pode ter favorecido a eficácia de todas as medidas de distância consideradas.

De toda forma, enfatiza-se a necessidade de uma análise cuidadosa da distribuição espacial dos dados em conjunto com a avaliação da densidade ao escolher ou ajustar métodos de clusterização. Também entende-se que, embora as meta-heurísticas possam ser uma ferramenta poderosa, a natureza dos dados desempenha um papel crítico em seu desempenho. Portanto, uma abordagem personalizada e a consideração das características específicas do conjunto de dados são essenciais para alcançar bons resultados de agrupamento em uma variedade de contextos.

5.3 Estudo de caso: agrupamento homogêneo de bovinos

Essa subseção contém os experimentos e resultados obtidos para o processo de agrupamento de dois conjuntos de dados de bovinos de uma empresa agropecuária, cujo objetivo é obter os bovinos com as características mais homogêneas por cluster. Nesse caso, tem-se ainda como foco a determinação da melhor meta-heurística para o agrupamento desses conjuntos, considerando a análise minuciosa dos resultados obtidos.

Foram utilizadas a mesma configuração de parâmetros das meta-heurísticas AG, PSO e ACO, as mesmas métricas de distância, sendo a Euclidiana, Manhttan e Chebyshev, e as mesmas etapas da Seção 5.1, sendo realizado a

normalização das dados, análise de componentes principais e definição da quantidade de cluster, que foi realizada por meio da análise dos índices de avaliação.

O primeiro conjunto de dados, denominado como B1, é referente ao ano de 2021 e possui 70 amostras e 13 atributos, e o segundo conjunto de dados, denominado B2, é referente ao ano de 2022 e conta com 64 amostras e 12 atributos. Ambos os conjuntos possuem os mesmos 12 atributos e o B1 possui o atributo EGP a mais. A sigla de cada atributo, seu significado e sua respectiva medida são apresentados no Quadro 10.

Quadro 10: Siglas dos atributos, significado e respectiva medida.

Sigla	**Significado**	**Medida**
C	Avalia a quantidade de carne na carcaça	Escore
P	Avalia a capacidade do bovino chegar a um grau de acabamento mínimo de carcaça, com peso vivo não elevado	Escore
M	Avalia o desenvolvimento da massa muscular como um todo, observada em pontos como antebraço, paleta, lombo, garupa e, principalmente, a largura e profundidade dos quartos traseiros	Escore
U	Avalia o tamanho e o formato do prepúcio e/ou umbigo	Escore
R	Avalia o padrão racial de acordo com as normas de sua respectiva associação promocional	Escore
PL	Avalia o comprimento e a espessura do pelo	Escore
AOL	Medida da área total do músculo longo dorsal, obtida no sítio anatômico no espaço intercostal entre a 12ª e a 13ª costela.	Centímetro quadrado
GIM	Pontos de gordura visíveis por ultrassom no músculo *Longissimus Dorsus* (contrafilé).	Porcentagem
EGS	Espessura da gordura subcutânea medida no espaço entre a 12ª e a 13ª costela do bovino.	Milímetros
EGP	Espessura da gordura subcutânea medida entre a picanha e a alcatra do bovino.	Milímetros
CE	Avalia a circunferência escrotal	Centímetro
PS	Peso vivo do bovino	Quilo
IF	Idade	Dias

Os atributos PS e CE são medidos diretamente no bovino, sendo o atributo PS dado pelo peso real do bovino e o CE pela medida da circunferência do escroto

do bovino no momento da avaliação. Os atributos AOL, MAR, EGS e EGP são obtidos por meio de ultrassonografia. Já os atributos C, PR, M, R, PL e U são atribuídos por meio de avaliação considerando as definições dadas pelo PROMEBO (Programa de Melhoramento de Bovinos de Carne), da Associação Nacional de Criadores Herd-Book Collares (2018) . O atributo IF é a quantidade de dias do bovino desde seu nascimento até o dia da avaliação.

Além disso, os atributos C, P, M e R são discretos, não categóricos, e variam entre os valores inteiros de 1 a 5, em que 1 é o pior resultado para o atributo e 5 o melhor. O atributo U também varia de 1 a 5, mas nesse caso 1 é o melhor tamanho de umbigo, enquanto 5 é o pior. Já o PL também é discreto, mas varia apenas de 1 a 3, com 1 sendo atribuído aos bovinos com melhor pelagem e 3 aos bovinos com a pior pelagem.

Embora essa variação possa ser interpretada como uma escala de notas atribuídas, é importante ressaltar que esses valores não são categóricos, mas sim discretos, representando diferentes níveis em uma determinada escala. Dessa forma, como os dados estão limitados a um conjunto específico de valores discretos, eles não podem ser considerados categóricos no sentido tradicional, já que representam medidas numéricas em uma escala definida.

As Tabelas 6 e 7 apresentam a estatística de todos os atributos, contendo a média (mean), desvio padrão (std), valor mínimo (min), valores por percentis (25%, 50% e 75%) e valor máximo (max) de cada um deles. Tomando B1 como exemplo, tem-se a média das valores atribuidos as bovinos para o atributo conformidade (C) é 3,82, com um desvio padrão de 1,20, valor mínimo em 1,0 e máximo em 5,0, com 25% dos dados abaixo de 3,0, 50% abaixo de 4,0 e 75% abaixo de 5,0.

Tabela 6: Descrição estatística do conjunto de dados B1.

	C	PR	M	U	R	PL	AOL	GIM	EGS	EGP	CE	PS	IF
Mean	3,82	3,72	4,14	2,11	2,82	1,27	94,26	2,98	3,65	4,75	33,08	469,25	475
Std	1,20	0,79	0,76	0,43	0,41	0,44	11,62	0,72	1,06	1,24	2,97	59,54	54,21
Min	1,0	2,0	3,0	1,0	1,0	1,0	69,77	1,23	1,82	2,26	28,00	360,00	362,00
25%	3,0	3,0	4,0	2,0	3,0	1,0	86,56	2,58	2,62	4,06	31,00	421,50	413,00
50%	4,0	4,0	4,0	2,0	3,0	1,0	91,69	3,10	3,87	4,74	33,00	460,00	468,50
75%	5,0	4,0	5,0	2,0	3,0	2,0	100,16	3,45	4,46	5,46	35,00	513,75	500,00
Max	5,0	5,0	5,0	3,0	3,0	2,0	126,19	4,32	5,97	7,69	43,00	595,00	550,00

Tabela 7: Descrição estatística do conjunto de dados B2.

	C	PR	M	U	R	PL	AOL	GIM	EGS	CE	PS	IF
Mean	3,73	4,03	4,03	1,78	2,61	1,61	86,47	2,70	3,48	32,28	450,33	491,20
Std	1,21	0,66	0,93	0,57	0,25	0,74	18,75	0,78	1,18	3,53	110,24	113,91
Min	1,0	2,0	2,0	1,0	1,0	1,0	56,17	0,69	1,87	26,00	285,0	331,00
25%	3,0	4,0	3,0	1,0	2,0	1,0	70,89	2,28	2,36	30,00	364,0	412,00
50%	4,0	4,0	4,0	2,0	3,0	1,0	86,42	2,86	3,59	32,50	431,0	453,00
75%	5,0	4,0	5,0	2,0	3,0	2,0	99,10	3,30	4,24	34,50	549,0	599,00
Max	5,0	5,0	5,0	3,0	3,0	3,0	130,84	4,20	6,92	43,00	712,0	735,00

Dada as diferentes escalas dos atributos, foi realizada a padronização dos dados numéricos dos dois conjuntos utilizando a Equação (1), de forma a colocá-los em uma mesma escala. Essa abordagem uniformiza a escala dos atributos, proporcionando uma análise mais consistente e precisa.

Feita a padronização dos dados, aplicou-se a técnica PCA, conforme apresentado na Seção 2.1.5. A Figura 29 demonstra a curva de variância explicada em relação ao número de componentes. Diante da complexidade inerente à manipulação de dados com um grande número de dimensões, reduziu-se a dimensionalidade dos dois conjuntos para 4 dimensões. Optou-se por reter uma quantidade de componentes que, juntos, abrangem mais de 90% da variância explicada total, visando equilibrar a redução de dimensionalidade com a retenção de uma percentagem substancial da informação original contida nos dados.

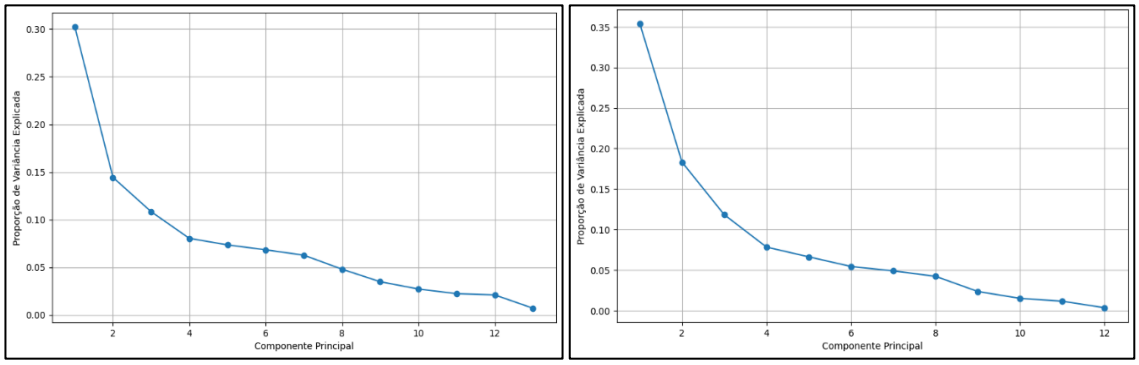


Figura 29: Gráfico de variância explicada por componentes principais para os conjuntos de dados B1 e B2.
Fonte: Autoria Própria (2024).

Para os Conjuntos B1 e B2 a definição do melhor número de clusters foi dada por meio da utilização dos índices de avaliação SH, DB e CH. Como se sabe, esses índices fornecem medidas quantitativas da coesão intra-cluster e da separação inter-cluster, permitindo uma avaliação objetiva da qualidade dos agrupamentos para diferentes números de clusters. Nesse caso, foi testada uma variação de 2 a 6 no número de clusters, a fim de identificar com qual número se obtém o melhor agrupamento.

