Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Departamento de Astronomia

Eduardo Batista

Otimização de Algoritmos de *Machine Learning* para a Detecção de Aglomerados Abertos

São Paulo
2026

Eduardo Batista

Otimização de Algoritmos de *Machine Learning* para a Detecção de Aglomerados Abertos

Trabalho de Conclusão de Curso apresentado ao Departamento de Astronomia do Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Astronomia.

Vertente: Pesquisa Básica

Orientadora: Prof.ª Dr.ª Jane Gregorio-Hetem (IAG/USP)

Coorientador: Prof. Dr. Phillip Andreas Brenner Galli (IAG/USP)

São Paulo

2026

Agradecimentos

Agradeço à minha mãe, Ana Salete Ribeiro da Silva e ao meu pai, José Ricardo Batista, por terem dedicado suas vidas a garantir aos filhos oportunidades que eles próprios não tiveram. Vocês são fundamentais em qualquer conquista que eu possa ter e meus maiores exemplos como seres humanos.

À minha orientadora, Profa. Dra. Jane Gregorio-Hetem, que me acompanha desde o início da graduação. Agradeço pela atenção constante, por cada reunião e conversa, e pelas inúmeras oportunidades ao longo desses anos.

Aos amigos do meu grupo de pesquisa, pela convivência, pelas discussões e pelos momentos de descontração que tornaram essa jornada muito mais leve. Aos meus amigos de turma e de outros cursos, que compartilharam comigo todos os momentos da graduação.

Aos meus amigos de infância, em especial Gustavo, Kaio e Paulo, com quem cresci e divido a vida há tantos anos, não poderia ter escolhido amizades melhores.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro concedido por meio do processo nº 2024/23573-0.

Ao Instituto de Astronomia, Geofísica e Ciências Atmosféricas da USP e a todos os seus funcionários, técnicos e equipes de limpeza e segurança, pelo suporte e pelo ambiente acadêmico que foram fundamentais para a minha formação.

Esta tese/dissertação foi escrita em LaTeX com a classe IAGTESE, para teses e dissertações do IAG.

"*É necessário sempre acreditar que um sonho é possível*
Que o céu é o limite e você, truta, é imbatível"

Racionais Mc's

Resumo

Aglomerados estelares abertos são fundamentais para compreender a formação e a evolução estelar na Via Láctea. Embora o censo dessas estruturas ainda seja incompleto, o lançamento da missão *Gaia*, que tem uma precisão astrométrica sem precedentes, transformou a área, revelando centenas de novos aglomerados.

Para detectá-los em meio ao denso campo estelar, algoritmos de *machine learning* baseados em densidade, como o DBSCAN, são extremamente efetivos ao comparar membros de aglomerados e estrelas de campo. Contudo, sua eficácia depende criticamente dos parâmetros iniciais escolhidos de forma arbitrária, o que frequentemente gera resultados inconsistentes.

Para superar essa limitação, este trabalho propõe uma varredura sistemática no espaço de parâmetros do DBSCAN. O objetivo é testar e otimizar a capacidade de recuperação de aglomerados ao entender o comportamento de estruturas em diferentes parâmetros, tanto em ambiente simulado quanto em dados observacionais reais. Para isso, utilizamos o algoritmo `OCELOT` para incluir na análise aglomerados sintéticos de alta fidelidade em campos reais observados pelo *Gaia*. A partir do comportamento e da persistência dessas estruturas, treinamos um classificador *XGBoost* capaz de distinguir aglomerados físicos reais de estruturas espúrias. Ao ser avaliado em um conjunto de teste com 50 campos estelares, nosso modelo recuperou com sucesso 81,3% dos aglomerados reais, superando significativamente a sensibilidade de 62% da abordagem tradicional do DBSCAN com parâmetros fixos.

Abstract

Open star clusters are fundamental for understanding star formation and evolution in the Milky Way. Although the census of these structures remains incomplete, the launch of the *Gaia* mission, which provides unprecedented astrometric precision, has transformed the field, revealing hundreds of new clusters.

To detect them amidst dense stellar fields, density-based *machine learning* algorithms, such as DBSCAN, are highly effective at comparing cluster members and field stars. However, their effectiveness critically depends on arbitrarily chosen initial parameters, which frequently yields inconsistent results.

To overcome this limitation, this work proposes a systematic scan of the DBSCAN parameter space. The objective is to test and optimize cluster recovery capabilities by understanding the behavior of structures under different parameters, both in simulated environments and in actual observed fields. To this end, we use the `OCELOT` algorithm to inject high-fidelity synthetic clusters into fields observed by *Gaia*. Based on the behavior and persistence of these structures, we trained an *XGBoost* classifier capable of distinguishing real physical clusters from spurious structures. When evaluated on a test set of 50 stellar fields, our model successfully recovered 81.3% of the real clusters, significantly outperforming the 62% sensitivity of the traditional fixed-parameter DBSCAN approach.

Lista de Figuras

2.1 Representação esquemática da classificação do DBSCAN para um *cluster* com $\min_{\mathrm{PTS}}$=3. 20
2.2 Principais configurações detectadas devido à interação do algoritmo com o campo estelar. 24
2.3 Mapas de avaliação paramétrica do DBSCAN sob um campo estelar puramente composto por ruído de fundo. 26

3.1 Comparação da distribuição Galáctica de 4000 aglomerados catalogados por Hunt e Reffert (2023) e 4000 aglomerados simulados. 28
3.2 Exemplo de um aglomerado simulado com fotometria dispersa por avermelhamento diferencial. 29

4.1 Distribuição dos aglomerados presentes na Simulação 01. 33
4.2 F1-Score individual para cada aglomerado simulado da Simulação 01. . . . 35
4.3 Mapas de pertinência paramétrica para os grupos obtidos na Simulação 01. 37
4.4 Painel de diagnóstico para a Simulação 01. 37
4.5 Matriz de confusão obtida sobre o conjunto de validação. 40
4.6 Comparação entre as distribuições dos aglomerados simulados e detectados para Simulação 01. 41

A.1 Distribuição dos aglomerados obtidos pela metodologia na região de CMa. 54

Lista de Tabelas

3.1 Distribuições dos parâmetros de aglomerados simulados. 30

A.1 Parâmetros dos aglomerados CMa detectados pela metodologia. 56

Sumário

1. Introdução 17

2. Fundamentos Teóricos 19

2.1 DBSCAN 19

2.1.1 Normalização dos dados 21

2.1.2 Densidade Mínima 21

2.2 Evolução Estrutural no Espaço Paramétrico 22

2.3 Métricas de Validação 25

3. Simulação de Aglomerados Estelares 27

3.1 Simulação de um campo de aglomerados estelares 31

4. Resultados 33

4.1 Validação da Métrica de Seleção 33

4.1.1 Validação em ambiente simulado: F1-Score 34

4.1.2 Persistência paramétrica e critérios de seleção 35

4.2 Classificação por *XGBoost* 38

4.2.1 Conjunto de treinamento 38

4.2.2 Validação e resultados 39

5. Conclusões e Perspectivas Futuras 43

Referências 47

Apêndice 51

A. Aplicação do algoritmo na Associação Estelar Canis Major 53

A.1 A região de Canis Major OB1 53

Capítulo 1

Introdução

O censo de aglomerados abertos da Via Láctea permanece incompleto mesmo na vizinhança solar. Kharchenko et al. (2013) catalogaram cerca de 3000 desses objetos com distâncias até 1,8 kpc. Posteriormente, diversas buscas encontraram novos aglomerados devido aos dados cada vez mais precisos, revelando novas estruturas mesmo em regiões consideradas bem estudadas, indicando que o censo de aglomerados abertos está longe de ser completo.

Aglomerados abertos são grupos de estrelas com origem em comum, formados a partir do colapso de uma mesma nuvem molecular. Por esse motivo, esses objetos compartilham uma evolução dinâmica em comum, tornando as posições e os movimentos correlacionados. Diferentemente, estrelas de campo apresentam distribuição dispersa e não correlacionada no espaço astrométrico, de modo que as correlações nos dados para membros de um mesmo aglomerado se manifestam como sobredensidades no espaço de parâmetros. Essa característica implica que a qualidade astrométrica dos dados é diretamente determinante para a detectabilidade dessas estruturas, aumentando a capacidade de separação entre aglomerados e campo. Nesse contexto, a missão *Gaia* (Gaia Collaboration et al., 2016) transformou a busca de aglomerados estelares ao fornecer astrometria em posição (α, δ), movimento próprio $(\mu_{\alpha^*}, \mu_\delta)$ e paralaxe (ϖ) para mais de um bilhão de fontes estelares com uma precisão sem precedentes. O desafio passa a ser, portanto, desenvolver métodos capazes de identificar eficientemente essas sobredensidades em um espaço de cinco dimensões densamente populado por estrelas de campo.

Algoritmos baseados em densidade como o DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) (Ester et al., 1996) exploram diretamente essa propriedade ao buscarem regiões do espaço de parâmetros com densidade local acima de um limiar definido

por ϵ (raio de busca) e $\min_{\rm PTS}$ (número mínimo de vizinhos). Contudo, a eficácia do DBSCAN depende criticamente da escolha desses parâmetros. Castro-Ginard et al. (2019) mostram que o fator limitante de detecção é o contraste entre o aglomerado e o campo estelar, de modo que objetos de baixo contraste simplesmente não são recuperados para nenhuma combinação razoável de parâmetros fixos. A abordagem tradicional de selecionar um único par $(\epsilon, \min_{\rm PTS})$ por região ignora que aglomerados com densidades intrínsecas distintas são detectáveis em regiões distintas do espaço paramétrico, introduzindo um viés sistemático contra estruturas mais difusas.

Para superar essa limitação, nossa metodologia utiliza uma varredura sistemática de diferentes combinações de parâmetros $(\epsilon, \min_{\rm PTS})$. Ao testar múltiplos parâmetros, avaliamos se uma estrutura continua sendo identificada sob diferentes critérios de densidade de busca, definindo essa estabilidade como a persistência da estrutura, que passa a ser o próprio critério de detecção. Grupos que se mantêm estáveis em uma ampla faixa do espaço de parâmetros são candidatos a aglomerados físicos reais, pois sua existência não depende de uma escolha arbitrária de busca. Por outro lado, agrupamentos esparsos apresentam baixa persistência e mostram-se ser flutuações estatísticas do campo estelar.

Para distinguir aglomerados reais de estruturas espúrias, utilizamos um classificador baseado em aprendizado de máquina treinado em campos com aglomerados simulados de alta fidelidade física, reconhecendo padrões dessas regiões para atribuir a cada grupo detectado uma probabilidade de ser um aglomerado real. A metodologia é então avaliada em um conjunto independente de campos estelares reais.

Capítulo 2

Fundamentos Teóricos

A identificação de aglomerados estelares em meio a uma região densamente povoada por estrelas de campo exige métodos robustos capazes de distinguir estruturas a partir de dados multidimensionais. Para isso, as técnicas utilizadas se baseiam na análise dos parâmetros astrométricos do catálogo *Gaia DR3* (Gaia Collaboration et al., 2023), que, em conjunto, permitem caracterizar a posição e a cinemática estelar. O princípio para a detecção dessas estruturas é a busca por excessos de densidade no espaço de parâmetros, pois enquanto as estrelas de campo distribuem-se de forma mais dispersa e aleatória, os aglomerados manifestam-se como concentrações significativamente mais densas e correlacionadas. Entre as abordagens mais eficazes para explorar essa característica estão os algoritmos de agrupamento baseados em densidade, como o DBSCAN.

2.1 DBSCAN

O DBSCAN, proposto por Ester et al. (1996), é um algoritmo de *machine learning* projetado para identificar agrupamentos de formas arbitrárias em regiões imersas em ruído, como um aglomerado aberto em um campo estelar. O algoritmo utiliza o conceito de densidade local para agrupar pontos, permitindo que aglomerados sejam detectados sem a necessidade de especificar previamente o número de agrupamentos, como ocorre em métodos particionais. Essa característica torna o DBSCAN especialmente útil em situações em que o conhecimento prévio sobre a estrutura dos dados é limitado ou em que os aglomerados podem ter formas complexas.

