Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Comparação de critérios para determinação do número de *clusters*

José Cezário Mariano Junior

Orientador: Professor André Luiz Fernandes Cançado

Brasília

2018

José Cezário Mariano Junior

Comparação de critérios para determinação do número de *clusters*

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Professor André Luiz Fernandes Cançado

Brasília

2018

José Cezário Mariano Junior
Comparação de critérios para determinação do número de *clusters*/ José Cezário Mariano Junior. – Brasília, 2018-
41 p. : il. (algumas color.) ; 30 cm.

Orientador: Professor André Luiz Fernandes Cançado

Relatório Final – Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística
Trabalho de Conclusão de Curso de Graduação, 2018.

1. análise de agrupamento. 2. critérios de inferência do número de grupos.

José Cezário Mariano Junior

Comparação de critérios para determinação do número de *clusters*

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Trabalho aprovado. Brasília, 7 de dezembro de 2018:

Professor André Luiz Fernandes Cançado
Orientador

Professor Antônio Eduardo Gomes
Membro da Banca

Professor Donald Matthew Pianto
Membro da Banca

Brasília
2018

Resumo

Este trabalho apresenta um estudo de critérios de inferência do número correto de grupos em conjuntos de dados, considerando diferentes métodos de agrupamento aplicados a diversas configurações de conjuntos de dados. Foram analisados 22 conjuntos de dados com dimensões e número de grupos variáveis, com grupos gaussianos, elipsoidais e em espiral. Os agrupamentos foram realizados usando o pacote *NbClust* (linguagem R), utilizando os métodos k-médias, Ward, ligação completa e centroide, todos com distância euclidiana, e usando os critérios CH, Silhueta, DB, Hartigan, Tracew, Trcovw e Gap para inferir o número correto de grupos. Os critérios CH, Silhueta e DB apresentaram bons resultados para conjuntos de dados com grupos gaussianos. Os critérios Hartigan, Tracew e Trcovw apresentaram bons resultados apenas para conjuntos com poucos grupos gaussianos de baixa dimensão. O critério Gap não apresentou resultados satisfatórios em nenhuma das análises realizadas. De forma geral, os resultados não foram satisfatórios para conjuntos de dados com grupos de geometria mais complexa ou de dimensões mais elevadas, o que pode ser consequência da simplicidade dos métodos de agrupamento usados.

Palavras-chave: agrupamento, critérios, inferência, grupos, NbClust.

Abstract

This study presents an evaluation of different cluster validity indices, considering different clustering methods applied to data sets with different configurations. The study was conducted on 22 data sets of different dimensions, number of clusters and type of clusters (Gaussian, ellipsoidal and spiral clusters). The clustering process was performed using the *NbClust* R-package using k-means and hierarchical clustering (Ward, complete linkage and centroid) and Euclidean distance, comparing the results from CH, Silhouette, DB, Hartigan, Tracew, Trcovw and Gap validity indices. CH, Silhouette and DB indices were able to find the correct number of clusters in data sets with Gaussian clusters. Hartigan, Tracew and Trcovw were able to correctly find the number of clusters only for low-dimension Gaussian data sets. The Gap index could not find the correct number of clusters in any of the data sets analysed. In general, results were not satisfactory as dimension and geometry of the data sets got higher and more complex, which may be due to the simplicity of the clustering methods applied in the study.

Keywords: clustering, index, inference, clusters, NbClust.

Sumário

1 Introdução

Análise de agrupamento (*clustering*) consiste em uma técnica exploratória para identificação de grupos (*clusters*) homogêneos de objetos em um conjunto de dados (HANDL; KNOWLES, 2007). O conceito de grupo é uma generalização, para dimensões elevadas, da noção humana de grupos (bi ou tridimensionais) como estruturas de objetos que são naturalmente conectados (HANDL; KNOWLES, 2007). Objetos de um mesmo grupo são mais similares entre si do que objetos de outros grupos (CHARRAD et al., 2014). O objetivo da análise de agrupamento é a identificação não supervisionada de grupos em conjuntos de dados complexos, geralmente de difícil interpretação humana (HANDL; KNOWLES, 2007). A análise de agrupamento permite o estabelecimento de hipóteses sobre as relações entre objetos do conjunto de dados, com aplicações em mineração de dados, processamento de imagens e reconhecimento de padrões (JOHNSON; WICHERN, 2007).