A variação de valores permite não apenas para definir qual o melhor número de clusters, mas também para determinar qual meta-heurística e com qual métrica de distância se obtém os melhores resultados para os índices analisados. O agrupamento final para os Conjuntos B1 e B2 foi feito apenas pela meta-heurística com a métrica de distância que obteve os melhores resultados considerando os três índices de avaliação.

As três meta-heurísticas com as diferentes métricas foram processadas 10 vezes individualmente e as Figuras de 30 e 33 apresentam a variação dos índices de avaliação em relação à quantidade cluster para cada meta-heurística com para B1 e da Figura 33 a 35 para B2. Os resultados para cada processamento são apresentados no Anexo A e B. Os resultados para SH e DB são representados pelas cores azul e laranja, respectivamente, enquanto o de CH é o de cor verde.

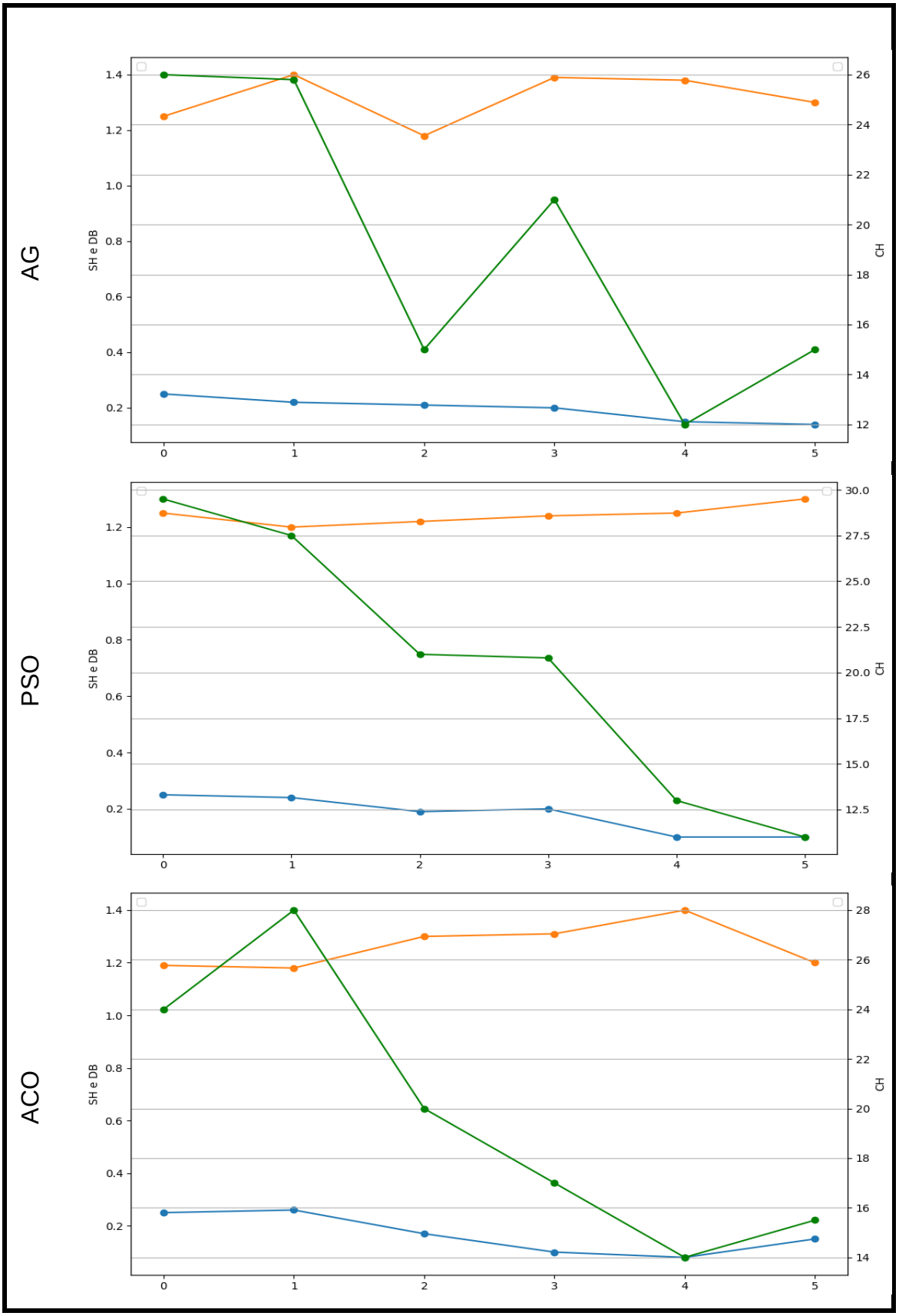


Figura 30: Gráfico da variação dos índices de avaliação em relação à quantidade cluster para cada meta-heurística com a Distância Euclidiana para B1.

Fonte: Autoria Própria (2024).

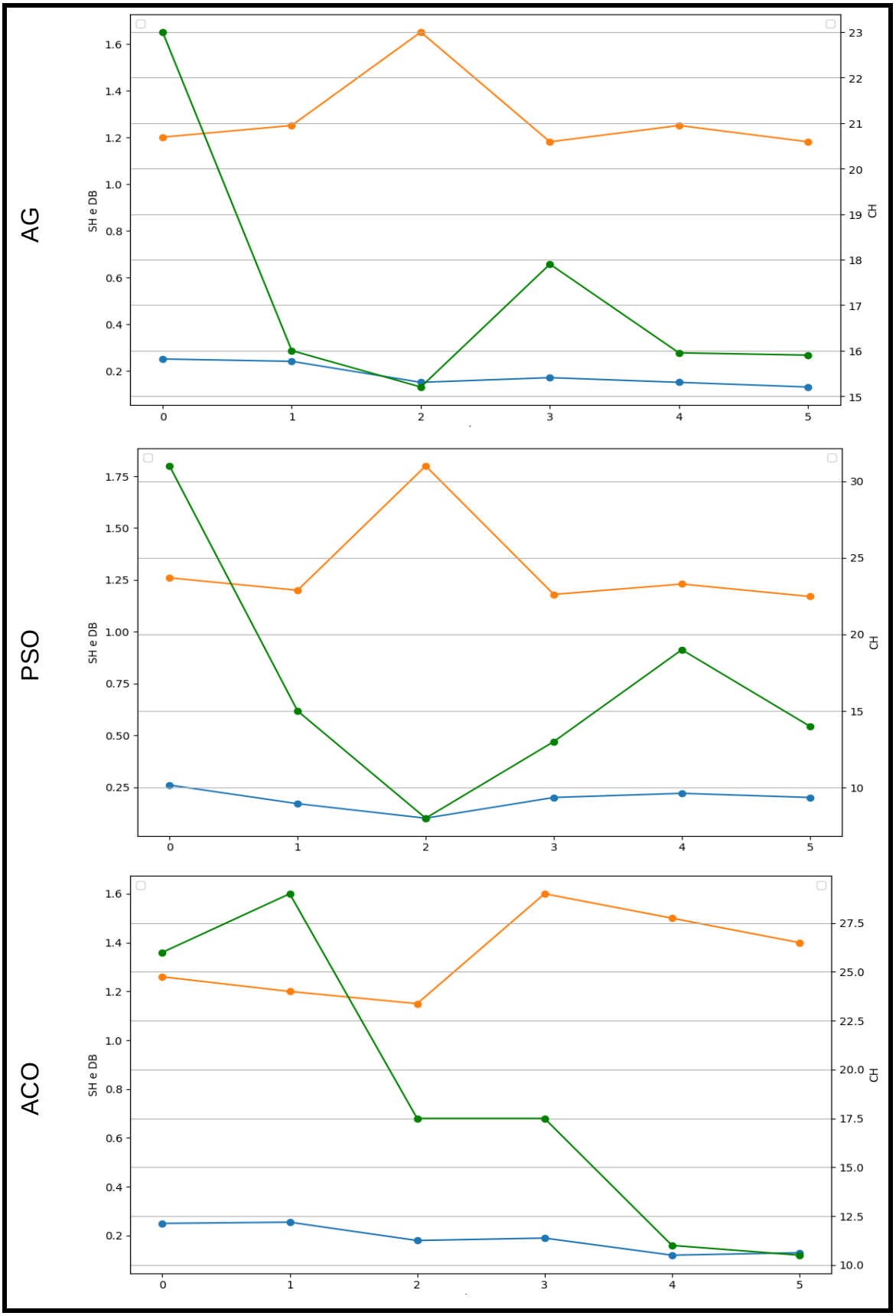


Figura 31: Gráfico da variação dos índices de avaliação em relação à quantidade cluster para cada meta-heurística com a Distância Manhathan para B1.

Fonte: Autoria Própria (2024).