Para implementar esse conceito, é utilizada a densidade em torno de um ponto específico. O tamanho da região em torno desse ponto é delimitado por um parâmetro chamado

epsilon (ϵ), que define a distância máxima para considerar outro ponto como vizinho (Figura 2.1). Assim, dados dois pontos, a distância entre eles é calculada usando uma métrica escolhida pelo usuário, como a distância Euclidiana. Outro parâmetro fundamental é o $\min_{\mathrm{PTS}}$, que representa o número mínimo de pontos necessários para que uma região seja considerada densa o suficiente para formar um cluster. Com esses dois parâmetros, os pontos de uma base de dados são classificados em três categorias representadas na Figura 2.1:

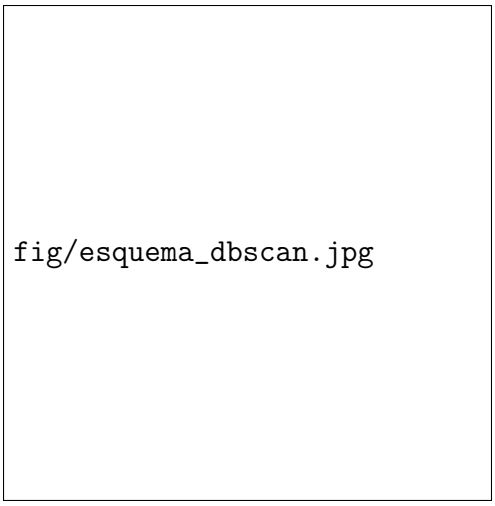

Figura 2.1: Representação esquemática da classificação do DBSCAN para um *cluster* com $\min_{\mathrm{PTS}}$=3. Símbolos em verde, azul e vermelho representam pontos centrais, pontos de borda e ruídos, respectivamente. Figura retirada de Castro-Ginard et al. (2019).

1. Pontos Centrais: Esses são os pontos que possuem uma densidade suficiente ao seu redor, ou seja, o número de vizinhos com distância menor que ϵ é maior ou igual a $\min_{\mathrm{PTS}}$. Eles formam o núcleo dos aglomerados e são essenciais para expandir o agrupamento.

2. Pontos de borda: São pontos que têm distância menor que ϵ de um ponto central, mas não possuem densidade suficiente para serem considerados centrais. Eles estão nos arredores de um *cluster* e ajudam a definir seus limites.

3. Ruído (estrelas de campo no caso de aglomerados estelares): Pontos que não estão conectados a nenhum *cluster*. Esses pontos estão isolados ou em regiões de baixa densidade e são ignorados no agrupamento.

Assim, é perceptível que os resultados obtidos pelo DBSCAN são dependentes da escolha adequada dos parâmetros ϵ e $\min_{\mathrm{PTS}}$, que deve seguir um critério bem fundamentado.

2.1.1 Normalização dos dados

Como a aplicação do DBSCAN depende fundamentalmente do cálculo de distâncias entre pontos, é essencial que os parâmetros utilizados estejam em uma escala comparável. Caso contrário, atributos com valores numericamente maiores dominam o cálculo da distância e o critério de vizinhança definido por ϵ deixa de refletir a proximidade conjunta dos cinco parâmetros astrométricos.

Neste trabalho, adotamos a padronização dos dados, na qual cada parâmetro é centrado em sua média $\bar{x}$ e dividido pelo desvio padrão σ do campo estelar:

$$z = \frac{x - \bar{x}}{\sigma}, \tag{2.1}$$

de modo que cada parâmetro passa a ter média zero e desvio padrão unitário. Os cinco parâmetros astrométricos do *Gaia* (α, δ, $\mu_{\alpha^*} = \mu_\alpha \cos\delta$, μ_δ, ϖ) são transformados independentemente, originando cinco novas coordenadas adimensionais que denotamos por:

$$z_1 \equiv z_\alpha, \quad z_2 \equiv z_\delta, \quad z_3 \equiv z_{\mu_{\alpha^*}}, \quad z_4 \equiv z_{\mu_\delta}, \quad z_5 \equiv z_\varpi. \tag{2.2}$$

A distância euclidiana entre dois pontos i e j é então calculada nesse espaço normalizado:

$$d(i,j) = \sqrt{\sum_{k=1}^{5} \left(z_{k,i} - z_{k,j}\right)^2}. \tag{2.3}$$

A partir daqui, todas as referências a ϵ assim como as densidades derivadas dele dizem respeito a este espaço adimensional. Em particular, ϵ não possui unidade física pois representa uma distância em unidades de desvio padrão conjunto dos cinco parâmetros astrométricos.

2.1.2 Densidade Mínima

A densidade de objetos (ρ) no espaço normalizado é definida como o número de pontos por unidade de hipervolume no espaço transformado de cinco dimensões. Para o algoritmo DBSCAN, a densidade mínima de busca (denominada $\rho_{\min}$) necessária para formar um aglomerado pode ser expressa como:

$$\rho_{\min} = \frac{\min_{\mathrm{PTS}}}{V_n(\epsilon)}, \tag{2.4}$$

onde $\min_{\mathrm{PTS}}$ é o número de pontos necessário para que o objeto seja considerado um ponto central e $V_n(\epsilon)$ representa o volume de uma hiperesfera de raio ϵ em n dimensões. O volume de uma hiperesfera de raio ϵ em n dimensões é dado por:

$$V_n(\epsilon) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\epsilon^n. \tag{2.5}$$

Para o caso particular de $n = 5$ dimensões, que corresponde aos cinco parâmetros astrométricos normalizados, o volume se torna:

$$V_5(\epsilon) = C \cdot \epsilon^5, \tag{2.6}$$

onde a constante C é dada por $C = \frac{\pi^{2,5}}{\Gamma(3,5)} \approx 5{,}26$ e Γ é a função Gamma. Substituindo a Equação 2.6 na Equação 2.4 e rearranjando os termos, obtemos a relação que mostra todos os parâmetros do DBSCAN em que um objeto com densidade $\rho \geq \rho_{\min}$ é obtido.

$$\min_{\mathrm{PTS}} = \rho_{\min} \cdot C \cdot \epsilon^5 \tag{2.7}$$

Esta relação mostra que, para um dado valor de $\rho_{\min}$ constante, existem infinitas combinações de ϵ e $\min_{\mathrm{PTS}}$ que satisfazem a equação. Ou seja, esse conjunto de combinações forma uma curva característica no espaço paramétrico, ao longo da qual um aglomerado com densidade $\rho \geq \rho_{\min}$ pode ser detectado.

Entretanto, essa curva apresenta apenas objetos em que o número de pontos e o volume considerado aumentam proporcionalmente ($\rho_{\min}$ = constante), o que não é necessariamente verdade para aglomerados estelares, em que a densidade de objetos é variável. Desse modo, um aglomerado que tem densidade variável não deve ser identificado exatamente ao longo da curva dada pela Equação 2.7, mas próxima.

2.2 Evolução Estrutural no Espaço Paramétrico

A relação matemática estabelecida pela Equação 2.7 demonstra que a varredura bidimensional do DBSCAN, definida por ϵ e $\min_{\mathrm{PTS}}$, tem uma implicação física direta, ditando a capacidade do algoritmo de identificar estruturas com diferentes densidades ($\rho_{\min}$). Em

um campo observado qualquer, a distribuição estelar não é uniforme nos parâmetros astrométricos, apresentando flutuações que representam desde o ruído das estrelas de campo até as regiões densamente povoadas de aglomerados reais. Dessa forma, ao percorrermos o espaço de parâmetros, obtemos diferentes configurações de estruturas que evoluem de maneira não linear e que podem ou não ser os aglomerados físicos que buscamos. Assim, distinguir aglomerados reais (maior pertinência em diferentes densidades) e objetos que emergem da densidade de campo (menor pertinência) é o principal desafio ao utilizar o DBSCAN.

Para entender como estrelas de campo podem produzir aglomerados, consideramos uma amostra astrométrica com aproximadamente 53 000 objetos de uma região galáctica do disco fino, centrada nas coordenadas $\alpha = 206°$ e $\delta = 49°$ e que não contém aglomerados conhecidos (Figura 2.2). Esse é um teste de campo nulo, que busca caracterizar os objetos de campo no mapa paramétrico. O mapa paramétrico é uma representação bidimensional do espaço de parâmetros do DBSCAN, em que os eixos correspondem a ϵ e min_{PTS}, e a cor de cada ponto codifica o valor de alguma métrica calculada para aquela combinação específica (como o número de aglomerados detectados). Mostraremos que, mesmo na ausência de qualquer aglomerado real, o mapa paramétrico revela estruturas intrínsecas ao campo quando submetido a diferentes limiares de densidade (ρ_{min}) chamados de fases de alta restrição, fragmentação e aglutinação.

O primeiro desses cenários, que denominamos de fase de alta restrição, ocorre quando a densidade de busca exigida pelo algoritmo é superior à densidade volumétrica das estruturas físicas presentes na região ($\rho_{\text{min}} \gg \rho_{\text{aglomerados}}$). Neste regime, a combinação de parâmetros (tipicamente configurada com valores de ϵ extremamente baixos e/ou min_{PTS} muito elevados) impede que qualquer região multidimensional satisfaça o critério de vizinhança mínima pela baixa densidade, resultando na classificação de toda a amostra astrométrica como ruído de campo, conforme apresentado na Figura 2.2a.

Contudo, à medida que a densidade de busca diminui gradativamente até se aproximar da densidade média do fundo estelar ($\rho_{\text{min}} \approx \rho_{\text{campo}}$), o algoritmo entra na fase de fragmentação. Neste regime, o DBSCAN torna-se sensível a flutuações estatísticas locais da distribuição estelar. Estrelas de campo que, aleatoriamente, demonstram uma densidade espacial e cinemática ligeiramente maior que o campo no espaço de cinco dimensões, satisfazem os critérios de vizinhança. Observa-se, então, um aumento abrupto e artificial

no número de aglomerados detectados (Figura 2.2b). Trata-se de pequenos agrupamentos, desprovidos de significado físico ou coesão gravitacional, gerados unicamente por pequenas sobredensidades estatisticamente esperadas em uma grande quantidade de dados.

Se o limiar de densidade for reduzido ainda mais, caindo abaixo da densidade média do campo ($\rho_{\text{min}} < \rho_{\text{campo}}$), o comportamento do algoritmo transita para a fase de aglutinação. Este regime é impulsionado por uma vulnerabilidade inerente ao DBSCAN conhecida como efeito de encadeamento. Devido à sua formulação baseada em conectividade transitiva, os falsos agrupamentos gerados na fase anterior começam a se fundir. Como resultado, o número de aglomerados diminui rapidamente pois todos os aglomerados da fase de aglutinação se fundem em um único com um número de estrelas extremamente alto, como apresentado na Figura 2.2c.

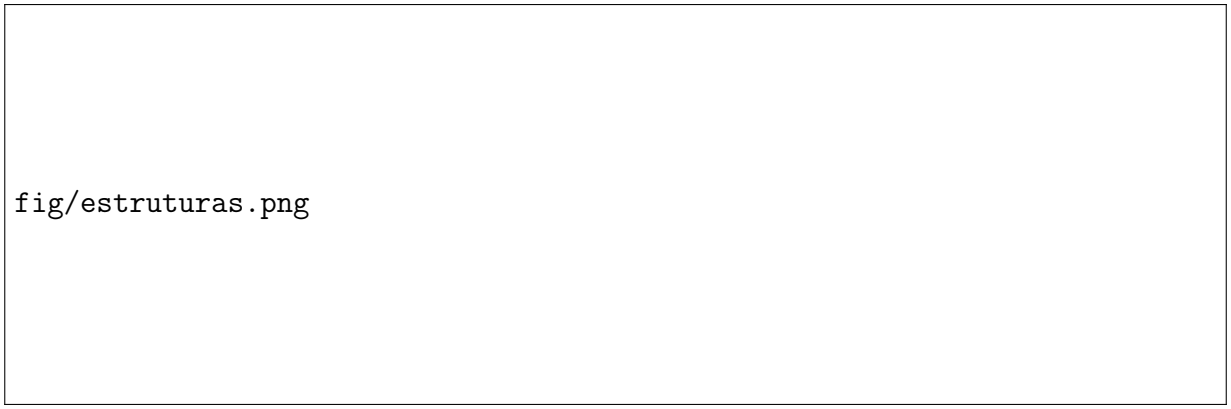

Figura 2.2: Principais configurações detectadas devido à interação do algoritmo com o campo estelar. (a) Fase de alta restrição ($\rho_{\text{min}} \gg \rho_{\text{campo}}$): a densidade exigida é tão superior à do fundo que nenhum subgrupo é validado, classificando toda a amostra como ruído (em cinza). (b) Fase de fragmentação ($\rho_{\text{min}} \approx \rho_{\text{campo}}$): a densidade de busca se aproxima à do campo, fazendo com que sobredensidades locais aleatórias gerem múltiplos aglomerados espúrios. (c) Fase de aglutinação ($\rho_{\text{min}} \ll \rho_{\text{campo}}$): a densidade mínima de busca é muito inferior à do campo, resultando no efeito de encadeamento que funde as estrelas em um grupo estelar único.