O agrupamento é realizado com base em medidas de similaridade ou distância, sem que sejam de fato conhecidos o número correto de grupos ou a estrutura dos grupos (JOHNSON; WICHERN, 2007). Os algoritmos de agrupamento geralmente requerem que o usuário indique o número de grupos a serem considerados (métodos não hierárquicos) ou realizam agrupamentos com número de grupos variando de n (total de objetos no conjunto de dados) até um (todos os objetos em um único grupo) (CHARRAD et al., 2014). Dessa forma, é necessário validar o resultado da análise de agrupamento. Para tanto, diversos critérios de inferência do número de grupos são propostos na literatura. Os critérios geralmente utilizam informações sobre o grau de compactação dentro de cada grupo e o nível de isolamento entre grupos, bem como características geométricas e estatísticas dos dados, o número de objetos e medidas de similaridade ou distância (CHARRAD et al., 2014).

A eficácia dos métodos de agrupamento e dos critérios de validação de agrupamento pode ser afetada pela geometria do conjunto de dados: sobreposição, diferenças no tamanho e na forma dos grupos (HANDL; KNOWLES, 2007).

Assim, torna-se interessante identificar em que tipos de conjuntos de dados os diferentes métodos e critérios de agrupamento são mais eficazes na inferência do número correto de grupos.

1.1 Objetivo Geral

Analisar critérios de inferência do número correto de grupos em um conjunto de dados, considerando diferentes algoritmos de agrupamento aplicados a diversas configurações de conjuntos de dados.

1.2 Objetivos Específicos

- Descrever métodos de agrupamento e identificar critérios de inferência do número correto de grupos em um conjunto de dados; e
- Comparar a eficácia dos critérios descritos na inferência do número correto de grupos em conjuntos de dados sintéticos de diferentes características.

2 Revisão da Literatura

2.1 Medidas de dissimilaridade

As medidas de dissimilaridade apresentadas nas subseções seguintes consideram dois vetores p-dimensionais $\mathbf{x} = [x_1, \ldots, x_p]'$ e $\mathbf{y} = [y_1, \ldots, y_p]'$ resultantes da medição de p características em cada um dos objetos. As medidas de dissimilaridade, ou distâncias, podem ser aplicadas a variáveis quantitativas (discretas ou contínuas) ou qualitativas (nominais ou ordinais), natureza que deve ser levada em consideração na escolha da distância utilizada.

2.1.1 Distância Euclidiana

A distância euclidiana é a distância usual entre dois vetores, descrita pela Equação 2.1.

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{p} (x_j - y_j)^2 \right]^{1/2}. \tag{2.1}$$

2.1.2 Distância Máxima

A distância máxima é a maior diferença entre dois componentes de $\mathbf{x}$ e $\mathbf{y}$, descrita pela Equação 2.2.

$$d(\mathbf{x}, \mathbf{y}) = \sup_{1 \leq j \leq p} |x_j - y_j|. \tag{2.2}$$

2.1.3 Distância de Manhattan

A distância de Manhattan (*city-block*) é a distância absoluta entre dois vetores, dada pela Equação 2.3.

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} |x_j - y_j|. \tag{2.3}$$

2.1.4 Distância de Minkowski

A distância de Minkowski é a norma de ordem m, descrita pela Equação 2.4.

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{p} (x_j - y_j)^m \right]^{1/m}. \tag{2.4}$$

Para $m = 1$, a distância de Minkowski equivale à distância de Manhattan. Para $m = 2$, tem-se a distância euclidiana. De forma geral, o valor de m controla o peso dado às maiores e menores diferenças (JOHNSON; WICHERN, 2007).

2.1.5 Distância de Canberra

A distância de Canberra é descrita pela Equação 2.5.

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} \frac{|x_j - y_j|}{|x_j| + |y_j|}. \tag{2.5}$$

Valores para os quais a distância resulta em uma indeterminação matemática são considerandos como valores faltantes.

2.2 Algoritmos de agrupamento

Os algoritmos de agrupamento são a solução proposta ao problema de identificar de forma não supervisionada o número de grupos em conjuntos de dados complexos, geralmente de difícil interpretação humana. Devido a limitações de tempo e de poder computacional, os algoritmos propostos na literatura buscam por agrupamentos suficientemente bons (de acordo com medidas de dissimilaridade), porém não necessariamente ótimos, uma vez que a solução ideal de selecionar o melhor agrupamento entre todos os possíveis agrupamentos é geralmente inviável.

Os métodos de agrupamento podem ser classificados em hierárquicos e não-hierárquicos.