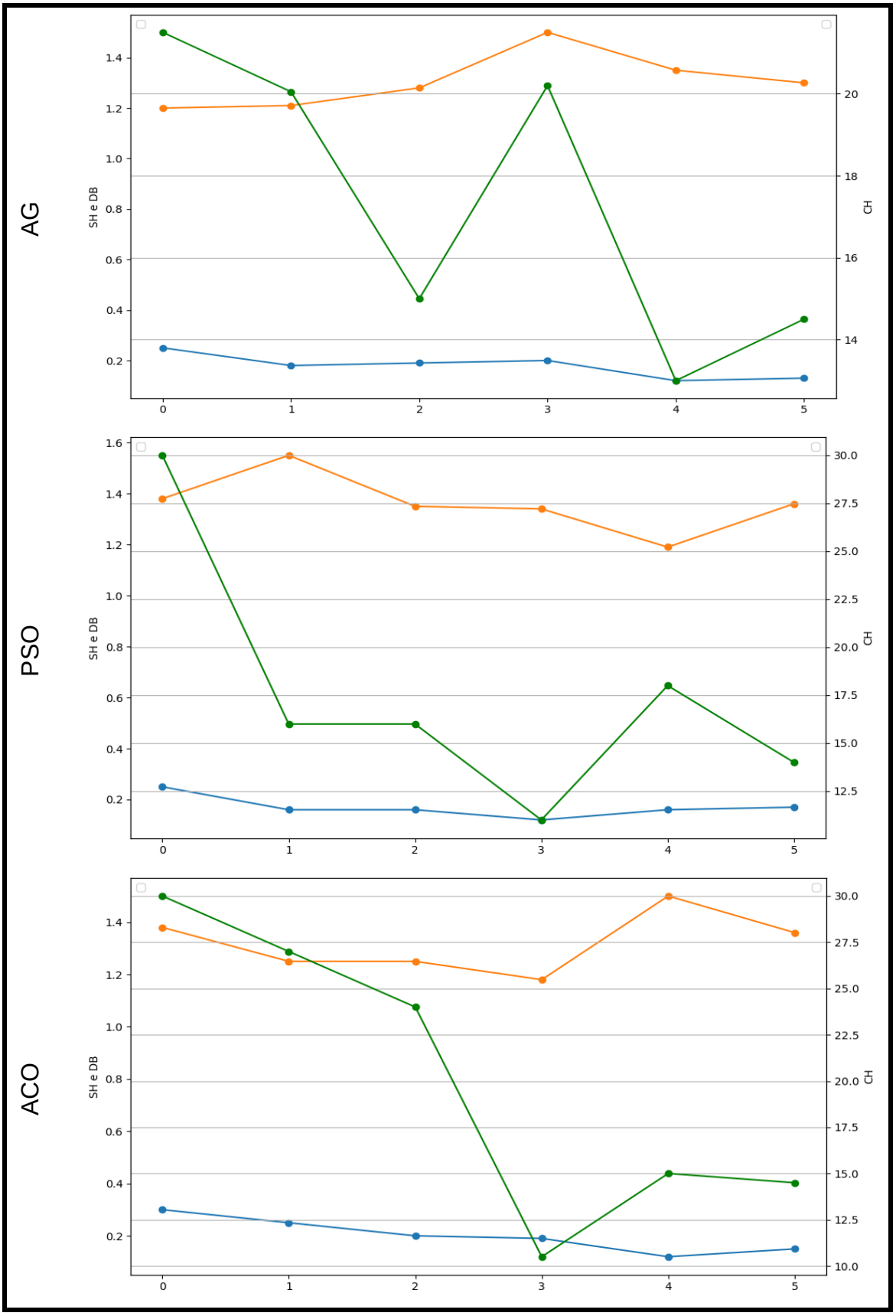


Figura 32: Gráfico da variação dos índices de avaliação em relação à quantidade cluster para cada meta-heurística com a Distância Chebyshev para B1.

Fonte: Autoria Própria (2024).

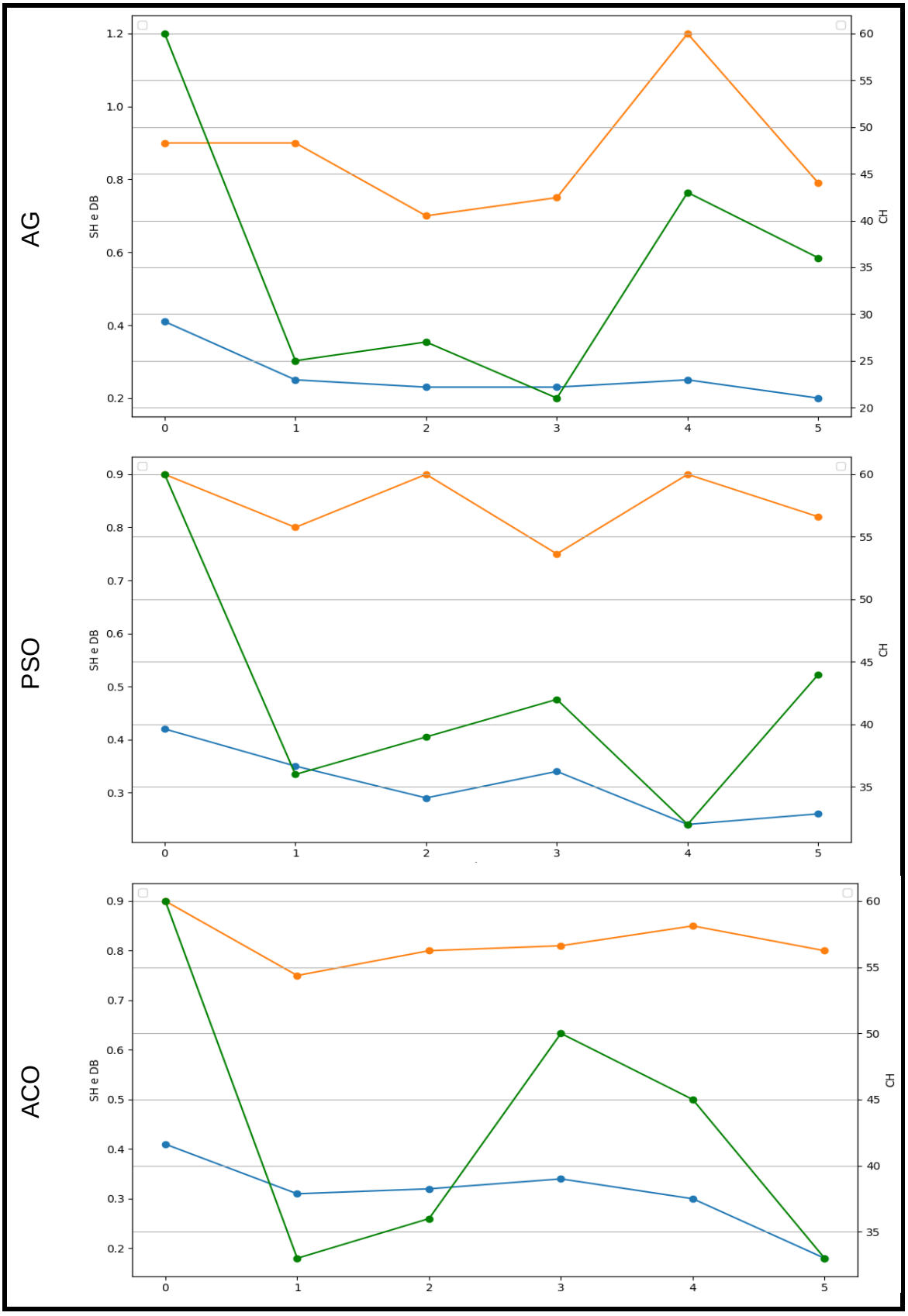


Figura 33: Gráfico da variação dos índices de avaliação em relação à quantidade cluster para cada meta-heurística com a Distância Euclidiana para B2.

Fonte: Autoria Própria (2024).

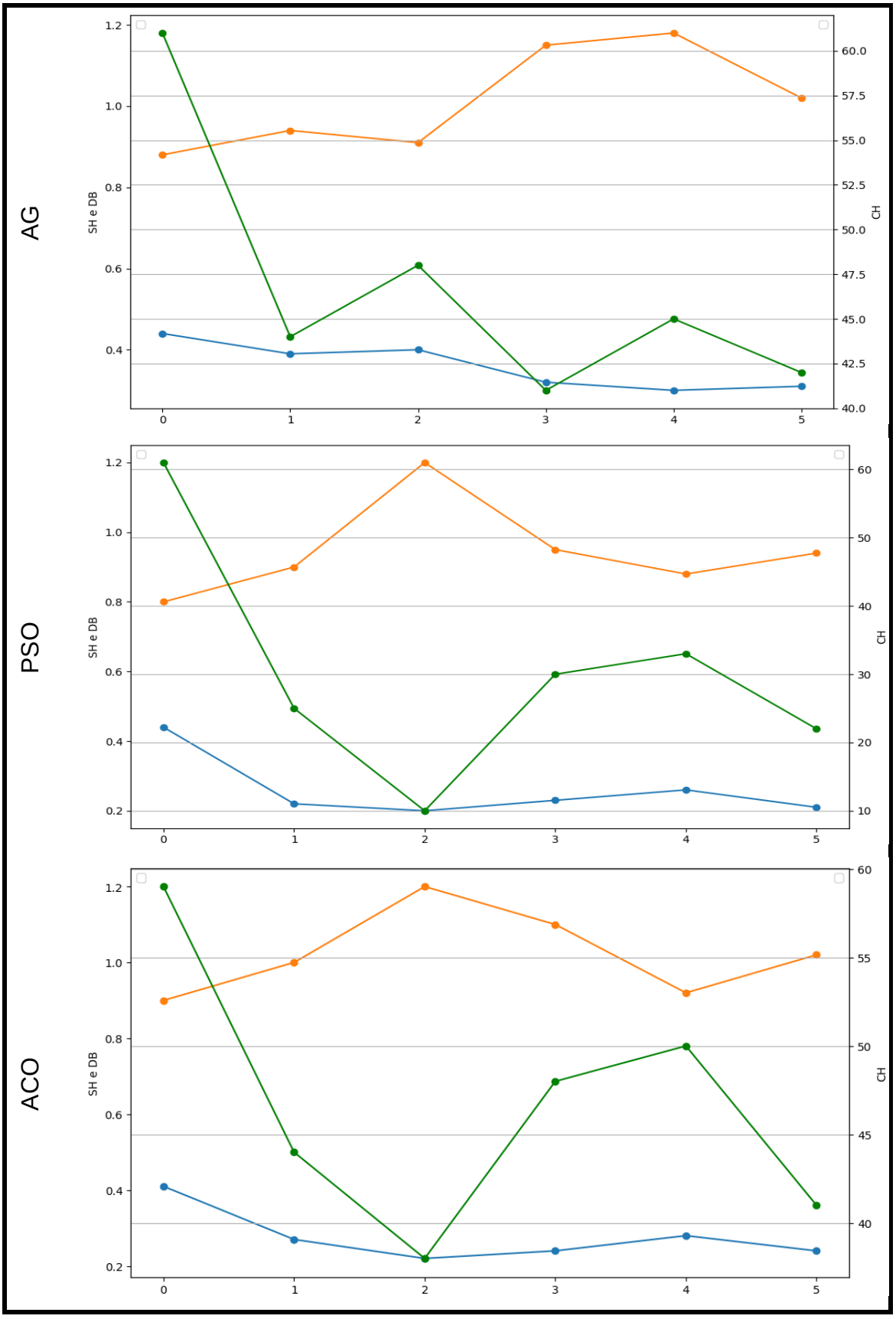


Figura 34: Gráfico da variação dos índices de avaliação em relação à quantidade cluster para cada meta-heurística com a Distância Manhathan para B2.

Fonte: Autoria Própria (2024).