Assim, é perceptível que apenas a existência de objetos de campo gera diferentes estruturas no espaço paramétrico. Mostra-se, portanto, estritamente necessário utilizar técnicas para desfavorecer tais regiões nos mapas paramétricos, isolando e privilegiando as configurações que contenham aglomerados fisicamente pertinentes.

2.3 Métricas de Validação

Como demonstrado na Seção 2.2, a sensibilidade aos parâmetros iniciais representa um desafio na aplicação do algoritmo DBSCAN em campos estelares. Diferentemente de métodos particionais, em que o número de aglomerados é definido previamente, a abordagem baseada em densidade exige a escolha de limiares (ϵ e $\min_{\mathrm{PTS}}$) que atualmente não possuem critérios de escolha na literatura. Se não monitorado, o algoritmo inevitavelmente transita por fases de fragmentação ou aglutinação do fundo estelar, gerando resultados irreais.

Para superar essa limitação e remover a subjetividade da escolha paramétrica, desenvolvemos uma abordagem sistemática de varredura bidimensional. Avaliamos um amplo conjunto de combinações de ϵ e $\min_{\mathrm{PTS}}$, extraindo diferentes informações ao calcular métricas topológicas e estruturais para cada configuração. O espaço paramétrico é varrido em intervalos fixos, com $\epsilon \in [0{,}01; 0{,}8]$ em passos de $\Delta\epsilon = 0{,}005$ e $\min_{\mathrm{PTS}} \in [10; 120]$ em passos de $\Delta\min_{\mathrm{PTS}} = 1$, em que cada conjunto de combinações corresponde a uma execução independente do algoritmo. Os resultados são projetados na forma de mapas de calor (*heatmaps*), em que os eixos coordenados representam os parâmetros ϵ e $\min_{\mathrm{PTS}}$, enquanto a cor de cada célula codifica o valor da métrica avaliada naquela configuração, permitindo visualizar a resposta do algoritmo em todo o espaço paramétrico de forma simultânea. Iniciamos essa metodologia medindo, para cada combinação de parâmetros, duas propriedades: o número de agrupamentos detectados e o número de membros do maior aglomerado detectado.

A primeira métrica extraída da varredura é a contagem total de agrupamentos detectados. Conforme discutido, a fase de fragmentação ocorre quando diversas regiões de sobredensidades aparecem e logo se aglutinam em um objeto massivo, ou seja, o número de aglomerados cresce e diminui rapidamente ao chegar próximo à densidade do campo. Aplicamos essa métrica na região de campo nulo discutida anteriormente, obtendo que o número de aglomerados é alto apenas próximo à densidade de campo, o que indica que esses agrupamentos só emergem na região de fragmentação (Figura 2.3a).

Entretanto, o algoritmo permanece vulnerável à fase de aglutinação em buscas com baixa densidade. Para monitorar esse comportamento, extraímos o número de membros do maior aglomerado detectado em cada iteração ($N_{\max}$).

Aglomerados abertos reais e simulados (como veremos posteriormente) presentes no

disco galáctico possuem limites de massa e número de membros bem estabelecidos na literatura. Quando o algoritmo entra na fase de aglutinação devido ao efeito de encadeamento, o valor de $N_{\max}$ cresce rapidamente, incorporando milhares de estrelas de campo indiscriminadamente, pois a densidade de busca é menor que a densidade de campo. Para mitigar esse efeito, implementamos um valor máximo de pertinência para um aglomerado (neste trabalho, consideramos um aglomerado com no máximo 2000 membros).

A aplicação desse corte físico atua como um filtro de qualidade. Se o maior aglomerado ultrapassa esse limiar, ele é descartado, pois é considerado ruído aglutinado. Comparamos os mapas paramétricos do número de aglomerados sem (Figura 2.3b) e com (Figura 2.3c) o limite máximo de 2000 membros para um aglomerado, evidenciando que a limitação permite a visualização das regiões de maior importância, como as estruturas geradas pelo campo, sem perda de informação.

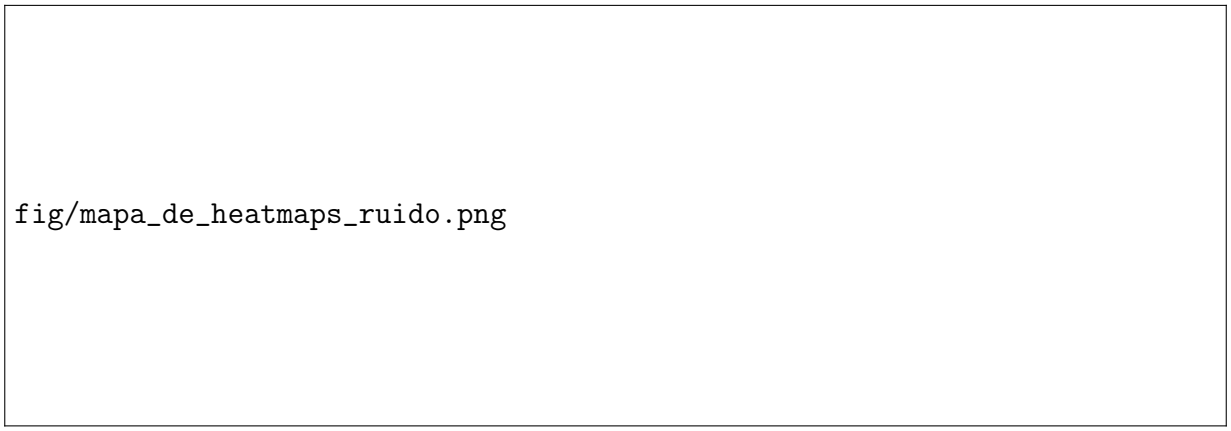

Figura 2.3: Mapas de avaliação paramétrica do DBSCAN sob um campo estelar puramente composto por ruído de fundo. (a) Número total de aglomerados detectados em escala logarítmica. (b) Tamanho do maior aglomerado ($N_{\max}$) sem restrição física, demonstrando o efeito de encadeamento na fase de aglutinação. (c) Tamanho do maior aglomerado com limite máximo de 2000 membros, removendo o falso aglomerado gigante e permitindo visualizar melhor os agrupamentos menores.

Com o comportamento do campo estelar no espaço paramétrico estabelecido, torna-se necessário compreender como aglomerados estelares reais se manifestam nesse mesmo espaço. Para isso, o capítulo seguinte descreve a construção de aglomerados sintéticos injetados em campos reais do catálogo *Gaia* DR3 (Gaia Collaboration et al., 2023), o que nos permitirá identificar, em ambiente controlado, qual região do mapa de parâmetros há detectabilidade, distinguindo estruturas físicas genuínas das flutuações estatísticas do campo.

Capítulo 3

Simulação de Aglomerados Estelares

Para a simulação dos aglomerados estelares sintéticos, utilizamos o código aberto `OCELOT`[1] (Open-source star ClustEr muLti-purpOse Toolkit; Hunt et al., 2026), que implementa um processo de alto grau de realismo físico para gerar e injetar aglomerados sintéticos diretamente nos dados reais do *Gaia*. As escolhas metodológicas e os intervalos dos parâmetros de entrada foram adaptados a partir dos estabelecidos por Hunt et al. (2025).

O processo de simulação inicia com a definição das propriedades globais do aglomerado. A massa total do aglomerado M é amostrada a partir de uma lei de potência que prioriza aglomerados de baixa massa em relação às altas massas (Tabela 3.1), refletindo a distribuição de massa observada na população de aglomerados abertos da Galáxia (Lamers et al., 2005). A idade $[\log t]$ e a metalicidade [M/H] completam os parâmetros fundamentais. A distribuição uniforme para a idade foi delimitada ao intervalo de $\log t \in [6{,}0; 9{,}3]$ (considerando idades entre 1 milhão e 2 bilhões de anos) para refletir o tempo de vida típico de aglomerados abertos no disco galáctico (Portegies Zwart et al., 2010; Krumholz et al., 2019). A partir destas propriedades, a população estelar individual é construída de maneira probabilística. As massas das estrelas são amostradas seguindo a Função de Massa Inicial (IMF) de Kroupa (2001), por meio da biblioteca Python `imf`[2]. Em seguida, para cada estrela, são atribuídas magnitudes absolutas nas bandas fotométricas do *Gaia* (G, G_{BP}, G_{RP}) através da interpolação de isócronas PARSEC (Bressan et al., 2012), garantindo consistência com a idade e metalicidade pré-definidas.

A posição Galáctica do aglomerado é definida em coordenadas galactocêntricas (X, Y, Z).

[1] `https://github.com/emilyhunt/ocelot`

[2] `https://github.com/keflavich/imf`

O intervalo amostrado para X e Y (Tabela 3.1) prioriza a direção do Anticentro Galáctico, estendendo-se rumo às regiões do disco externo. Limitamos a distância dos aglomerados a 5 kpc para otimizar o tempo de processamento. Esta escolha de parâmetros permite avaliar o desempenho de algoritmos de detecção em zonas de menor densidade estelar, as quais permanecem cruciais para a compreensão da estrutura da Via Láctea. A Figura 3.1 compara a distribuição de aglomerados reais obtidos por Hunt e Reffert (2023) com os aglomerados simulados com as distribuições utilizadas nesse trabalho.

A componente vertical Z é amostrada a partir de uma mistura de distribuições para representar a população fina do disco, em que se concentram a maioria dos aglomerados estelares jovens. A estrutura espacial interna é modelada supondo um perfil de densidade radial de King (King, 1962), parametrizado por um raio de núcleo r_c e um raio de maré r_t. A distribuição tridimensional das estrelas é realizada pelo amostrador interno do `OCELOT`.

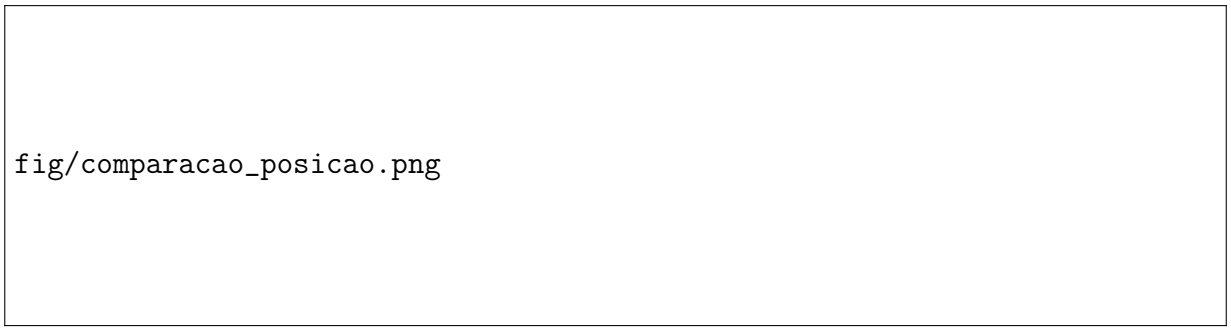

Figura 3.1: Comparação da distribuição Galáctica de (a) 4000 aglomerados catalogados por Hunt e Reffert (2023) e (b) 4000 aglomerados simulados pela distribuição espacial dada na Tabela 3.1

A dinâmica interna é simulada atribuindo a cada estrela uma velocidade amostrada de uma distribuição Gaussiana tridimensional. A dispersão de velocidade é calculada com base na massa do aglomerado, seu raio (aproximado por r_{50}) e seu estado virial, parametrizado por Q ($Q = 0{,}5$ indica equilíbrio virial). Amostramos Q a partir de uma distribuição normal centrada em 0,5, permitindo explorar aglomerados ligeiramente sub ou superviriais. O aglomerado recebe também uma velocidade de órbita global (V_ϕ, V_ρ, V_z) no referencial Galactocêntrico. Os parâmetros físicos são convertidos nos parâmetros astrométricos observáveis pelo *Gaia* $(\delta, \alpha, \mu_\delta, \mu_{\alpha^*}, \varpi)$ utilizando as ferramentas do `astropy` (Astropy Collaboration et al., 2013).