2.2.1 Métodos Hierárquicos

Os métodos hierárquicos baseiam-se em uma série de aglomerações ou divisões sucessivas, podendo então ser classificados em dois tipos:

- Agrupamento Hierárquico Aglomerativo (HAC): inicialmente, cada objeto é alocado em seu próprio grupo e pares de grupos são unidos à medida em que se sobe na hierarquia de grupos.

- Agrupamento Hierárquico Divisivo: todos os objetos são inicialmente alocados em um único grupo e divisões são efetuadas à medida em que se desce na hierarquia de grupos.

Os métodos hierárquicos requerem a definição de uma medida de similaridade (ou distância), como as apresentadas na Seção 2.1, e de um critério de aglomeração, como os apresentados nas subseções a seguir (métodos HAC).

2.2.1.1 Ligação simples

A distância D_{ij} entre dois grupos C_i e C_j é a menor distância entre dois objetos $\mathbf{x}$ e $\mathbf{y}$, com $\mathbf{x} \in C_i$ e $\mathbf{y} \in C_j$ (CHARRAD et al., 2014):

$$D_{ij} = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y}). \tag{2.6}$$

A ligação simples tende a criar grupos irregulares e alongados, pois alguns grupos podem ser unidos simplesmente porque um único objeto de um grupo está próximo a um único objeto de outro grupo, embora vários outros objetos de cada grupo estejam distantes uns dos outros (efeito de encadeamento).

2.2.1.2 Ligação completa

A distância D_{ij} entre dois grupos C_i e C_j é a maior distância entre dois objetos $\mathbf{x}$ e $\mathbf{y}$, com $\mathbf{x} \in C_i$ e $\mathbf{y} \in C_j$ (CHARRAD et al., 2014):

$$D_{ij} = \max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y}). \tag{2.7}$$

2.2.1.3 Ligação média

A distância D_{ij} entre dois grupos C_i e C_j é a distância média de pares de objetos $\mathbf{x}$ e $\mathbf{y}$, com $\mathbf{x} \in C_i$ e $\mathbf{y} \in C_j$ (CHARRAD et al., 2014):

$$D_{ij} = \sum_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \frac{d(\mathbf{x}, \mathbf{y})}{n_i . n_j}, \tag{2.8}$$

com n_i e n_j iguais ao número de objetos nos grupos C_i e C_j, respectivamente.

2.2.1.4 Mediana

A distância D_{ij} entre dois grupos C_i e C_j é calculada da seguinte forma (CHARRAD et al., 2014):

$$D_{ij} = \frac{D_{ik} + D_{il}}{2} - \frac{D_{kl}}{4}, \tag{2.9}$$

sendo o grupo C_j formado pela agregação dos grupos C_k e C_l.

2.2.1.5 Centroide

A distância D_{ij} entre dois grupos C_i e C_j é dada pelo quadrado da distância euclidiana entre os centroides dos dois grupos, ou seja, entre os vetores média dos dois grupos, $\overline{x_i}$ e $\overline{x_j}$, respectivamente (CHARRAD et al., 2014):

$$D_{ij} = ||\overline{x_i} - \overline{x_j}||^2. \tag{2.10}$$

Esse método é mais robusto a pontos isolados do que os demais métodos apresentados.

2.2.1.6 Método de Ward

O método de Ward minimiza a variância total intra-grupo. A cada passo, o par de grupos de menor distância (soma do erro quadrático) é unido, de forma a proporcionar um aumento mínimo na variância total intra-grupo após a união e a minimizar a perda de informação ao unir os grupos. Este método é tido como um método hierárquico precursor aos métodos não-hierárquicos de agrupamento, pois se baseia na noção de que os grupos de objetos apresentam formato aproximadamente elíptico (JOHNSON; WICHERN, 2007).

2.2.1.7 Método de McQuitty

A distância D_{ij} entre dois grupos C_i e C_j é a média ponderada das distâncias entre grupos (CHARRAD et al., 2014):

$$D_{ij} = \frac{D_{ik} + D_{il}}{2}, \tag{2.11}$$

sendo o grupo C_j formado pela agregação dos grupos C_k e C_l.

2.2.2 Métodos Não-hierárquicos

Os métodos não-hierárquicos são projetados para agrupar os objetos em uma coleção de k grupos, sendo k definido inicialmente ou determinado em tempo de execução pelo algoritmo de agrupamento. Os métodos não-hierárquicos geralmente podem ser aplicados a conjuntos de dados bem maiores dos que os suportados por métodos hierárquicos (JOHNSON; WICHERN, 2007).