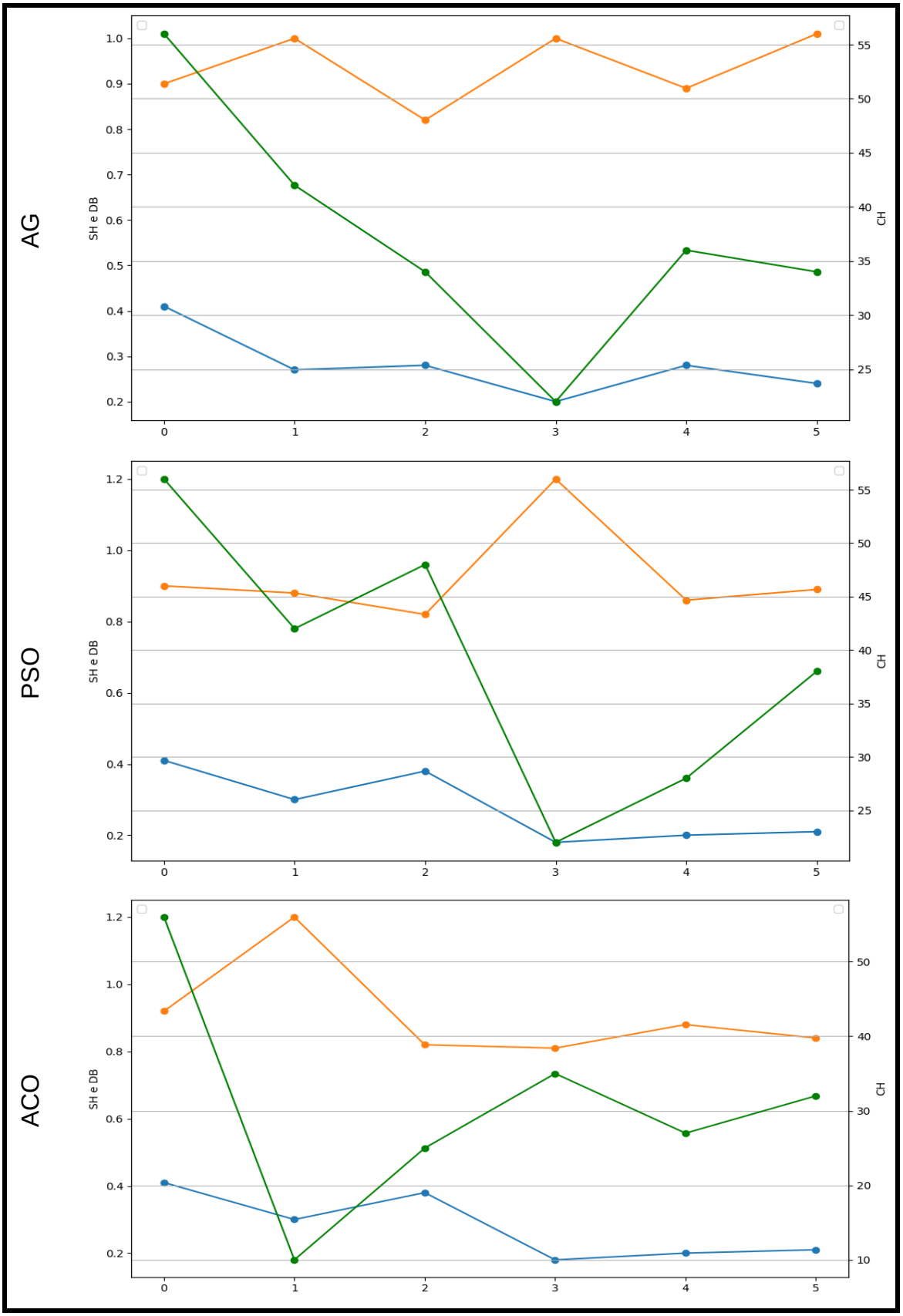


Figura 35: Gráfico da variação dos índices de avaliação em relação à quantidade cluster para cada meta-heurística com a Distância Chebyshev para B2.

Fonte: Autoria Própria (2024).

Considerando que o melhores resultados são os que apresentam o valor mais próximo de 1 para SH, mais perto de 0 para DB e o maior número para CH, conforme mostra a Figura 32, para o Conjunto B1 o número de clusters *k*=2 foi o que obteve os melhores resultados e a meta-heurística PSO com distância Manhattan é a que apresentou os melhores resultados entre todas. É possível observar ainda que *k*=2 está entre os melhores resultados para todas as meta-heurísticas, independente da métrica de distância, o que reitera que este seja o melhor número.

Para o Conjunto B2, o número de clusters *k*=2 também foi o que obteve os melhores resultados e a meta-heurística com distância Manhathan é a que apresenta os melhores resultados entre todas. Novamente, é possível observar que *k*=2 está entre os melhores resultados para todas as meta-heurísticas, tanto para a distância Euclidiana, Manhattan ou Chebyshev.

Como o desempenho das meta-heurísticas pode variar entre diferentes execuções devido a fatores estocásticos, foram feitas mais 10 processamentos independentes da PSO com distância Manhattan, que foi que obteve os melhores resultados na etapa anterior, para os dois conjuntos. Essa abordagem visa obter uma avaliação mais confiável do desempenho da meta-heurística para ambos os conjuntos.

As Tabelas 8 e 9 apresentam o melhor valor médio de cada atributo para cada cluster, denominados como C1 e C2, resultantes da aplicação da meta-heurística PSO com distância Manhattan que obtiveram os melhores resultados nos índices de avaliação. Esses valores médios representaram as características específicas dos clusters, fornecendo informações sobre como os dados foram agrupados e como os clusters se diferenciam em relação aos atributos considerados.

Tabela 8: Valores médios de cada atributo por cluster obtidos pela PSO – Manhattan para o Conjunto B1.

	C	PR	M	U	R	PL	AOL	GIM	EGS	EGP	CE	PS	IF
C1	4,8	4,1	4,56	2,2	2,93	1,2	102,18	3,15	3,94	5,30	35,13	520,3	488,06
C2	3,1	3,45	3,82	2,05	2,75	1,32	88,32	2,85	3,44	4,33	31,55	430,97	434,27

Tabela 9: Valores médios de cada atributo por cluster obtidos pela PSO – Manhattan para o Conjunto B2.

	C	PR	M	U	R	PL	AOL	GIM	EGS	CE	PS	IF
C1	4,35	4,16	4,48	1,75	2,75	1,75	98,72	2,75	4,03	34,13	519,86	552,75
C2	2,92	3,85	3,42	1,82	2,42	1,42	70,28	2,63	2,76	29,83	358,46	409,85

Ao analisar o atributo C, em que o escore 5 refere-se ao melhor resultado possível, tem-se que para o conjunto B1 o cluster C1 obteve valor 54,83% superior (4,8) em relação ao cluster C2 (3,1). O mesmo ocorreu para o conjunto B2, em que C1 ficou com a média de 4,35 e C2 com 2,92, o que representa uma diferença de 48,97% de C1 em relação a C2. Esses valores evidenciam que o C1 dos dois conjuntos é formado, em sua maioria, por bovinos que estão acima da média apresentada nas Tabelas 6 e 7, enquanto o C2 ficou com os bovinos que tiveram menor escore para este atributo.

Para o atributo PR, referente à precocidade do bovino, observa-se que o Conjunto B1 obteve os valores 4,1 para o C1 e 3,45 para C2. Novamente, existe uma diferença de 18,84% entre os dois grupos, com o C1 sendo formado por bovinos considerados mais precoces e C2 formado por bovinos com uma precocidade média. Esse cenário é semelhante para o Conjunto B2, que apresenta uma diferença de 8,05% entre os clusters, com o C1 (4,16) sendo formado por bovinos mais precoces do que o cluster C2 (3,85). Novamente, nota-se que C1 ficou acima da média e C2 abaixo.

Os valores do atributo M também evidenciam uma diferença entre os clusters C1 e C2. Para o Conjunto B1, houve uma diferença de 19,37% entre os clusters, com o valor de 4,56 para o C1 demonstrando que a musculatura dos bovinos deste grupo foi considerada superior à média do conjunto todo, enquanto o cluster C2 (3,82) ficou abaixo da média. O mesmo ocorreu para o conjunto B2, que apresenta 4,46 para C1 e 3,42 para C2, estando C1 acima e C2 abaixo da média apresentada na Tabela 7.

Analisando os valores obtidos para o atributo U, referente à altura do umbigo, percebe-se que para os dois clusters os valores foram próximos tanto para o Conjunto B1, com 7,31% de diferença de C1 (2,2) e C2 (2,05), quanto para B2, com 4,0% entre C1 (1,75) e C2 (1,82). Nesse caso, não há uma distinção clara entre os dois grupos para ambos os conjuntos, mas ao considerar os valores para

atributo apresentados na Tabela 6 e 7, entende-se que os valores não poderiam ser distantes entre os clusters, dado que o escore atribuido a maioria dos bovinos dos dois conjuntos havia sido 2.

No que tange às características raciais, representada pelo atributo R, para o Conjunto B1, o C1 (2,93) ficou 6,54% acima de C2 (2,75), e para o Conjunto B2, o C1 obteve 2,75 e C2 2,42, que representa uma diferença de 13,63%. Para os dois conjuntos, considerando as Tabelas 6 e 7, C1 ficou acima da média e C2 abaixo. Considerando os percentis para os dois conjuntos, entende-se que a maioria dos bovinos dos clusters C1 foram avaliados como 3, enquanto os de C2 apresentou maior quantidade de bovinos com escore menor.

O atributo PL também apresentou valores próximos entre os dois clusters para os dois conjuntos de dados. O valor de 1,2 para C1 e 1,32 para C2 para o Conjunto B1 demonstra que a maioria dos bovinos em ambos os clusters tem escore 1. Para o Conjunto B2, com C1 apresentando 1,42 e C2 1,75, evidencia-se a existência de quantidade maior de bovinos com notas acima de 1 em ambos os clusters. Importante reiterar que a Tabela 6 demonstra que para a maioria dos bovinos dos do Conjunto B1 foi atribuída 1 para este atributo, o que justifica o fato dos dois clusters terem valores tão próximos, enquanto para o Conjunto B2 tem-se maior variedade de escores.