A etapa de transformação para o espaço observacional inclui a aplicação realista da extinção interestelar. Para isso, o valor médio da extinção A_V é obtido a partir das estimativas fornecidas pelo próprio catálogo *Gaia* DR3 na posição e distância do aglomerado.

O efeito do avermelhamento diferencial é simulado utilizando a ferramenta integrada do pacote `OCELOT` (Hunt et al., 2026), que gera um mapa de ruído fractal (ou ruído rosa Bak et al., 1987) para a região de interesse. A cada estrela é atribuída uma variação ΔA_V proporcional ao valor do ruído na sua posição projetada (Figura 3.2). Diferente de uma perturbação aleatória independente, o ruído fractal introduz uma correlação espacial, ou seja, estrelas próximas no céu passam a compartilhar valores de extinção similares. Esse comportamento busca reproduzir a estrutura filamentar da poeira interestelar (Edenhofer et al., 2024), produzindo uma dispersão no diagrama cor-magnitude condizente com a observada em aglomerados reais sujeitos a alto avermelhamento.

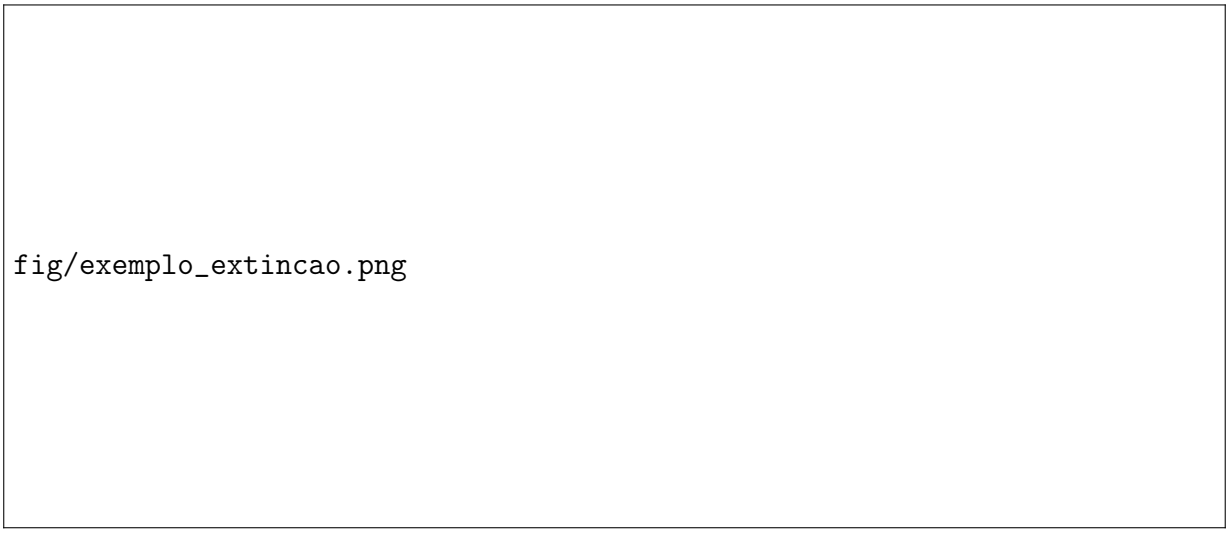

Figura 3.2: Exemplo de um aglomerado simulado com fotometria dispersa por avermelhamento diferencial. O aglomerado tem idade de 10 milhões de anos, massa de 1500 $M_\odot$, extinção de 1,5 mag com 0,2 mag de avermelhamento diferencial e está à distância de 5 kpc. (a) Posição espacial dos membros sobrepostos ao mapa de ruído fractal usado para determinação do avermelhamento diferencial. (b) Diagrama cor-magnitude simulado final para o aglomerado, incluindo o impacto do avermelhamento diferencial e erros fotométricos do Gaia. A linha preta mostra a isócrona que corresponde à idade do aglomerado.

Efeitos instrumentais e observacionais do *Gaia* também são incorporados. Primeiramente, sistemas binários são tratados seguindo as estatísticas de binaridade de Moe e Di Stefano (2017), estrelas são pareadas e, caso sua separação angular projetada seja inferior ao limite de resolução do *Gaia* (aproximadamente 0,6″ conforme Fabricius et al. (2021)), o par é observado como uma única fonte com fotometria combinada. Em segundo lugar, incertezas astrométricas e fotométricas realistas são atribuídas a cada estrela. O `OCELOT` utiliza uma abordagem baseada em dados reais, pois para cada estrela simulada, busca-se no catálogo do *Gaia* DR3, dentro da mesma região do céu, uma estrela real com magnitude

G similar e atribui a ela os erros medidos dessa estrela de referência (Fabricius et al., 2021), capturando a variação dos erros com a magnitude de maneira empírica, a posição no céu e a resolução angular.

O resultado final é um aglomerado sintético que replica, com alto grau de realismo, as características observáveis de um aglomerado real nos dados do *Gaia* DR3, incluindo número de estrelas, dispersão astrométrica, extinção interestelar e erros instrumentais. A Tabela 3.1 sumariza as distribuições de probabilidade adotadas para cada parâmetro fundamental do aglomerado. Para alguns deles, em particular as componentes verticais e cinemáticas, adotamos uma combinação ponderada entre uma distribuição normal $\mathcal{N}$ e uma uniforme $\mathcal{U}$.

Tabela 3.1 - Distribuições dos parâmetros de aglomerados simulados.

Parâmetro	Distribuição	Mín.	Máx.	Unidade
X^a	$\mathcal{U}$	$-23{,}1$	0	kpc
Y^a	$\mathcal{U}$	-15	15	kpc
Z^a	$0{,}8\mathcal{N}(0;0{,}25) + 0{,}2\mathcal{U}$	$-2{,}5$	2,5	kpc
V_ρ	$0{,}8\mathcal{N}(0;35) + 0{,}2\mathcal{U}$	-150	150	$\mathrm{km\,s^{-1}}$
$V_{\phi\text{-circ}}$	$0{,}8\mathcal{N}(0;25) + 0{,}2\mathcal{U}$	-150	150	$\mathrm{km\,s^{-1}}$
V_Z	$0{,}8\mathcal{N}(0;15) + 0{,}2\mathcal{U}$	-100	100	$\mathrm{km\,s^{-1}}$
M	Leis de potência[b]	40	5000	$\mathrm{M_\odot}$
$\log t$	$\mathcal{U}$	6	10	dex
$[M/H]$	$\mathcal{U}$	-1	0,5	dex
A_V	mapas de poeira	0	20	mag
$\Delta A_V/A_V$	$\mathcal{U}$	0	0,5	mag
Q	$\mathcal{N}(0{,}5;0{,}2)$	0,1	10	–
r_c	$\mathcal{U}$	0,1	8	pc
r_J	pot. galáct.	–	–	pc
r_i/r_J	$\mathcal{U}$	0,5	2	–

Notas. $\mathcal{U}$ denota uma distribuição uniforme entre os valores mínimo e máximo. $\mathcal{N}(a;b)$ denota uma distribuição normal truncada entre os valores mínimo e máximo, com média a e desvio padrão b. Para alguns parâmetros, foi utilizada uma combinação ponderada de ambas as distribuições. [(a)] Amostras com distância heliocêntrica total $d > 5$ kpc ou $d < 0{,}5$ kpc foram descartadas. [(b)] Amostra de uma lei de potência com inclinações (em ordem de massa) de $-1{,}013$, $-2{,}606$ e 0,0, com pontos de quebra em $813\,\mathrm{M_\odot}$ e $3364\,\mathrm{M_\odot}$. Tabela adaptada de Hunt et al. (2026).

3.1 Simulação de um campo de aglomerados estelares

A organização espacial e a seleção dos campos estelares utilizados para a injeção dos aglomerados sintéticos baseiam-se no esquema de pixelização HEALPix (Hierarchical Equal Area isoLatitude Pixelization), uma ferramenta para o tratamento de dados astronômicos em toda a esfera celeste (Gorski et al., 1999). Diferentemente de grades baseadas em coordenadas equatoriais, que sofrem com distorções de área e singularidades próximas aos polos, o HEALPix divide a superfície esférica em pixels de áreas exatamente iguais, garantindo que a densidade estelar e a probabilidade de detecção de um aglomerado sejam analisadas independentemente da latitude galáctica. Ao definir a posição central de um aglomerado simulado sob uma resolução de particionamento $N_{side} = 64$, identificamos o pixel correspondente e incluímos automaticamente os seus oito vizinhos imediatos. Este procedimento define campos de simulação com áreas angulares de aproximadamente 7,55 graus quadrados, mitigando falhas de detectabilidade de membros do aglomerado devido a efeitos de borda espacial.

Para representar regiões complexas de formação estelar, cada simulação contém de um a cinco aglomerados injetados no mesmo agrupamento HEALPix, com o número exato de aglomerados é amostrado aleatoriamente. Adicionalmente, consultamos o catálogo para identificar aglomerados reais já conhecidos na mesma faixa de paralaxe da região injetada, o que permite testar diretamente a capacidade do algoritmo de distinguir as estruturas simuladas de outras já catalogadas.

Foram gerados 50 campos sintéticos, cobrindo diferentes direções Galácticas amostradas a partir das mesmas distribuições espaciais (X, Y, Z) descritas na Tabela 3.1. As consultas ao catálogo *Gaia* DR3 foram limitadas a estrelas com $G < 19$, faixa em que a completeza fotométrica e a qualidade astrométrica do catálogo são suficientes para a detecção de aglomerados no regime de contraste explorado neste trabalho.

Capítulo 4

Resultados

4.1 Validação da Métrica de Seleção

Os 50 campos estelares descritos na Seção 3.1 foram processados em cenários distintos, variando posição Galáctica, densidade de fundo, número de aglomerados injetados e parâmetros físicos, de modo a construir uma amostra representativa das condições encontradas no disco da Via Láctea. Utilizaremos os resultados obtidos na Simulação 01 para exemplificar a metodologia no texto, que compreende um campo contendo quatro aglomerados simulados com diferentes características, conforme ilustrado na Figura 4.1.

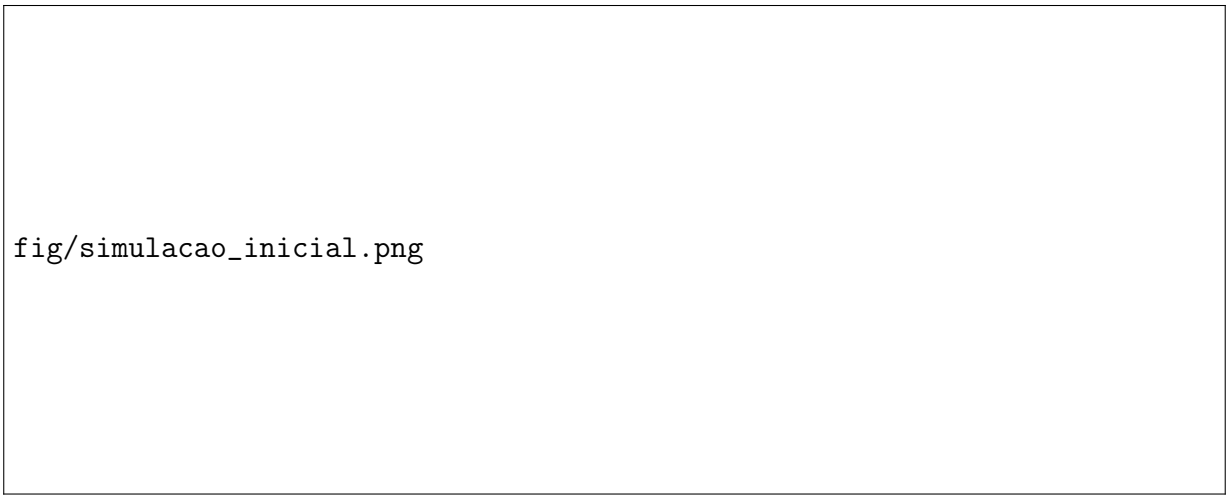

Figura 4.1: Distribuição dos aglomerados presentes na Simulação 01. (a) Distribuição angular (b) Distribuição de movimento próprio.