2.2.2.1 k-médias

O método k-médias é um algoritmo iterativo que minimiza a soma de quadrados intra-grupo para um número k de grupos. É um dos métodos de agrupamento mais

populares, sendo implementado em uma variedade de aplicativos e pacotes (CHARRAD et al., 2014).

Basicamente, o algoritmo é composto por três passos (JOHNSON; WICHERN, 2007):

1. Particione os n objetos em k grupos iniciais ou especifique k centroides iniciais (sementes);
2. Percorra a lista de objetos, designando cada objeto ao grupo cujo centroide é o mais próximo (distância euclidiana). Recalcule os centroides de cada grupo que tenha ganhado ou perdido um objeto; e
3. Repita o passo anterior até que nenhuma nova designação ocorra.

O resultado do método k-médias é geralmente influenciado pela partição (ou sementes) escolhida no passo inicial. Por esse motivo, recomenda-se a execução do algoritmo algumas vezes partindo de soluções iniciais diferentes.

2.3 Critérios de inferência do número de grupos

Os diferentes algoritmos de agrupamento podem resultar em grupos diferentes em um mesmo conjunto de dados. O resultado de um único algoritmo pode mudar de acordo com os parâmetros usados no algoritmo ou com a ordem de apresentação dos objetos.

Assim, é necessário validar o resultado da análise de agrupamento. Os critérios de inferência do número grupos de um conjunto de dados, mostrados na Tabela 1, foram propostos justamente como padrão para validação do resultado da análise de agrupamento. Eles combinam informações sobre o grau de compactação dentro de cada grupo e o nível de isolamento entre grupos, bem como características geométricas e estatísticas dos dados, o número de objetos e medidas de similaridade ou distância. Charrad et al. (2014) apresentam uma descrição de todos os critérios relacionados na Tabela 1.

Tabela 1 – Critérios de validação da análise de agrupamento.

Critério	Indicação do número ótimo de grupos
C-Index DB Gplus McClain SD SDbw	Valor mínimo do critério
CCC CH Correlação Bisserial Dunn Gamma KL Silhueta Tau Ratkowsky	Valor máximo do critério
Ball Friedman Hartigan Scott Trcovw	Diferença máxima entre níveis hierárquicos do critério
D-Index Gama de Hubert	Método gráfico
Beale	Número de grupos tal que o critério é maior ou igual ao nível de significância α
Duda	Número mínimo de grupos tal que o critério é maior do que um valor crítico
Frey	Número de grupos antes que o valor do critérios seja menor do que 1.00
Gap	Número mínimo de grupos tal que o o valor crítico é maior ou igual a zero
Marriot Tracew	Valor máximo da segunda diferença entre níveis do critério
Pseudo t^2	Número mínimo de grupos tal que o critério é menor do que um valor crítico
Rubin	Valor mínimo da segunda diferença entre níveis do critério

Fonte – Adaptado de Charrad et al. (2014)

Devido ao alto custo computacional na execução de algumas análises, notadamente na aplicação dos critérios Gap, Gamma de Hubert, Gplus e Tau, o estudo se restringiu a analisar agrupamentos resultantes da aplicação de apenas alguns critérios. A escolha dos critérios analisados foi feita de acordo com o número de citações no Google Scholar dos artigos que os apresentam, conforme mostra a Tabela 2.

Os critérios Hartigan, Silhueta, DB, CH, Tracew, Trcovw e Gap foram aplicados

Tabela 2 – Número de citações aos artigos de apresentação de cada critério no Google Scholar.

Critério	Número de citações no Google Scholar
D-Index	6
CCC	27
Frey	29
McClain	73
Ratkowsky	92
Gplus	169
Tau	169
Gamma	238
Marriot	253
C-Index	281
SD	309
Beale	316
SDbw	351
Scott	494
KL	589
Friedman	766
Rubin	766
Ball	1269
Correlação Bisserial	1747
Dunn	1865
Gap	3339
Trcovw	3737
Tracew	3737
CH	3816
DB	4711
Gama de Hubert	5164
Silhueta	6538
Hartigan	9030
Duda	20742
Pseudo t^2	20742

Fonte – Google Scholar, consultado em outubro de 2018, conforme referências de Charrad et al. (2014)

aos agrupamentos realizados no estudo. Uma breve descrição dos critérios usados é feita nas próximas subseções. A seguinte notação é usada:

- n = número de objetos;
- p = número de características medidas em cada objeto;
- k = número de grupos;

- $X = \{x_{ij}\}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, p$, matriz de dados $n \times p$, de p características medidas de n objetos;
- $\bar{x}$ = centroide da matriz de dados X;
- n_j = número de objetos no grupo C_j;
- c_j = centroide do grupo C_j;
- x_i = vetor p-dimensional de características do i-ésimo objeto no grupo C_j;
- $W_k = \sum_{j=1}^{k} \sum_{\mathbf{i} \in C_j} (x_i - c_j)(x_i - c_j)^T$, matriz de dispersão intra grupo para dados agrupados em k grupos;
- $B_k = \sum_{j=1}^{k} n_j (c_j - \bar{x})(c_j - \bar{x})^T$, matriz de dispersão entre grupos para dados agrupados em k grupos.