O atributo AOL não possui um intervalo determinado, sendo atribuído por ultrassonografia e medido em cm^2. Para esse atributo, o Conjunto B1 obteve valores de 102,18 para C1 e 88,32 para C2, enquanto o Conjunto B2 obteve 98,72 para C1 e 70,28 para C2. Percebe-se que o C1 ficou acima da média dos seus respectivos conjuntos, com valores mais próximos do valor máximo (126,19) do que do mínimo (69,77). Já o C2 dos dois conjuntos ficou abaixo da média, com valores mais próximos do mínimo do seu respectivo conjunto.

Os valores para o atributo GIM também são atribuídos por ultrassonografia, sendo medido em porcentagem. Os resultados para o Conjunto B1 demonstram que os bovinos do cluster C1 possuem um valor 10,52% mais elevado que o C2 para este atributo, com 3,15 e 2,85, respectivamente. O mesmo ocorre para o Conjunto B2, em que C1 (2,75) apresenta valor 4,65% maior que C2 (2,63). Considerando a descrição completa de cada conjunto, entende-se que o C1 está sendo formado pelos bovinos com as maiores medidas, enquanto o C2 com os bovinos com as menores.

Para o atributo EGP, o Conjunto B1 obteve 5,30 para o C1 e 4,33 para C2. Considerando todos os valores desse atributo presentes no conjunto todo, em que a média é 4,75 e 25% das amostras possuem nota superior a 5,46 e 50% superior a 4,74, entende-se que o C1 é formado, em sua maioria, por bovinos que tiveram maiores medidas no conjunto todo, enquanto C2 foi formado pelos bovinos com que obtiveram medidas inferiores à 4,74.

Com relação ao atributo CE, tem-se que o C1 (35,13) tem bovinos com maior circunferência escrotal do que os bovinos do grupo C2 (31,55) para o Conjunto B1. O mesmo ocorre para o Conjunto B2, em que C1 (34,13) possui maior valor em relação à C2 (29,83). De acordo com a descrição dos dados nas Tabelas 6 e 7, é possível inferir que os bovinos alocados no C1 possuem circunferências escrotais superiores aos 50% da distribuição para os dois conjuntos, enquanto os do C2 apresentam medidas inferiores a esse percentil.

Ao analisar o atributo PF para o Conjunto B1, observa-se que o C1 (520,3) tem bovinos 20,72% mais pesados do que os de C2 (430,97). Igualmente, o C1 (519,86) do Conjunto B2 obteve valor 45,02% maior do que o C2 (358,46). Novamente, conforme descrito nas Tabelas 6 e 7, os bovinos atribuídos ao C1 exibem pesos superiores à mediana da distribuição, enquanto aqueles do C2 apresentam pesos abaixo desse valor.

O último atributo considerado foi IF, que refere-se a idade dos bovinos. O C1 e C2 do Conjunto B1 ficaram com 488,06 e 434,27, respectivamente, e para o Conjunto B2 o Cluster C1 ficou com 552,75 e o C2 com 409,85. Entende-se que os bovinos alocados no C1, em sua maioria, possuem uma idade mais elevada que a média de ambos os conjuntos, conforme descrito nas Tabelas 6 e 7, enquanto o Conjunto C2 é formado pelos bovinos que estão, em sua maioria, abaixo da média.

De modo geral, a discrepância observada entre os clusters nos Conjuntos B1 e B2, em relação a todos os atributos, evidencia uma segmentação clara entre os grupos, indicando que a meta-heurística aplicada foi eficaz na separação dos bovinos com base em suas características. Essa disparidade sugere que os clusters foram capazes de identificar e capturar as variações significativas nas características dos bovinos, resultando em perfis distintos para cada grupo.

Além disso, a diferenciação entre os clusters é fundamental para entender as variações existentes dentro da população de bovinos e orientar estratégias de

seleção e manejo para otimizar a produção e a qualidade da carne. Ao identificar grupos distintos de bovinos com características específicas, os produtores podem direcionar suas ações de manejo, nutrição e reprodução de forma mais precisa, visando aprimorar atributos desejáveis, como conformação, peso, deposição de gordura e fertilidade.

É importante reiterar que, por mais capacitado que seja o profissional que atribui os valores aos atributos C, PR, M, R, PL e U, essa atribuição ainda possui um viés subjetivo, sujeito a inconsistências e discordâncias a partir de outro ponto de vista. Assim, se torna válido considerar que, ainda que seja padronizado, podem existir variações naturais nos dados, já que diferentes profissionais podem atribuir valores distintos ao mesmo bovino. Portanto, é essencial interpretar os resultados com cautela e considerar outros aspectos ao realizar análises e tomar decisões com base nos dados fornecidos.

No caso dos atributos atribuídos por ultrassonografia, estes estão isentos do viés de avaliação pessoal do profissional avaliador. O mesmo vale para para a circunferência escrotal, peso e idade, que são mensurados através de procedimentos objetivos, reduzindo assim a subjetividade na atribuição dos valores. Cabe ainda destacar que, considerando os resultados de cada atributo individualmente, não necessariamente valores mais altos são melhores, já que cada atributo possui a interpretação distintas para suas medidas e escores.

As Tabelas 10 e 11 apresentam um compilado dos valores dos atributos e seus respectivos significados, que estão na coluna “Avaliação”, segundo o método de avaliação do PROMEBO. Para os valores atribuídos como escore, considera-se o número inteiro mais próximo do valor apresentado nas Tabelas 8 e 9, entendendo que a maioria dos bovinos daquele grupo possuíam esse valor atribuído. No caso dos atributos medidos por ultrassonografia, fita métrica, balança e em dias, a avaliação de maior e menor ou melhor e pior é feita comparando um cluster com o outro, tendo em vista que não são estabelecidos intervalos fechados para realizar essa avaliação.

Tabela 10: Valor atribuído e avaliação de cada cluster do Conjunto B1.

Atributo	Cluster	Valor	Avaliação
C	C1	5	Ótimo/animal superior
	C2	3	Um pouco acima da média
PR	C1	4	Um pouco acima da média
	C2	3	Animal Médio
M	C1	5	Ótimo/animal superior
	C2	4	Um pouco acima da média
U	C1	2	Pequeno
	C2	2	Pequeno
R	C1	3	Dentro do padrão racial
	C2	3	Dentro do padrão racial
PL	C1	1	Bovinos com pelo curto e liso
	C2	1	Bovinos com pelo Curto e liso
AOL	C1	102,18	Maior rendimento de carcaça
	C2	88,32	Menor rendimento de carcaça
GIM	C1	3,15	Maior porcentagem de gordura intramuscular
	C2	2,85	Menor porcentagem de gordura intramuscular
EGS	C1	3,94	Maior espessura de gordura subcutânea
	C2	3,44	Menor espessura de gordura subcutânea
EGP	C1	5,30	Maior espessura de gordura na área da picanha
	C2	4,33	Menor espessura de gordura na área da picanha
CE	C1	35,13	Maior circunferência escrotal
	C2	31,55	Menor circunferência escrotal
PS	C1	520,3	Maior peso corporal
	C2	430,97	Menor peso corporal
IF	C1	488,06	Maior idade
	C2	434,27	Menor idade

Para o Conjunto B1, o Cluster 1 (C1) demonstra características excepcionais ou superiores em termos de conformação (C), precocidade (PR) e musculosidade (M), bem como para características ideais para o tamanho do umbigo (U), características raciais (R) e pelagem (PL) . Além disso, apresenta valores maiores em atributos como área de olho da picanha (AOL), gordura intramuscular (GIM), espessura de gordura subcutânea (EGS) e na área da picanha (EGP), circunferência escrotal (CE), peso corporal (PS) e idade (IF), indicando uma qualidade superior e um perfil mais robusto em relação aos bovinos do Cluster 2 (C2).

O contraste entre os dois clusters evidencia uma diferenciação significativa em termos de características morfológicas e de desempenho. O Cluster 1 (C1) destaca-se como um grupo com bovinos de alta qualidade e

potencial produtivo, representando uma categoria superior em todos os aspectos analisados. Suas características excepcionais, como maior tamanho, musculatura mais desenvolvida, e espessura de gordura mais generosa, sugerem bovinos mais robustos e com potencial para um melhor rendimento em termos de produção de carne.

O Cluster 2 (C2) exibe características que, embora não sejam tão boas quanto às de C1, não podem ser consideradas como ruins, dado que estão na média ou padrão determinado para cada taributo, indicando um porte físico bom, porém não excepcional. Esse grupo pode ser considerado como uma categoria intermediária, com bovinos que possuem qualidades sólidas, mas que podem não atingir o mesmo nível de desempenho ou produção que os do Cluster 1 (C1). Isso sugere que, embora os bovinos do C2 tenham potencial produtivo, eles podem exigir estratégias de manejo específicas para otimizar seu desempenho e alcançar seu pleno potencial genético.

Tabela 11: Valor atribuído e avaliação de cada cluster do Conjunto B2.

Atributo	Cluster	Valor	Avaliação
C	C1	4	Um pouco acima da média
	C2	3	Animal Médio
PR	C1	4	Um pouco acima da média
	C2	4	Um pouco acima da média
M	C1	4	Um pouco acima da média
	C2	3	Animal Médio
U	C1	2	Pequeno
	C2	2	Pequeno
R	C1	3	Dentro do padrão racial
	C2	2	Fora do padrão racial
PL	C1	2	Condição intermediária
	C2	1	Bovinos com pelo Curto e liso
AOL	C1	98,72	Maior rendimento de carcaça
	C2	70,28	Menor rendimento de carcaça
GIM	C1	2,75	Maior porcentagem de gordura intramuscular
	C2	2,63	Menor porcentagem de gordura intramuscular
EGS	C1	4,03	Maior espessura de gordura subcutânea
	C2	2,76	Menor espessura de gordura subcutânea
CE	C1	34,13	Maior circunferência escrotal
	C2	29,83	Menor circunferência escrotal
PS	C1	519,86	Maior peso corporal
	C2	358,46	Menor peso corporal
IF	C1	552,75	Maior idade
	C2	409,85	Menor idade

No caso do Conjunto B2, o C1 apresenta características que indicam um desempenho superior ou ligeiramente acima da média em comparação com o C2. Em relação à conformação (C), precocidade (PR) e musculosidade (M), os bovinos de C1 obtiveram uma avaliação “um pouco acima da média”, enquanto os de C2 exibem algumas características que sugerem um nível médio. O atributo U indica que a maioria dos bovinos de ambos os clusters possuem um tamanho de umbigo pequeno.