O objetivo desta Seção é estabelecer uma metodologia de seleção de aglomerados que possa ser aplicada a regiões reais, nas quais o conjunto de membros verdadeiros não está disponível, diferentemente de campos simulados. A metodologia descrita se desenvolve em duas etapas complementares. Primeiro, em ambiente controlado (campos simulados),

utilizamos o conhecimento dos membros dos aglomerados simulados para caracterizar as regiões de detectabilidade no mapa paramétrico. Em seguida, mostramos que a mesma assinatura de detecção pode ser obtida sem esse conhecimento prévio a partir de quantidades derivadas exclusivamente das saídas do DBSCAN, o que torna a metodologia aplicável a regiões reais.

4.1.1 Validação em ambiente simulado: F1-Score

Enquanto na Seção 2.2 descrevemos como o algoritmo reage a um campo exclusivamente composto por estrelas de campo, nesta Seção investigamos as características dos aglomerados presentes em campos estelares no espaço paramétrico. A densidade intrínseca de cada aglomerado determina em qual região de $(\epsilon, \min_{\mathrm{PTS}})$ ele é detectável (Equação 2.7). Para quantificar essa correspondência no contexto simulado, calculamos para cada combinação o F1-Score, métrica que avalia quão bem um conjunto recuperado corresponde a um conjunto de referência. Ele é definido como a média harmônica entre a precisão P (fração dos detectados que são membros reais) e a revocação R (fração dos membros reais que foram detectados):

$$\mathrm{F1} = \frac{2PR}{P+R}, \quad P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \tag{4.1}$$

onde TP, FP e FN denotam verdadeiros positivos, falsos positivos e falsos negativos. O F1-Score varia entre 0 (recuperação nula) e 1 (recuperação perfeita) e exige o conhecimento prévio dos membros reais de cada aglomerado, restringindo seu uso a campos simulados.

Novamente o espaço paramétrico é varrido em intervalos fixos para todos os campos, com variações $\Delta\epsilon = 0{,}005$ e $\Delta\min_{\mathrm{PTS}} = 1$. Dessa forma, utilizamos o DBSCAN e calculamos as métricas em 17.649 combinações de parâmetros distintas para cada campo.

A Figura 4.2 apresenta os mapas de F1-Score individual para cada aglomerado da Simulação 01. Cada estrutura apresenta uma região distinta e localizada de recuperação elevada no espaço paramétrico, correspondente às combinações em que ela é consistentemente identificada. Estruturas distintas ocupam regiões distintas e parcialmente sobrepostas do espaço paramétrico de acordo com sua densidade intrínseca, o que implica que uma única combinação $(\epsilon, \min_{\mathrm{PTS}})$ é, em geral, insuficiente para recuperar todas as estruturas presentes em um mesmo campo.

Figura 4.2: F1-Score individual para cada aglomerado simulado da Simulação 01. Cada painel revela a região do espaço paramétrico $(\epsilon, \min_{\mathrm{PTS}})$ em que o aglomerado é recuperado. A curva sobreposta delimita a fase de fragmentação, dada pela Equação 2.7.

4.1.2 Persistência paramétrica e critérios de seleção

A observação de que um mesmo aglomerado é recuperado em uma região contínua e ampla do espaço paramétrico motiva uma estratégia baseada em persistência paramétrica. Em vez de selecionar uma combinação única de $(\epsilon, \min_{\mathrm{PTS}})$ para obter todos os aglomerados, calculamos os resultados em todo o espaço de parâmetros. Em seguida, comparamos os aglomerados obtidos em diferentes combinações de parâmetros, por exemplo, um aglomerado A_i de uma combinação $(\epsilon_i, \min_{\mathrm{PTS}i})$ e um aglomerado B_j de uma combinação $(\epsilon_j, \min_{\mathrm{PTS}j})$. Aqueles que compartilham membros suficientes são agrupados por serem considerados representações de uma mesma estrutura física. A sobreposição entre dois aglomerados é medida pela razão de Jaccard (Jaccard, 1901):

$$J = \frac{|A_i \cap B_j|}{\min(|A_i|,\, |B_j|)}, \tag{4.2}$$

onde $|A_i|$ e $|B_j|$ são os números de membros de cada aglomerado e $|A_i \cap B_j|$ é o número de estrelas em comum. Pares de aglomerados com $J \geq 0{,}50$ são considerados representações do mesmo objeto físico. Ao usar o mínimo dos dois tamanhos no denominador, a razão de Jaccard agrupa aglomerados compatíveis mesmo que o número de membros seja variável em diferentes combinações.

Após o agrupamento, cada grupo está associado a um conjunto de combinações $(\epsilon, \min_{\mathrm{PTS}})$ no qual ele é detectável. Consideramos apenas os grupos presentes em pelo menos 1% do total de combinações (ou seja, que apareçam em pelo menos 176 combinações), descartando detecções esparsas associadas a flutuações estatísticas pontuais do campo. A Figura 4.3 mostra os mapas de pertinência paramétrica dos grupos obtidos na Simulação 01. Os grupos (a)-(d) apresentam regiões de detectabilidade amplas e estáveis ao longo do espaço paramétrico, compatíveis com o comportamento paramétrico de estruturas físicas reais (Figura 4.2), enquanto os grupos (e)-(f) têm sua pertinência concentrada em faixas estreitas ao longo da curva de fragmentação, sugerindo que são artefatos gerados por sobredensidades do campo estelar.

Para avaliar em que medida os grupos detectados correspondem a estruturas já conhecidas, comparamos cada grupo com os catálogos de aglomerados abertos de Cantat-Gaudin e Anders (2020), Dias et al. (2021) e Hunt e Reffert (2023). A comparação é feita por meio de uma distância entre os centroides normalizada no espaço de cinco dimensões:

$$d = \sqrt{\sum_k \left(\frac{\mu_k^{\mathrm{grupo}} - \mu_k^{\mathrm{alvo}}}{\sigma_k^{\mathrm{grupo}}} \right)^2}, \tag{4.3}$$

onde μ_k^{grupo} e $\sigma_k^{\mathrm{grupo}}$ são o centroide e a dispersão do grupo no parâmetro k, e μ_k^{alvo} é o centroide do aglomerado-alvo. Adotamos $d \leq 5$ como limiar de correspondência. Assim grupos que satisfazem esse critério em relação a um aglomerado simulado são classificados como aglomerados detectados, os grupos que satisfazem o critério em relação a objetos de algum catálogo são classificados como aglomerados reais e os demais são classificados como candidatos a novos aglomerados.

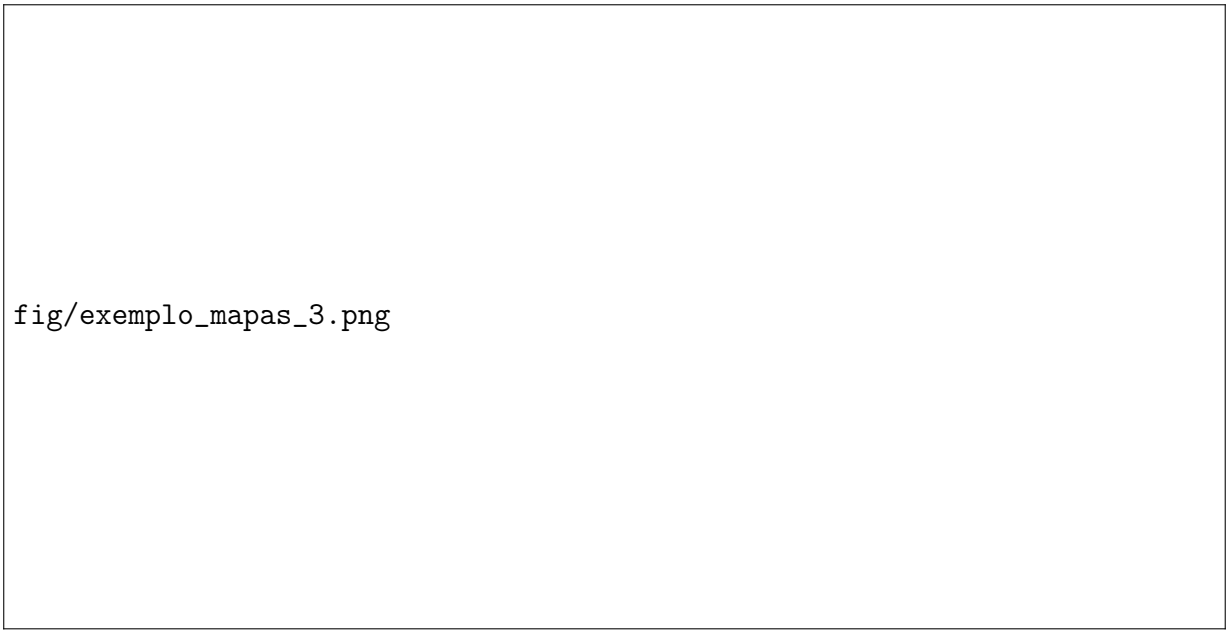

Figura 4.3: Mapas de pertinência paramétrica para os grupos obtidos na Simulação 01. Cada painel mostra as combinações $(\epsilon, \min_{\mathrm{PTS}})$ em que o grupo está presente (região colorida), com a curva de fragmentação sobreposta (sólido preto). (a)–(d) Grupos com pertinência ampla e estável, compatíveis com aglomerados físicos reais. (e)–(f) Grupos cuja pertinência se concentra ao longo da curva de fragmentação, indicando artefatos gerados por sobredensidades do campo estelar.

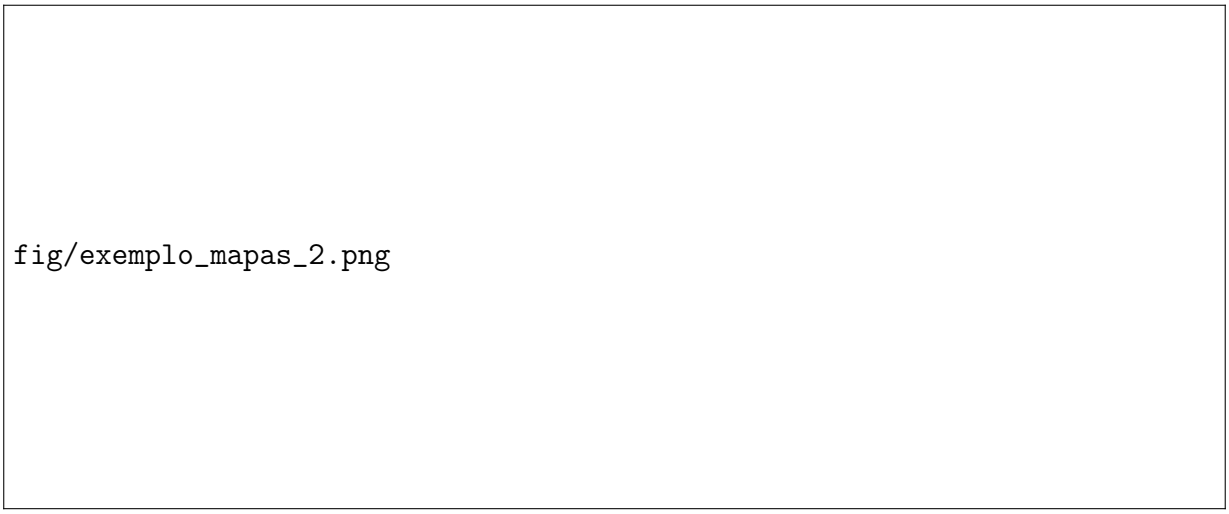

Figura 4.4: Painel de diagnóstico para a Simulação 01. (a) Número de aglomerados detectados (escala logarítmica). (b) Número de membros do maior aglomerado com limite máximo de 2000 membros. A curva sobreposta mostra a região de fragmentação.