É interessante observar que na avaliação das características raciais (R), a maioria dos bovinos do Cluster C1 está dentro do padrão racial esperado, enquanto a maioria dos bovinos de C2 estão classificados como "fora do padrão racial", indicando uma heterogeneidade genética ou variação fenotípica dentro deste grupo. Já em relação à pelagem, o C1 apresenta uma condição intermediária, indicando uma variedade de tipos de pelagem, enquanto o C2 é caracterizado principalmente por bovinos com pelo curto e liso, que é o ideal. Essas diferenças sugerem variações distintas nas características morfológicas dos bovinos em cada grupo.

Quanto aos atributos relacionados à área de olho da picanha (AOL), gordura intramuscular (GIM), espessura de gordura subcutânea (EGS) e na área da picanha (EGP), circunferência escrotal (CE), peso corporal (PS), os resultados da avaliação presente na Tabela 11 indicam qualidade superior de C1 em relação aos bovinos de C2. Isso sugere uma maior capacidade de deposição de gordura nos bovinos de C1, o que pode ser relevante em sistemas de produção que valorizam o rendimento de carcaça e a qualidade da carne.

No que diz respeito à circunferência escrotal (CE), que é uma medida relacionada à fertilidade e precocidade sexual dos machos, os bovinos de C1 apresentam valores mais elevados em comparação com os de C2, o que pode indicar uma maior potencial reprodutivo dos bovinos desse grupo. Em relação a idade, tem-se que os bovinos de C1 são predominantemente mais velhos do que os do cluster C2.

Considerando todos os atributos analisados, parece haver uma relação entre os atributos dos conjuntos de dados, dado que para ambos os conjuntos os bovinos do Cluster C1 apresentam as melhores avaliações. Essa relação sugere que os bovinos do Cluster C1 podem ter um desenvolvimento físico mais avançado e, possivelmente, um maior potencial reprodutivo em comparação com os do

Cluster C2.

Essas relações entre os atributos destacam a importância de considerar múltiplas características ao avaliar os bovinos para seleção e melhoramento genético, ou mesmo para direcioná-los para abate. Segundo Hlavatý et al. (2023), uma abordagem holística que leve em conta diversas características pode levar a decisões mais informadas e eficazes na gestão do rebanho, contribuindo significativamente para o desenvolvimento de rebanhos mais produtivos e adaptados às necessidades do mercado.

Ao considerar todas as características dos bovinos de C1 dos dois conjuntos, tem-se que esses bovinos poderiam ser direcionados tanto para programas de reprodução seletiva, visando aprimorar as características desejáveis identificadas, como para sistemas de produção que valorizam o rendimento de carne e a qualidade do produto final. A alta pontuação em uma variedade de atributos sugere que esses bovinos possuem um potencial significativo para contribuir tanto para o melhoramento genético do rebanho quanto para a maximização dos lucros na cadeia de produção de carne.

Por sua vez, o C2 consiste em bovinos que, embora não apresentem os mesmos níveis de excelência observados no C1, ainda possuem características dentro de padrões aceitáveis. Isso sugere que esses bovinos podem ser direcionados para sistemas de produção que demandem bovinos com desempenho satisfatório, mas não necessariamente excepcional. Além disso, os bovinos deste grupo podem requerer estratégias nutricionais específicas para maximizar seu potencial de ganho de peso e rendimento de carcaça.

De toda forma, independentemente da destinação final do bovino, o processo de particionamento permite uma abordagem mais personalizada em relação à nutrição e suplementação alimentar, dado que bovinos agrupados em diferentes clusters podem apresentar necessidades nutricionais distintas. Ridge, Foster e Daigle (2020) mencionam que ao considerar suas diferenças, os produtores podem ajustar de forma mais precisa a dieta e a suplementação de cada grupo de bovinos, maximizando seu potencial de crescimento, saúde e desempenho dos bovinos.

Ao identificar os clusters com características superiores, os produtores podem selecionar os bovinos mais promissores para reprodução, visando a transmissão de genes favoráveis e o aprimoramento das características desejadas

na progênie. Isso contribui para o avanço do melhoramento genético do rebanho, promovendo uma produção mais eficiente e sustentável ao longo do tempo, sendo uma prática fundamentada em princípios de seleção artificial e que tem sido amplamente adotada em programas de melhoramento genético animal em todo o mundo (Mueller & Van Eenennaam, 2022).

Para ambos os conjuntos, por exemplo, o cluster C1, ao qual os bovinos com maior circunferência escrotal foram atribuídos, também conta com os bovinos com as melhores avaliações para os demais atributos. Isso faz sentido ao considerar que a CE, sendo uma medida relacionada à fertilidade e ao desenvolvimento sexual dos machos, está correlacionada com outras características importantes para a produção de bovinos, como o peso corporal, a musculosidade e até mesmo a qualidade da carne (Pereira et al., 2023).

Por fim, entende-se que o destino final desses bovinos está diretamente ligado aos objetivos e às estratégias de criação estabelecidas pelo criador. No entanto, o particionamento dos bovinos em clusters distintos com base em características específicas oferece uma base sólida para melhor tomada de decisões. A partir dessas informações, os produtores podem direcionar os bovinos para diferentes finalidades, seja para reprodução, produção de carne de alta qualidade, melhoramento genético contínuo ou outros fins específicos. Esse processo de segmentação orientada permite uma gestão mais eficaz do rebanho, maximizando os benefícios econômicos e produtivos para o criador.

6 CONCLUSÕES

A proposta deste trabalho foi investigar e comparar a eficácia das meta-heurísticas Algoritmo Genético, Colônia de Formigas e Otimização por Enxame de Partículas em problemas de agrupamento, utilizando três métricas de distância distintas: Euclidiana, Manhattan e Chebyshev. Um dos objetivos específicos foi realizar o agrupamento de bovinos, com a finalidade de identificar a meta-heurística que apresenta os melhores resultados para esses conjuntos de dados. Por meio da análise dos resultados, buscamos analisar se as meta-heurísticas e as métricas de distância selecionadas são capazes de proporcionar uma representação precisa das relações intrínsecas entre os dados.

As meta-heurísticas AG, PSO e ACO demonstraram um bom desempenho ao agrupar conjuntos de dados *benchmarks*. A validação desses agrupamentos foi obtida tanto por meio das métricas de avaliação quantitativas quanto pela avaliação visual, reforçando a eficácia das meta-heurísticas. Importante ressaltar que essa etapa dos testes computacionais possibilitou o entendimento inicial de como as configurações dos parâmetros das meta-heurísticas influenciam no seu desempenho, permitindo ajustar as estratégias de otimização de acordo com a complexidade e a natureza dos dados.

Com relação aos conjuntos analisados na Seção 5.2, observou-se, novamente, um comportamento semelhante entre as meta-heurísticas. A análise dos diversos cenários ressalta a complexidade da relação entre as estratégias adotadas pelas meta-heurísticas e as características intrínsecas dos dados, com a distribuição dos dados emergindo como um fator mais influente do que a densidade no desempenho das meta-heurísticas. A variabilidade nos resultados sugere que as meta-heurísticas podem reagir de forma distinta a depender da natureza dos dados, destacando a importância de um estudo mais profundo sobre a aplicabilidade delas em diferentes contextos de dados.

Esses resultados proporcionaram uma visão abrangente do desempenho das meta-heurísticas em uma variedade de cenários. De modo geral, observa-se que as três meta-heurísticas demonstram potencial para serem aplicadas com sucesso na tarefa de agrupamento de dados, independentemente da métrica de distância utilizada. A análise abrangente realizada destaca a robustez e a adaptabilidade dessas abordagens em diferentes contextos, fornecendo *insights* valiosos para a seleção e implementação de técnicas de agrupamento em diversas

aplicações práticas.

No que se refere ao agrupamento dos bovinos, a segmentação dos bovinos em clusters distintos permitiu uma compreensão mais profunda das variações existentes dentro dos dois conjuntos. Essa identificação de grupos com características homogêneas pode ser importante para estratégias de manejo mais eficientes e para a otimização da produção e qualidade da carne, já que bovinos mais gordos e robustos podem ser direcionados para mercados que valorizam animais com maior peso, enquanto os bovinos mais magros podem receber uma dieta e cuidados que visem melhorar seu ganho de peso.

As características dos dois grupos identificados podem ser utilizadas não apenas para a seleção de reprodutores, mas também para a gestão do rebanho em termos de nutrição, saúde e reprodução. Esse particionamento dos bovinos em clusters também abre caminho para a implementação de estratégias de manejo mais direcionadas e personalizadas, considerando as necessidades específicas de cada grupo.

A comparação entre os clusters destacou áreas de melhoria e oportunidades de otimização no sistema de produção. Por exemplo, o Cluster 1 dos dois conjuntos apresentou consistentemente melhor desempenho em termos de rendimento de carcaça e qualidade da carne, indicando que práticas ou características genéticas que podem ser aplicadas de forma mais ampla para melhorar o desempenho geral do rebanho.

Já o Cluster 2 dos dois conjuntos não demonstrou o mesmo nível de desempenho em comparação com o Cluster 1, sugerindo a necessidade de uma avaliação mais aprofundada das práticas de manejo, genéticas ou ambientais que possam estar afetando negativamente esse grupo. Nesse caso, é importante identificar as causas subjacentes dessa disparidade entre os clusters para a implementação de estratégias de melhoria específicas, visando aumentar a produtividade e a eficiência deste grupo como um todo.

Face ao exposto, constatou-se que as três meta-heurísticas, utilizando as diferentes métricas de distância, foram capazes de identificar as relações entre os dados, validando a escolha das técnicas empregadas. Os achados confirmam a capacidade das estratégias de otimização adotadas de fornecer soluções adequadas para problemas de agrupamento, evidenciando seu potencial para aplicações práticas, em especial para a segmentação dos bovinos.

6.1 Pesquisas Futuras

Para estudos futuros, sugere-se realizar análises mais detalhadas sobre a composição genética dos clusters e sua associação com características específicas de desempenho, como taxa de crescimento, eficiência alimentar e resistência a doenças, o que poderia ajudar a identificar marcadores genéticos ou genes candidatos associados a características de interesse e facilitar a seleção de bovinos com maior potencial produtivo e adaptabilidade.

Outra área promissora para pesquisa futura seria investigar o impacto das práticas de manejo, nutrição e ambiente na expressão dos genes e no desempenho dos bovinos dos diferentes clusters. Compreender como fatores externos podem modular a expressão genética e influenciar características de interesse pode fornecer *insights* para o desenvolvimento de estratégias de manejo mais eficazes e sustentáveis.

Por fim, estudos longitudinais que acompanhem o desempenho dos bovinos ao longo do tempo e avaliem a estabilidade dos clusters ao longo das diferentes fases de produção seriam importantes para entender melhor as mudanças na estrutura populacional e suas implicações para a gestão do rebanho a longo prazo, podendo contribuir significativamente para o desenvolvimento de estratégias de produção animal mais eficientes, sustentáveis e adaptadas às necessidades específicas de cada sistema de produção.

REFERÊNCIAS BIBLIOGRÁFICAS

Ahmadyfard, A., & Modares, H. (2008). Combining PSO and k-means to enhance data clustering. In *2008 international symposium on telecommunications*. IEEE.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Alam, A., Muqeem, M., & Ahmad, S. (2021). Comprehensive review on Clustering Techniques and its application on High Dimensional Data. *International Journal of Computer Science & Network Security*, *21*(6).

Asem-Hiablie, S., Rotz, C. A., Stout, R., & Fisher, K. (2017). Management characteristics of beef cattle production in the western United States. *The Professional Animal Scientist*, *33*(4).

Associação Nacional de Criadores Herd-Book Collares (2018). *Manual do Usuário PROMEBO*. Pelotas, RS. Recuperado de: https://angus.org.br/wp-content/uploads/2018/04/Manual-do-Usu%C3%A1rio-PROMEBO.pdf

Balogh, J. M., & Jámbor, A. (2020). The environmental impacts of agricultural trade: A systematic literature review. *Sustainability*, *12*(3).

Bandyopadhyay, S. (2011). Genetic algorithms for clustering and fuzzy clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(6).

Barceló-Rico, F., Díez, J. L., & Bondia, J. (2011). A comparative study of codification techniques for clustering heart disease database. *Biomedical Signal Processing and Control*, *6*(1), 64-69.

Ben-David, S., & Ackerman, M. (2008). Measures of clustering quality: A working set of axioms for clustering. *Advances in neural information processing systems*, *21*.

Boussaid, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization metaheuristics. *Information sciences*, *237*.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-

dimensional data: A review. *Computational Statistics & Data Analysis*, *71*.

Cacchiani, V., Iori, M., Locatelli, A., & Martello, S. (2022). Knapsack problems—An overview of recent advances. Part II: Multiple, multidimensional, and quadratic knapsack problems. *Computers & Operations Research*, *143*.

Caruso, G., Gattone, S. A., Fortuna, F., & Di Battista, T. (2018). Cluster analysis as a decision-making tool: a methodological review. In *Decision Economics: In the Tradition of Herbert A. Simon's Heritage: Distributed Computing and Artificial Intelligence, 14th International Conference*. Springer International Publishing.

Chen, C. Y., & Ye, F. (2012). Particle swarm optimization algorithm and its application to clustering analysis. In *2012 Proceedings of 17th Conference on Electrical Power Distribution*. IEEE.

Chinnamgari, S. K. (2019). R Machine Learning Projects: Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5. *Packt Publishing Ltd.*

Chiou, Y. C., & Lan, L. W. (2001). Genetic clustering algorithms. *European Journal of Operational Research*, *135*.

Cooper, A., & Folta, T., (2017). Entrepreneurship and high-technology clusters. *The Blackwell handbook of entrepreneurship*.

Ezugwu, A. E., Shukla, A. K., Nath, R., Akinyelu, A. A., Agushaka, J. O., Chiroma, H., & Muhuri, P. K. (2021). Metaheuristics: a comprehensive overview and classification along with bibliometric analysis. *Artificial Intelligence Review*, *54*.

Faisal, M., & Zamzami, E. M. (2020, June). Comparative analysis of inter-centroid K-Means performance using euclidean distance, canberra distance and manhattan distance. In *Journal of Physics: Conference Series*, 1566.

Gaertler, M. (2005). Clustering. *Network analysis: Methodological foundations*.

Ghazal, T. M., Hussain, M. Z., Said, R. A., Nadeem, A., Hasan, M. K., Ahmad, M., Khan, M. A., & Naseem, M. T. (2021). Performances of K-Means Clustering Algorithm with Different Distance Metrics. *Intelligent Automation & Soft*

Computing, *30*.

Gupta, M. K., & Chandra, P. (2020). An empirical evaluation of K-means clustering algorithm using different distance/similarity metrics. In *Proceedings of ICETIT 2019: Emerging Trends in Information Technology*. Springer International Publishing.

Hancer, E., Xue, B., & Zhang, M. (2020). A survey on feature selection approaches for clustering. *Artificial Intelligence Review*, *53*.

Hlavatý, R., Krejčí, I., Houška, M., Moulis, P., Rydval, J., Pitrová, J., & Tichá, I. (2023). Understanding the decision-making in small-scale beef cattle herd management through a mathematical programming model. *International Transactions in Operational Research*, *30.*

Hussain, K., Mohd Salleh, M. N., Cheng, S., & Shi, Y. (2019). Metaheuristic research: a comprehensive survey. *Artificial intelligence review*, *52*.

Landau, S., Leese, M., Stahl, D., & Everitt, B. S. (2011). *Cluster analysis*. John Wiley & Sons.

Li, S., Wei, Y., Liu, X., Zhu, H., & Yu, Z. (2022). A new fast ant colony optimization algorithm: the saltatory evolution ant colony optimization algorithm. *Mathematics*, *10*(6).

Liu, Q., Li, X., Liu, H., & Guo, Z. (2020). Multi-objective metaheuristics for discrete optimization problems: A review of the state-of-the-art. *Applied Soft Computing*, *93*, 106382.

Liu, Y., Wu, X., & Shen, Y. (2011). Automatic clustering using genetic algorithms. *Applied mathematics and computation*, *218*(4), 1267-1279.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010, December). Understanding of internal clustering validation measures. In *2010 IEEE international conference on data mining* (pp. 911-916). IEEE.

Jain, K., Dubes, C. (1988). Algorithms for clustering data. *Prentice-Hall*, Inc.

Jolly, K. (2018). Machine learning with scikit-learn quick start guide: classification, regression, and clustering techniques in Python. *Packt Publishing Ltd*.

Kapil, S., & Chawla, M. (2016, July). Performance evaluation of K-means clustering algorithm with various distance metrics. In *2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES)*.

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(3).

Nanda, S. J., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary computation*, *16*.

Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern recognition*, *33*(9).

Martinez, C. C., Maples, J. G., & Benavidez, J. (2021). Beef cattle markets and covid-19. *Applied Economic Perspectives and Policy*, *43*(1).

Mc Hugh, N., Evans, R. D., Amer, P. R., Fahey, A. G., & Berry, D. P. (2011). Genetic parameters for cattle price and body weight from routinely collected data at livestock auctions and commercial farms. *Journal of Animal Science*, *89*(1).

Mc Hugh, N., Fahey, A. G., Evans, R. D., & Berry, D. P. (2010). Factors associated with selling price of cattle at livestock marts. *Animal*, *4*(8).

Menendez, H. (2021). Clustering: finding patterns in the darkness. *Open Journal of Machine Learning*, *1*(1), 1-28.

Milan, S. T., Rajabion, L., Ranjbar, H., & Navimipour, N. J. (2019). Nature inspired meta-heuristic algorithms for solving the load-balancing problem in cloud environments. *Computers & Operations Research*, *110*.

Mueller, M. L., & Van Eenennaam, A. L. (2022). Synergistic power of genomic selection, assisted reproductive technologies, and gene editing to drive genetic improvement of cattle. *CABI Agriculture and Bioscience*, *3*(1).

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6).

Osaba, E., Yang, X. S., & Del Ser, J. (2020). Traveling salesman problem: a perspective review of recent research and new results with bio-inspired metaheuristics. *Nature-Inspired Computation and Swarm Intelligence*, 135-164.

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: Application and trends. *Artificial Intelligence Review*, *56*(7).

Patil, C., & Baidari, I. (2019). Estimating the optimal number of clusters k in a dataset using data depth. *Data Science and Engineering*, *4*, 132-140.

Pereira, L. S., Brunes, L. C., Baldi, F., Do Carmo, A. S., Soares, B. B., Magnabosco, V., & Magnabosco, C. U. (2023). Genetic association between feed efficiency, growth, scrotal circumference, and carcass traits in Guzerat cattle. *Tropical Animal Health and Production*, *55*(2).

Reddy, C. K., & Vinzamuri, B. (2018). A survey of partitional and hierarchical clustering algorithms. *Data clustering*.

Rodrigues Fortes, A., Ferreira, V., Barbosa Simões, E., Baptista, I., Grando, S., & Sequeira, E. (2020). Food systems and food security: the role of small farms and small food businesses in Santiago Island, Cabo Verde. *Agriculture*, *10*(6).

Ridge, E. E., Foster, M. J., & Daigle, C. L. (2020). Effect of diet on non-nutritive oral behavior performance in cattle: a systematic review. *Livestock Science*, *238*.

Runkler, T. A. (2005). Ant colony optimization of clustering models*. International Journal of Intelligent Systems*, 20(12).

Sanchez-Sabate, R., & Sabaté, J. (2019). Consumer attitudes towards environmental concerns of meat consumption: A systematic review. *International journal of environmental research and public health*, *16*(7).

Shelokar, P. S., Jayaraman, V. K., & Kulkarni, B. D. (2004). An ant colony approach for clustering. *Analytica chimica acta*, 509(2).

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and*

Networking, *2021*(1).

Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, *67*(10).

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, *53*(2).

Surono, S., & Putri, R. D. A. (2021). Optimization of fuzzy c-means clustering algorithm with combination of minkowski and chebyshev distance using principal component analysis. *International Journal of Fuzzy Systems*, *23*(1).

Susskind, D. (2020). A world without work: Technology, automation and how we should respond. *Penguin UK*.

Thrun, M. C. (2021). Distance-based clustering challenges for unbiased benchmarking studies. *Scientific reports*, *11*(1).

Trejos, J., Murillo, A., & Piza, E. (2004). Clustering by ant colony optimization. *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies*.