4.2 Classificação por XGBoost

Os critérios de persistência paramétrica apresentados na Seção 4.1 identificam grupos candidatos a aglomerados, mas nem toda região persistente no espaço paramétrico $(\epsilon, \mathrm{min_{PTS}})$ corresponde a uma estrutura física real. Como demonstrado na Seção 2.2, são detectados diversos grupos originados por sobredensidades aleatórias do campo e que também satisfazem o critério de pertinência mínima, especialmente próximo à curva de fragmentação (Figura 4.4a). É necessário, portanto, um algoritmo capaz de distinguir aglomerados reais e espúrios a partir das propriedades de cada grupo no espaço paramétrico. Para isso, utilizamos um classificador *XGBoost* (*eXtreme Gradient Boosting*; Chen e Guestrin, 2016).

O *XGBoost* é um algoritmo de aprendizado de máquina baseado em *boosting* de árvores de decisão. A partir de um conjunto de grupos cuja natureza já é conhecida (como campos simulados), descritos por características numéricas, o algoritmo aprende quais dessas características conseguem distinguir uma categoria da outra. Internamente, o *XGBoost* constrói uma sequência de árvores de decisão em que cada árvore corrige os erros das anteriores, resultando em um modelo capaz de separar as duas classes mesmo quando essa separação depende de combinações não triviais de diversas características ao mesmo tempo.

Dessa forma, para cada campo simulado de treinamento, extraímos 39 características numéricas dos grupos detectados. Entre os parâmetros extraídos, destacamos a seguir as características mais relevantes organizadas em três conjuntos. O primeiro descreve a geometria da região de detectabilidade no mapa paramétrico, como a área total, a espessura em ϵ para cada valor de $\mathrm{min_{PTS}}$, a espessura máxima, a compacidade e a excentricidade da região. O segundo descreve as propriedades intrínsecas dos grupos, como o tamanho médio e máximo, a densidade volumétrica no espaço de cinco dimensões e a pertinência percentual. O terceiro descreve o ambiente estelar local, com a densidade de fundo do campo e as dispersões em posição, movimento próprio e paralaxe, que caracterizam o contraste entre o grupo e o ruído de fundo da região.

4.2.1 Conjunto de treinamento

O classificador foi treinado em um conjunto rotulado de grupos extraídos diretamente dos grupos obtidos conforme descrito na Seção 4.1, para os quais a natureza física de cada

candidato é conhecida. Cada grupo recebe um rótulo binário ($y = 1$ para aglomerado físico, $y = 0$ para grupo espúrio). Consideramos como físicos tanto os grupos com correspondência a aglomerados simulados injetados quanto aqueles compatíveis com algum aglomerado já catalogado na região, exigindo em ambos os casos uma distância normalizada $d \leq 5$ (Equação 4.3) em relação ao alvo de referência. O conjunto resultante reúne 185 grupos candidatos extraídos de 50 campos simulados, nos quais 121 foram rotulados como reais (96 correspondendo a aglomerados injetados, 16 ao catálogo de Hunt e Reffert 2023, 6 ao de Cantat-Gaudin e Anders 2020 e 3 ao de Dias et al. 2021) e 64 como espúrios.

4.2.2 Validação e resultados

O classificador foi treinado sobre o conjunto rotulado descrito e, em seguida, avaliado sobre um conjunto de teste completamente independente, composto por outros 50 campos reais com tamanho de 7,55 graus quadrados. Esses campos são distintos dos 50 campos simulados utilizados no treinamento e não contêm aglomerados simulados injetados e os grupos detectados foram comparados diretamente com os catálogos de referência (Cantat-Gaudin e Anders, 2020; Dias et al., 2021; Hunt e Reffert, 2023), de modo que o desempenho medido reflete a capacidade do modelo de recuperar aglomerados reais em regiões não conhecidas durante o treinamento.

Sobre o conjunto de validação, o modelo obteve um F1-Score de 0,819, com precisão de 0,824 e revocação de 0,813. O valor obtido para a precisão indica que 82,4% das estruturas que o classificador identificou como reais de fato correspondiam a aglomerados catalogados. Por sua vez, a revocação de 0,813 reflete que o algoritmo foi capaz de recuperar com sucesso 81,3% de todos os aglomerados reais presentes nos campos de teste, minimizando a perda de objetos reais como falsos negativos. A matriz de confusão correspondente (Figura 4.5) registra 61 grupos corretamente classificados como reais (TP), 606 como não reais (TN), 13 falsos positivos (FP) e 14 falsos negativos (FN). O desempenho, comparável ao obtido durante o ajuste, indica que o classificador generaliza bem para campos fora do regime de treinamento, sem sinais relevantes de sobreajuste.

Cabe ressaltar que os 13 grupos classificados como falsos positivos (FP) não formam uma população homogênea. Como o classificador *XGBoost* atribui a cada grupo uma probabilidade contínua P_{real} de ser um aglomerado físico, é possível adotar um critério de probabilidade para determinar os candidatos a novos aglomerados. Grupos com $P_{\text{real}} >$

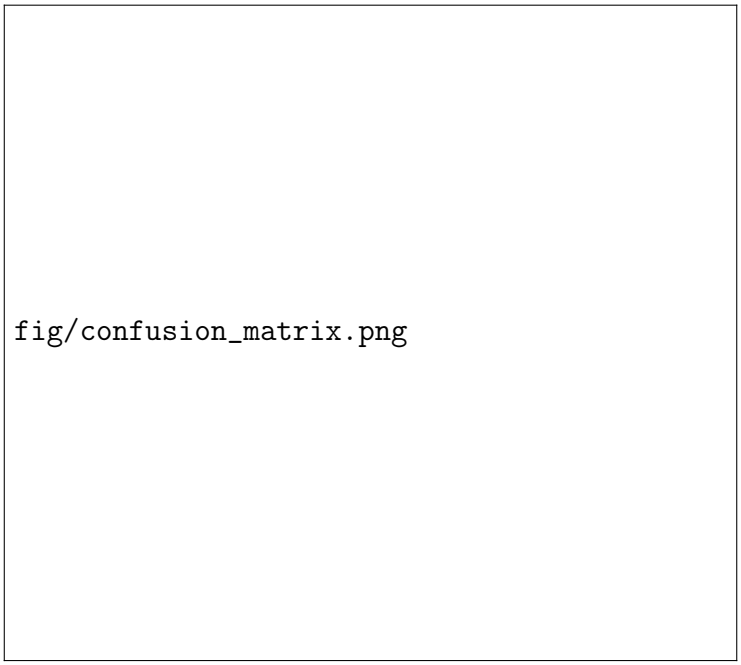

Figura 4.5: Matriz de confusão obtida sobre o conjunto de validação. TP: aglomerados reais corretamente classificados; TN: grupos não reais corretamente rejeitados; FP: grupos não reais classificados como reais; FN: aglomerados reais classificados como não reais.

0,90, mesmo sem correspondência em nenhum dos catálogos consultados, apresentam uma geometria de detectabilidade no espaço paramétrico tão compatível com a de aglomerados físicos reais que sua rejeição como espúrios é pouco provável e por isso são tratados como candidatos a aglomerados não catalogados na literatura. Por outro lado, os grupos com probabilidade entre 0,50 e 0,90, embora tenham cruzado o limiar de classificação como reais, apresentam menor robustez estatística e maior compatibilidade com flutuações do campo estelar próximas à fronteira de decisão do classificador e são considerados falsos positivos. Assim, os grupos com $P_{\mathrm{real}} > 0{,}90$ deixam de ser tratados como falsos positivos e passam a ser considerados verdadeiros positivos ainda não catalogados na literatura, enquanto os demais permanecem classificados como falsos positivos.

A análise das características mais relevantes para a classificação segundo a importância de ganho do modelo mostra que o perímetro normalizado da região de detectabilidade emerge como a variável mais discriminativa, seguida pela espessura média em ϵ e pela excentricidade da região. Essas características descrevem coletivamente a forma da região de detectabilidade no espaço paramétrico, ou seja, aglomerados reais tendem a ocupar regiões largas e regulares, que se mantêm estáveis ao longo de um amplo intervalo de densidades de busca. Em contraste, grupos espúrios produzem regiões estreitas e irregulares, con-

centradas próximo à curva de fragmentação e que desaparecem rapidamente ao variar os parâmetros.

Para os grupos classificados como reais, determinamos a lista final de membros a partir de um critério de persistência paramétrica, considerando como membro do aglomerado apenas os objetos presentes em mais de 50% das combinações de parâmetros em que a estrutura foi detectada. Aplicado à Simulação 01, o classificador *XGBoost* recuperou corretamente os grupos físicos e rejeitou os grupos espúrios identificados na Figura 4.3. A Figura 4.6 compara a distribuição angular dos aglomerados esperados (simulados e catalogados) com os grupos classificados como reais, evidenciando a eficiência do método ao recuperar todos os grupos nesse campo.

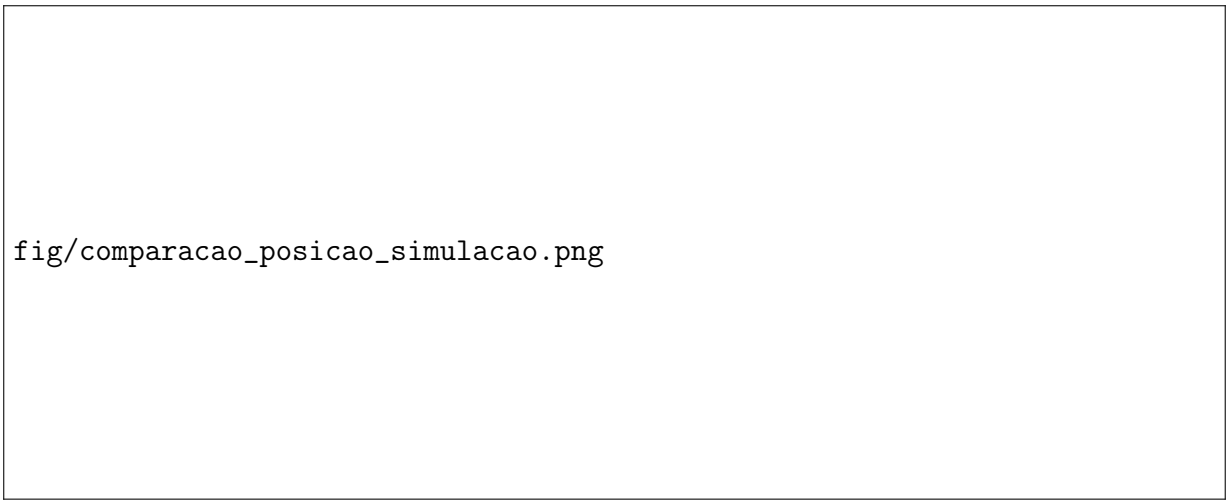

Figura 4.6: (a) Aglomerados detectados (simulados e dos catálogos) para a Simulação 01. (b) Grupos classificados como reais pelo *XGBoost* e membros selecionados com $P_{\mathrm{mem}} \geq 0{,}50$, no espaço dos parâmetros astrométricos.

Além dos campos de teste, aplicamos a metodologia à associação estelar Canis Major OB1, uma região de formação estelar com populações de diferentes idades e múltiplos aglomerados conhecidos. Como mostrado no Apêndice A, recuperamos a maioria dos aglomerados previamente catalogados na região e identificamos dois novos candidatos, o que evidencia a capacidade do método para descobertas em regiões já investigadas na literatura.