Ullah, Z., Al-Turjman, F., Mostarda, L., & Gagliardi, R. (2020). Applications of artificial intelligence and machine learning in smart cities. *Computer Communications*, *154*.

Vale, P., Gibbs, H., Vale, R., Christie, M., Florence, E., Munger, J., & Sabaini, D. (2019). The expansion of intensive beef farming to the Brazilian Amazon. *Global Environmental Change*, *57*.

Van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(3).

Van der Merwe, D. W., & Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. *Congress on Evolutionary Computation.*

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals*

of data science, *2*.

Zhou, S., Xu, Z., & Liu, F. (2016). Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. *IEEE transactions on neural networks and learning systems*, *28*(12).

Wang, W., & Siau, K. (2019). Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management*, *30*(1).

ANEXOS

ANEXO A - RESULTADO DO PROCESSAMENTO PARA O CONJUNTO B1

Neste anexo é apresentado os resultados obtidos nos 10 processamento realizados para cada uma das três meta-heurísticas utilizando as três métricas de distância para o Conjunto B1. Como cada processamento foi realizado para um número de cluster variando de 2 a 6, a coluna “Cluster” é formada pelo número de cluster que obteve o melhor resultado para cada processamento.

Proces	AG - EUCLIDIANA			Cluster
	SH	DB	CH	
1	0,22	1,24	24,6	2
2	0,21	1,25	23,7	2
3	0,22	1,24	24,6	2
4	0,23	1,22	26,0	2
5	0,23	1,23	25,5	2
6	0,23	1,23	25,4	2
7	0,22	1,24	24,7	2
8	0,19	1,31	23,4	3
9	0,23	1,23	25,1	2
10	0,22	1,24	24,7	2

Proces	AG - MANHATHAN			Nº de Cluster
	SH	DB	CH	
1	0,23	1,23	25,6	2
2	0,24	1,22	25,7	2
3	0,23	1,23	25,6	2
4	0,24	1,22	25,8	2
5	0,25	1,20	30,0	2
6	0,24	1,22	28,0	2
7	0,25	1,20	30,0	2
8	0,25	1,20	30,0	2
9	0,25	1,20	30,0	2
10	0,24	1,22	27,6	2

Proces	AG - CHEBYSHEV			Nº de Cluster
	SH	DB	CH	
1	0,23	1,23	25,4	2
2	0,23	1,23	25,4	2
3	0,22	1,24	24,6	2
4	0,23	1,23	25,5	2
5	0,23	1,23	25,4	2
6	0,23	1,23	25,4	2
7	0,22	1,24	24,7	2

8	0,23	1,23	25,6	2
9	0,23	1,23	25,1	2
10	0,22	1,24	24,8	2

Proces	PSO – EUCLIDIANA			Nº de Cluster
	SH	DB	CH	
1	0,22	1,24	24,6	2
2	0,21	1,25	23,7	2
3	0,19	1,30	23,1	3
4	0,19	1,31	23,6	3
5	0,23	1,23	25,5	2
6	0,23	1,23	25,4	2
7	0,22	1,24	24,7	2
8	0,22	1,24	24,6	2
9	0,22	1,24	24,6	2
10	0,21	1,25	23,7	2

Proces	PSO – MANHATHAN			Nº de Cluster
	SH	DB	CH	
1	0,24	1,21	29,1	2
2	0,25	1,20	30,0	2
3	0,23	1,23	25,8	2
4	0,24	1,21	28,2	2
5	0,25	1,20	30,0	2
6	0,24	1,22	28,0	2
7	0,25	1,20	30,0	2
8	0,25	1,20	30,0	2
9	0,25	1,20	30,0	2
10	0,24	1,22	28,2	2

Proces	PSO – CHEBYSHEV			Nº de Cluster
	SH	DB	CH	
1	0,23	1,23	25,4	2
2	0,23	1,23	25,4	2
3	0,22	1,24	24,8	2
4	0,23	1,23	25,5	2
5	0,23	1,23	25,4	2
6	0,23	1,23	25,4	2
7	0,22	1,24	24,7	2
8	0,23	1,23	25,5	2
9	0,23	1,23	25,4	2
10	0,22	1,24	24,8	2

Proces	ACO – EUCLIDIANA	Nº de Cluster

	SH	DB	CH	
1	0,19	1,30	23,0	3
2	0,19	1,31	23,6	3
3	0,19	1,31	23,6	3
4	0,22	1,24	24,7	2
5	0,22	1,24	24,8	2
6	0,23	1,23	25,4	2
7	0,23	1,23	25,4	2
8	0,23	1,23	25,5	2
9	0,23	1,23	25,4	2
10	0,23	1,23	25,4	2

Proces	ACO – MANHATHAN			Nº de Cluster
	SH	DB	CH	
1	0,22	1,24	24,6	2
2	0,22	1,24	24,6	2
3	0,22	1,24	24,7	2
4	0,22	1,24	24,7	2
5	0,22	1,24	24,8	2
6	0,19	1,31	23,6	3
7	0,23	1,23	25,4	2
8	0,19	1,31	23,6	3
9	0,22	1,24	24,6	2
10	0,22	1,24	24,6	2

Proces	ACO – CHEBYSHEV			Nº de Cluster
	SH	DB	CH	
1	0,23	1,23	25,4	2
2	0,23	1,23	25,4	2
3	0,22	1,24	24,6	2
4	0,23	1,23	25,5	2
5	0,23	1,23	25,4	2
6	0,19	1,31	23,6	3
7	0,22	1,24	24,7	2
8	0,14	1,47	18,2	4
9	0,23	1,23	25,1	2
10	0,22	1,24	24,8	2

ANEXO B - RESULTADO DO PROCESSAMENTO PARA O CONJUNTO B2

Neste anexo é apresentado os resultados obtidos nos 10 processamento realizados para cada uma das três meta-heurísticas utilizando as três métricas de distância para o Conjunto B2. Como cada processamento foi realizado para um número de cluster variando de 2 a 6, a coluna “Cluster” é formada pelo número de cluster que obteve o melhor resultado para cada processamento.

Proc	AG - EUCLIDIANA			Nº de Cluster
	SH	DB	CH	
1	0,40	0,95	58,2	2
2	0,40	0,94	57,8	2
3	0,40	0,94	58,0	2
4	0,40	0,92	58,6	2
5	0,40	0,94	58,1	2
6	0,42	0,90	60,0	2
7	0,40	0,90	59,7	2
8	0,42	0,90	60,0	2
9	0,42	0,90	60,0	2
10	0,39	0,92	58,6	2

Proc	AG - MANHATHAN			Nº de Cluster
	SH	DB	CH	
1	0,44	0,82	61,0	2
2	0,42	0,88	58,6	2
3	0,44	0,82	60,7	2
4	0,44	0,82	60,2	2
5	0,44	0,82	61,0	2
6	0,44	0,82	61,0	2
7	0,44	0,82	61,0	2
8	0,44	0,82	61,0	2
9	0,44	0,82	61,0	2
10	0,44	0,82	61,0	2

Proc	AG - CHEBYSHEV			Nº de Cluster
	SH	DB	CH	
1	0,40	0,98	55,0	2
2	0,40	0,95	55,3	2
3	0,40	0,95	55,3	2

4	0,40	0,95	55,3	2
5	0,40	0,93	56,2	2
6	0,40	0,93	56,2	2
7	0,40	0,93	56,2	2
8	0,40	0,93	56,2	2
9	0,40	0,93	56,2	2
10	0,40	0,93	56,2	2

Proc	PSO - EUCLIDIANA			Nº de Cluster
	SH	DB	CH	
1	0,43	0,89	60,0	2
2	0,43	0,89	60,0	2
3	0,40	0,90	59,7	2
4	0,43	0,89	60,0	2
5	0,43	0,89	60,0	2
6	0,43	0,89	60,0	2
7	0,43	0,89	60,0	2
8	0,43	0,89	60,0	2
9	0,43	0,89	60,0	2
10	0,43	0,89	60,0	2

Proc	PSO - MANHATHAN			Nº de Cluster
	SH	DB	CH	
1	0,44	0,82	60,7	2
2	0,44	0,82	60,7	2
3	0,44	0,80	61,0	2
4	0,44	0,80	61,0	2
5	0,44	0,80	61,0	2
6	0,44	0,80	61,0	2
7	0,44	0,82	60,7	2
8	0,44	0,80	61,0	2
9	0,44	0,80	61,0	2
10	0,44	0,80	61,0	2

Proc	PSO - CHEBYSHEV			Nº de Cluster
	SH	DB	CH	
1	0,40	0,92	55,3	2
2	0,40	0,92	55,3	2
3	0,41	0,90	56,0	2

4	0,41	0,90	56,0	2
5	0,41	0,90	56,0	2
6	0,41	0,90	56,0	2
7	0,40	0,92	55,3	2
8	0,41	0,90	55,8	2
9	0,41	0,90	56,0	2
10	0,41	0,90	56,0	2

Proc	ACO - EUCLIDIANA			Nº de Cluster
	SH	DB	CH	
1	0,40	0,90	59,7	2
2	0,41	0,90	56,0	2
3	0,40	0,90	59,7	2
4	0,40	0,90	59,7	2
5	0,40	0,90	59,7	2
6	0,41	0,90	56,0	2
7	0,40	0,90	60,0	2
8	0,40	0,90	60,0	2
9	0,40	0,90	59,7	2
10	0,40	0,90	59,7	2

Proc	ACO - MANHATHAN			Nº de Cluster
	SH	DB	CH	
1	0,41	0,88	59,8	2
2	0,40	0,90	59,7	2
3	0,40	0,90	59,7	2
4	0,41	0,88	59,8	2
5	0,41	0,88	59,8	2
6	0,41	0,88	59,8	2
7	0,41	0,88	59,8	2
8	0,41	0,88	59,8	2
9	0,41	0,88	59,8	2
10	0,40	0,90	59,7	2

Proc	ACO - CHEBYSHEV			Nº de Cluster
	SH	DB	CH	
1	0,40	0,92	55,0	2
2	0,40	0,92	55,0	2
3	0,40	0,93	54,7	2

4	0,40	0,92	55,3	2
5	0,40	0,92	55,3	2
6	0,40	0,93	54,3	2
7	0,40	0,92	55,0	2
8	0,40	0,92	55,3	2
9	0,40	0,92	55,3	2
10	0,40	0,92	55,3	2