Capítulo 5

Conclusões e Perspectivas Futuras

Neste trabalho desenvolvemos e validamos uma metodologia completa para a detecção de aglomerados estelares abertos por meio de dados *Gaia*, combinando a varredura sistemática do espaço paramétrico do DBSCAN e a classificação final entre grupos reais e espúrios por meio de um classificador *XGBoost*. Os resultados mostram que a metodologia recupera de forma confiável os aglomerados presentes nos campos analisados, tanto os injetados nas simulações quanto os já catalogados, mantendo um desempenho elevado quando aplicada a campos reais independentes.

Podemos comparar os resultados obtidos com Hunt e Reffert (2021), que compararam diferentes algoritmos de agrupamento aplicados a dados *Gaia* DR2, com o DBSCAN mostrando-se sensível a cerca de 62% dos aglomerados reais da amostra quando utilizado com uma única combinação de parâmetros para todos os campos, ainda que com precisão e especificidade muito altas. É importante ressaltar, contudo, que a análise feita por Hunt e Reffert (2021) foi conduzida sobre um cenário consideravelmente mais amplo, cobrindo praticamente todo o disco galáctico ($|b| < 25°$) e abrangendo aglomerados a distâncias de até 8,4 kpc, enquanto buscamos objetos com distância entre 0,5 e 5 kpc; a comparação é portanto indicativa, não direta. Ainda assim, a incorporação da persistência paramétrica e do classificador *XGBoost* representa uma melhora expressiva na recuperação de aglomerados em relação ao DBSCAN aplicado isoladamente, obtendo 81,3% dos aglomerados reais da amostra detectados e sugerindo que a metodologia adotada explora informações adicionais do espaço de parâmetros.

Além da comparação com o DBSCAN de parâmetros fixos, é relevante contextualizar nossos resultados frente a outros algoritmos de agrupamento aplicados a dados *Gaia*. Hunt e Reffert (2021) também avaliaram o desempenho do HDBSCAN (*Hierarchical DBSCAN*;

Campello et al., 2013) na mesma amostra, encontrando uma sensibilidade superior à do DBSCAN, de até 82%, ao custo de uma precisão e especificidade sensivelmente menores devido ao maior número de falsos positivos gerados pelo algoritmo. Nesse sentido, a metodologia aqui desenvolvida obtém uma sensibilidade comparável à do HDBSCAN (81,3%), mantendo simultaneamente uma precisão elevada (82,4%). Isso indica que a metodologia utilizada é capaz de reunir as vantagens de ambos os algoritmos sem dimuir a precisão observada no HDBSCAN.

A análise dos erros do classificador revela padrões com interpretação física direta e aponta direções naturais para continuidade do trabalho. Os falsos positivos correspondem a grupos persistentes no espaço paramétrico e com geometria compatível com a de aglomerados reais, mas que não foram associados a nenhum dos catálogos de referência utilizados. Esses objetos podem ser aglomerados reais presentes em catálogos não considerados neste trabalho ou candidatos genuinamente novos, ainda não reportados na literatura, e um possível estudo posterior seria confirmar se apresentam coerência física e atendem aos critérios gerais de aglomerados abertos.

Um exemplo disso é a aplicação à associação Canis Major OB1 (Apêndice A), uma região de formação estelar complexa, com populações de diferentes idades e densidades intrínsecas variadas, na qual o método identificou dois candidatos sem correspondência nos catálogos consultados.

Cabe destacar que o desempenho do classificador *XGBoost* está intrinsecamente ligado ao domínio de validade definido pelas distribuições físicas adotadas na simulação dos aglomerados sintéticos (Tabela 3.1). Aglomerados reais cujas propriedades se afastam significativamente desse domínio de treinamento, como estruturas pouco densas, aglomerados extremamente jovens ou em regiões de alta extinção podem não ser adequadamente representados pelo classificador, resultando em desempenho reduzido. Essa limitação foi observada na aplicação à associação Canis Major OB1 (Apêndice A), na qual os aglomerados CMa05 e CMa07, por serem pequenos e próximos do limite inferior de massa representado nas simulações, não foram recuperados pela metodologia. Assim, futuros trabalhos devem buscar estender o conjunto de treinamento para cobrir esses regimes físicos.

Como perspectivas de continuidade, buscamos aumentar o conjunto de características do classificador com informações como distância, extinção e idade, além de dados fotométricos do próprio *Gaia*. A disponibilização do *Gaia* DR4 também abre a possibilidade

de reavaliar os campos analisados com astrometria de maior precisão, permitindo comparações diretas com os resultados obtidos neste trabalho. Por fim, projetamos a expansão da metodologia para todo o disco galáctico. Essa varredura sistemática viabilizará uma comparação direta com os catálogos atuais de aglomerados abertos e a identificação de novos candidatos ainda não reportados na literatura.

Referências Bibliográficas

Astropy Collaboration et al., Astropy: A community Python package for astronomy, Astronomy & Astrophysics, 2013, vol. 558, p. A33

Bak P., Tang C., Wiesenfeld K., Self-organized criticality: An explanation of the 1/f noise, Phys. Rev. Lett., 1987, vol. 59, p. 381

Bressan A., Marigo P., Girardi L., Salasnich B., Dal Cero C., Rubele S., Nanni A., PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code, MNRAS, 2012, vol. 427, p. 127

Campello R. J. G. B., Moulavi D., Sander J., Density-Based Clustering Based on Hierarchical Density Estimates. In Advances in Knowledge Discovery and Data Mining , Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, p. 160

Cantat-Gaudin T., Anders F., Clusters and mirages: cataloguing stellar aggregates in the Milky Way, A&A, 2020, vol. 633, p. A99

Castro-Ginard A., Jordi C., Luri X., Cantat-Gaudin T., Balaguer-Núñez L., Hunting for open clusters in Gaia DR2: the Galactic anticentre, A&A, 2019, vol. 627, p. A35

Chen T., Guestrin C., XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794

Dias W. S., Monteiro H., Moitinho A., Lépine J. R. D., Carraro G., Paunzen E., Alessi B., Villela L., Updated parameters of 1743 open clusters based on Gaia DR2, Monthly Notices of the Royal Astronomical Society, 2021, vol. 504, p. 356

Edenhofer G., Zucker C., Frank P., Saydjari A. K., Speagle J. S., Finkbeiner D., Enßlin T. A., A parsec-scale Galactic 3D dust map out to 1.25 kpc from the Sun, A&A, 2024, vol. 685, p. A82

Ester M., Kriegel H.-P., Sander J., Xu X., A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining , KDD'96, AAAI Press, 1996, p. 226–231

Fabricius C., Luri X., Arenou F., Babusiaux C., Helmi A., Muraveva T., Reylé C., Spoto F., Vallenari A., Antoja T., Balbinot E., Barache C., Bauchet N., Bragaglia A., Busonero D., Cantat-Gaudin T., Carrasco J. M., Diakité S., Fabrizio M., Figueras F., Garcia-Gutierrez A., Garofalo A., Jordi C., Kervella P., Khanna S., Leclerc N., Licata E., Lambert S., Marrese P. M., Masip A., Ramos P., Robichon N., Robin A. C., Romero-Gómez M., Rubele S., Weiler M., Gaia Early Data Release 3. Catalogue validation, A&A, 2021, vol. 649, p. A5

Fernandes B., Montmerle T., Santos-Silva T., Gregorio-Hetem J., Runaways and shells around the CMa OB1 association, A&A, 2019, vol. 628, p. A44

Gaia Collaboration Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Bailer-Jones C. A. L., Biermann M., Evans D. W., Eyer L., Jansen F., Jordi C., Klioner S. A., Lammers U., Lindegren L., Luri X., Mignard F., Panem C., Gaia Data Release 2. Summary of the contents and survey properties, A&A, 2018, vol. 616, p. A1

Gaia Collaboration Prusti T., de Bruijne J. H. J., Brown A. G. A., Vallenari A., Babusiaux C., Bailer-Jones C. A. L., Bastian U., Biermann M., Evans D. W., et al. The Gaia mission, A&A, 2016, vol. 595, p. A1

Gaia Collaboration Vallenari A., Brown A. G. A., Prusti T., de Bruijne J. H. J., Arenou F., Babusiaux C., Biermann M., Creevey O. L., Ducourant C., Evans Gaia Data Release 3. Summary of the content and survey properties, A&A, 2023, vol. 674, p. A1

Gorski K. M., Wandelt B. D., Hansen F. K., Hivon E., Banday A. J., , 1999 The HEALPix Primer

Hunt E. L., Cantat-Gaudin T., Anders F., Malhotra S., Spina L., Castro-Ginard A., Cavallo L., The selection function of the Gaia DR3 open cluster census, A&A, 2026, vol. 706, p. A341

Hunt E. L., Cantat-Gaudin T., Anders F., Spina L., Cavallo L., Castro-Ginard A., Belokurov V., Brown A. G. A., Casey A. R., Drimmel R., Fouesneau M., Reffert S., The completeness of the open cluster census towards the Galactic anticentre, A&A, 2025, vol. 699, p. A273

Hunt E. L., Reffert S., Improving the open cluster census - I. Comparison of clustering algorithms applied to Gaia DR2 data, A&A, 2021, vol. 646, p. A104

Hunt E. L., Reffert S., Improving the open cluster census. II. An all-sky cluster catalogue with Gaia DR3, A&A, 2023, vol. 673, p. A114

Jaccard P., Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bulletin de la Société vaudoise des sciences naturelles, 1901, vol. 37, p. 547

Kharchenko N. V., Piskunov A. E., Schilbach E., Röser S., Scholz R.-D., Global survey of star clusters in the Milky Way. II. The catalogue of basic parameters, A&A, 2013, vol. 558, p. A53

King I., The structure of star clusters. I. an empirical density law, AJ, 1962, vol. 67, p. 471

Kroupa P., On the variation of the initial mass function, MNRAS, 2001, vol. 322, p. 231

Krumholz M. R., McKee C. F., Bland-Hawthorn J., Star Clusters Across Cosmic Time, ARA&A, 2019, vol. 57, p. 227

Lamers H. J. G. L. M., Gieles M., Bastian N., Baumgardt H., Kharchenko N. V., Portegies Zwart S., An analytical description of the disruption of star clusters in tidal fields with an application to Galactic open clusters, A&A, 2005, vol. 441, p. 117

Moe M., Di Stefano R., Mind Your Ps and Qs: The Interrelation between Period (P) and Mass-ratio (Q) Distributions of Binary Stars, ApJS, 2017, vol. 230, p. 15

Portegies Zwart S. F., McMillan S. L. W., Gieles M., Young Massive Star Clusters, ARA&A, 2010, vol. 48, p. 431

Santos-Silva T., Perottoni H. D., Almeida-Fernandes F., Gregorio-Hetem J., Jatenco-Pereira V., Mendes de Oliveira C., Montmerle T., Bica E., Bonatto C., Monteiro H., Dias W. S., Barbosa C. E., Fernandes B., Galli P. A. B., Borges Fernandes M., Kanaan A., Ribeiro T., Schoenell W., Canis Major OB1 stellar group contents revealed by Gaia, MNRAS, 2021, vol. 508, p. 1033

Zucker C., Speagle J. S., Schlafly E. F., Green G. M., Finkbeiner D. P., Goodman A. A., Alves J., A Large Catalog of Accurate Distances to Local Molecular Clouds: The Gaia DR2 Edition, The Astrophysical Journal, 2019, vol. 879, p. 125

Apêndice

Apêndice A

Aplicação do algoritmo na Associação Estelar Canis Major

A.1 A região de Canis Major OB1

A associação estelar Canis Major OB1 (CMa OB1) é uma região de formação estelar da Via Láctea, localizada a aproximadamente 1 kpc (Zucker et al., 2019). A região contém um grande número de objetos jovens com idades entre 5 e 100 milhões de anos, incluindo mais de 20 aglomerados conhecidos em sua linha de visada, dos quais nem todos estão fisicamente associados à região.

A história de formação estelar da CMa OB1 está associada a sucessivas explosões de supernovas há cerca de 1, 2 e 6 Myr (Fernandes et al., 2019). Essa coexistência de populações de diferentes idades torna a região interessante para o teste do algoritmo, pois diferentes grupos apresentam densidades intrínsecas, dispersões cinemáticas e contrastes em relação ao campo bastante distintos.

Santos-Silva et al. (2021) aplicaram o algoritmo HDBSCAN a dados do *Gaia* DR2 (Gaia Collaboration et al., 2018) sobre uma região de 4,1° de raio centrada em $(\alpha, \delta) = (106{,}7°, -10{,}6°)$, identificando 29 grupos estelares, dos quais 14 com distâncias entre 570 e 1650 pc foram caracterizados como possíveis membros da associação. Desses 14, dez foram identificados como aglomerados já conhecidos na literatura (incluindo Alessi 21, Gulliver 13, vdB 92, NGC 2318, NGC 2323, NGC 2335, NGC 2343, NGC 2353, FOF 2216 e FOF 2302) e cinco foram propostos como novos candidatos a aglomerados (CMa01, CMa02, CMa05, CMa08 e CMa23).

Aplicamos nossa metodologia utilizada na mesma região analisada por Santos-Silva et al. (2021), utilizando dados do *Gaia* DR3 sob critérios de seleção astrométrica e fo-

tométrica equivalentes. A varredura sistemática do espaço paramétrico $(\epsilon, \min_{\mathrm{PTS}})$ foi conduzida nos mesmos intervalos e os grupos detectados foram submetidos ao classificador *XGBoost* previamente treinado nos campos simulados. Para a validação, comparamos os grupos classificados como reais com os 14 grupos confiáveis reportados por Santos-Silva et al. (2021) e com os catálogos externos de Cantat-Gaudin e Anders (2020), Dias et al. (2021) e Hunt e Reffert (2023).

A Figura A.1 apresenta a distribuição espacial, de movimento próprio e de paralaxe dos grupos classificados como reais pela metodologia. A Tabela A.1 apresenta os parâmetros obtidos para os aglomerados detectados, indicando a correspondência com aglomerados conhecidos. No total, o método identificou 21 grupos persistentes classificados como reais, dos quais 19 apresentam correspondência direta com objetos catalogados e 2 são candidatos a novos aglomerados, que não têm correspondência com nenhum dos catálogos consultados.

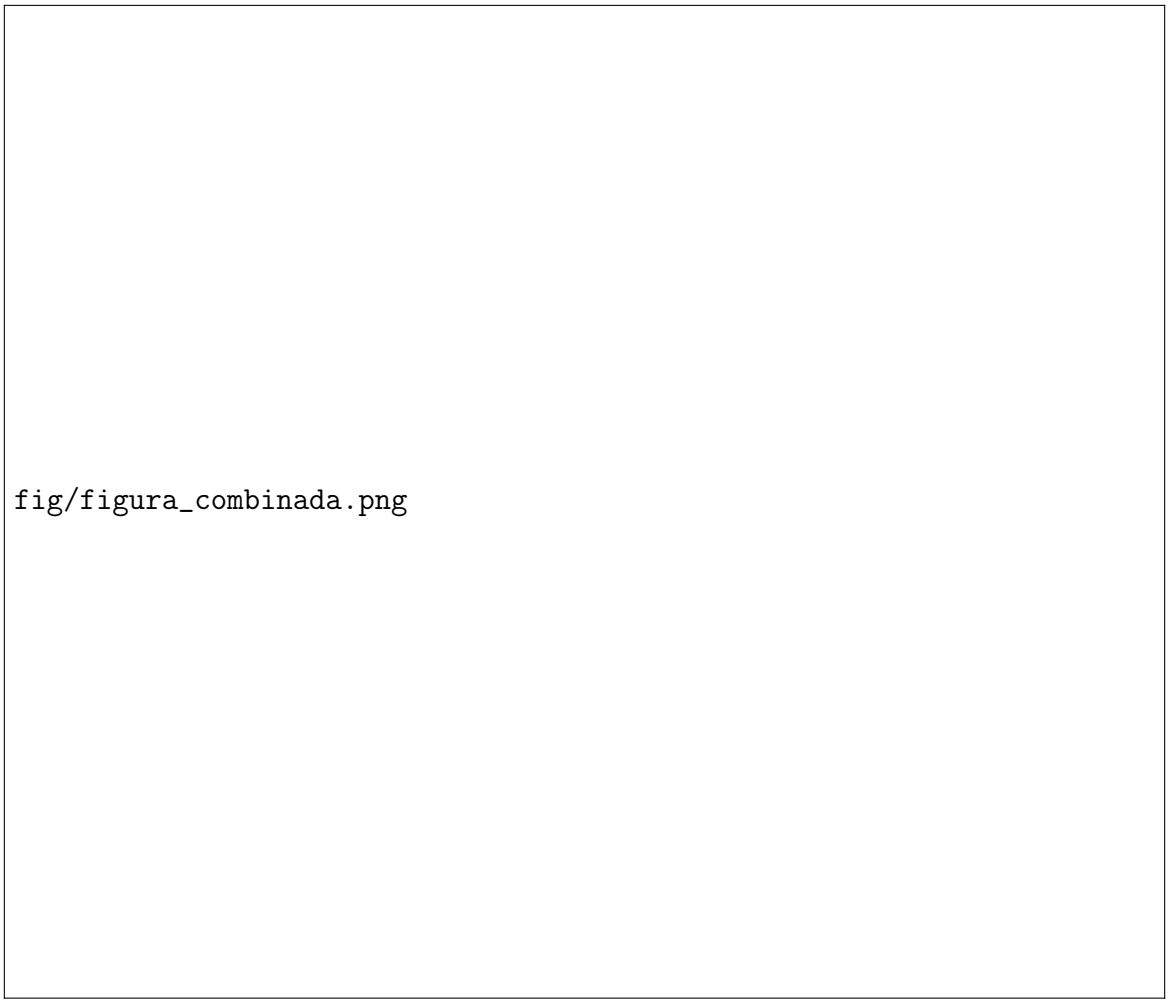

Figura A.1: Distribuição dos aglomerados obtidos pela metodologia na região de CMa. (a) Distribuição angular. (b) Distribuição de movimento próprio. (c) Histogramas violino da paralaxe para cada grupo.

Como resultado, obtivemos uma precisão de 0,87, revocação de 0,65 e F1-score de 0,74. Comparando especificamente com o conjunto de 14 grupos confiáveis obtidos por Santos-Silva et al. (2021), nossa metodologia recuperou 10 deles (CMa01, CMa03, CMa06, CMa08, CMa09, CMa13, CMa15, CMa17, CMa18 e CMa19). Os grupos não recuperados (CMa02, CMa05, CMa07 e CMa23) apresentam, em sua maioria, características que justificam a dificuldade de detecção.

Os CMa05 e CMa07 são grupos pequenos (28 e 26 membros, respectivamente, segundo Santos-Silva et al. 2021), próximos do limite inferior estabelecido pelo nosso classificador, treinado predominantemente com aglomerados de massa intermediária. Já CMa02 e CMa23 apresentam baixa densidade intrínseca e dispersões cinemáticas relativamente amplas, ocupando regiões estreitas do mapa paramétrico próximas à curva de fragmentação, levando à classificação como não real. Por outro lado, a metodologia recuperou estruturas adicionais não presentes no conjunto principal de Santos-Silva et al. (2021), incluindo aglomerados como FSR 1163, HXHWL 22, UBC 444, PHOC 5, Theia 2289 e OC 0381, identificados em catálogos mais recentes baseados em *Gaia* DR3.

A revocação inferior à observada nos campos simulados é esperada e reflete a complexidade real da associação CMa OB1, que contém grupos de baixíssimo contraste cinemático, populações sobrepostas no espaço e estruturas em processo de dissolução, características não totalmente representadas pelos aglomerados sintéticos utilizados no treinamento. Já a precisão elevada confirma que, mesmo em um ambiente desafiador, o classificador *XGBoost* mantém alta especificidade, rejeitando consistentemente artefatos de fragmentação.

A detecção de aglomerados em uma ampla faixa de paralaxes (entre $\varpi \approx 0{,}44$ e 1,12 mas) evidencia que a varredura sistemática do espaço paramétrico permite identificar estruturas em diferentes distâncias e densidades intrínsecas. Esta característica é particularmente relevante em regiões como CMa OB1, onde populações sobrepostas em linha de visada são comuns, e onde uma abordagem com parâmetros fixos do DBSCAN tende a privilegiar apenas uma faixa estreita de densidades.

Os dois candidatos novos (Candidato 01 e Candidato 02) apresentam persistência paramétrica compatível com aglomerados físicos, mas não possuem correspondência em nenhum dos catálogos consultados. Sua confirmação como aglomerados reais exigirá análises posteriores, em particular a caracterização do diagrama cor-magnitude, que permitirá verificar a coerência da população estelar com isócronas e estimar parâmetros físicos como idade e

Tabela A.1 - Parâmetros dos aglomerados CMa detectados pela metodologia.

Nome	P_{real}	α (graus)	δ (graus)	μ_{α^*} (mas/yr)	μ_δ (mas/yr)	ϖ (mas)	N_{det} (stars)	**Outro nome**
CMa01^{HS}	0,671	106,817	-7,678	-3,92	1,18	0,916	87	Theia 2649
CMa03^{CDHS}	0,791	104,934	-13,247	-2,94	0,31	0,627	116	Gulliver 13
CMa04^{S}	0,961	102,991	-9,969	-1,29	1,26	0,495	134	-
CMa06^{CDHS}	0,966	106,183	-11,416	-4,22	1,54	0,866	289	VdB 92
CMa08^{H}	0,702	107,237	-12,446	-3,19	0,84	0,853	151	-
CMa09^{CDHS}	0,960	104,889	-13,744	-1,78	-2,15	0,718	194	NGC 2318
CMa10^{S}	0,926	105,475	-13,527	-0,99	-1,42	0,480	120	
CMa13^{S}	0,959	107,264	-7,206	-0,87	-1,21	1,122	102	FOF 2302
CMa15^{CHDS}	0,928	108,670	-10,212	-1,07	0,79	0,834	252	NGC 2353
CMa17^{CHDS}	0,970	105,679	-8,361	-0,73	-0,82	0,991	839	NGC 2323
CMa18^{CHDS}	0,976	107,000	-10,638	0,22	-0,18	0,893	208	NGC 2343
CMa19^{CHDS}	0,945	106,860	-9,918	-0,77	-0,59	0,578	191	NGC 2335
CMa20^{S}	0,510	106,589	-9,301	-1,25	0,86	0,456	356	-
Candidato 01	0,919	106,917	-13,334	-1,32	0,80	0,444	26	-
Candidato 02	0,937	105,556	-8,378	-0,35	-0,45	0,999	125	-
FSR 1163^{CDH}	0,796	104,842	-9,077	0,16	1,25	0,460	63	-
HXHWL 22^{H}	0,798	108,980	-9,410	-0,59	0,95	0,868	102	-
OC 0381^{H}	0,602	105,229	-14,308	-1,44	1,66	0,627	29	-
PHOC 5^{H}	0,887	105,877	-7,801	-0,64	-1,04	0,792	62	-
Theia 2289^{H}	0,830	106,748	-12,372	-1,03	-0,25	0,609	132	-
UBC 444^{H}	0,770	105,472	-7,110	-1,50	-2,44	0,878	82	-

Notas: P_{real} é a probabilidade do grupo ser real atribuída pelo classificador *XGBoost* e N_{det} é o número de membros detectados pelo algoritmo. Clusters encontrados em $^{(C)}$ Cantat-Gaudin e Anders (2020); $^{(D)}$ Dias et al. (2021); $^{(H)}$ Hunt e Reffert (2023); $^{(S)}$ Santos-Silva et al. (2021).

extinção.

De forma geral, a aplicação a CMa OB1 valida a metodologia em um cenário observacional real, complementando os testes em campos simulados apresentados no Capítulo 4. A combinação de alta precisão com revocação média indica que o método é confiável na identificação positiva de aglomerados, ainda que conservador na recuperação de estruturas de baixo contraste. Assim, o método se mostra desejável em aplicações de larga escala, em que minimizar falsos positivos é prioritário para a construção de catálogos confiáveis a partir dos dados do *Gaia*.