2.3.1 Critério CH

O critério CH, proposto por Caliński e Harabasz (1974 apud CHARRAD et al., 2014), é dado pela equação 2.12.

$$CH(k) = \frac{tr(B_k)/(k-1)}{tr(W_k)/(n-k)} \tag{2.12}$$

O valor de k que maximiza $CH(k)$ é tomado como o número estimado de grupos.

2.3.2 Critério DB

O critério DB, proposto por Davies e Bouldin (1979 apud CHARRAD et al., 2014), é dado pela equação 2.13.

$$DB(k) = \frac{1}{k} \sum_{j=1}^{k} \max_{j \neq l} \left(\frac{\delta_j + \delta_l}{d_{jl}} \right) \tag{2.13}$$

com

- $j, l = 1, \ldots, k$;
- $d_{jl} = \left(\sum_{i=1}^{k} |c_{ji} - c_{li}|^{\nu} \right)^{\frac{1}{\nu}}$, distância entre os centroides dos grupos C_j e C_l;
- $\delta_j = \left(\frac{1}{n_j} \sum_{i \in C_j} \sum_{l=1}^{k} |x_{il} - c_{jl}|^{\nu} \right)^{\frac{1}{\nu}}$, dispersão interna do grupo C_j.

O valor de k que minimiza $DB(k)$ é tomado como o número estimado de grupos.

2.3.3 Critério Hartigan

O critério Hartigan, proposto por Hartigan (1975 apud CHARRAD et al., 2014), é dado pela equação 2.14.

$$Hartigan = \left(\frac{tr(W_k)}{tr(W_{k+1})} - 1\right)(n - k - 1) \tag{2.14}$$

com $k = 1, \ldots, n - 2$.

O número estimado de grupos corresponde à máxima diferença entre níveis hierárquicos.

2.3.4 Critério Trcovw

Apresentado por Milligan e Cooper (1985 apud CHARRAD et al., 2014), o critério Trcovw é dado pelo traço da matriz de covariâncias do agrupamento:

$$Trcovw = tr(Cov(W_k)) \tag{2.15}$$

A diferença máxima entre níveis hierárquicos é usada para estimar o número de grupos.

2.3.5 Critério Tracew

Analisado por Milligan e Cooper (1985 apud CHARRAD et al., 2014), o critério Tracew é um dos critérios mais populares em análise de agrupamento (CHARRAD et al., 2014). O critério é dado pelo traço da matriz de dispersão intra grupo:

$$Tracew = tr(W_k) \tag{2.16}$$

O valor máximo da segunda diferença entre níveis do critério é usado para estimar o número de grupos.

2.3.6 Critério Silhueta

Rousseeuw (1987 apud CHARRAD et al., 2014) apresentou o critério Silhueta, dado pela equação 2.17.

$$Silhueta = \frac{\sum_{i=1}^{n} S(i)}{n} \tag{2.17}$$

com

- $S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$

- $a(i)$ é a distância média entre o i-ésimo objeto do grupo C_r e os demais objetos do mesmo grupo;
- $b(i)$ é a distância média mínima entre o i-ésimo objeto do grupo C_r e os demais objetos do grupo $C_s, s \neq r$.

O número estimado de grupos é aquele em que o valor do critério é máximo.

2.3.7 Estatística Gap

A estatística Gap, proposta por Tibshirani, Walther e Hastie (2001 apud CHARRAD et al., 2014), é dada pela equação 2.18.

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}) - log(W_k) \tag{2.18}$$

sendo B a quantidade de conjuntos de dados de referência gerados segundo uma distribuição uniforme.

O número de grupos estimado é o menor k tal que

$$Gap(k) \geq Gap(k+1) - s_{k+1} \tag{2.19}$$

com

- $s_k = sd_k\sqrt{1 + 1/B}$;
- sd_k é o desvio-padrão de $log(W_{kb})$.

3 Metodologia

Uma vez realizadas a identificação e a descrição dos métodos de agrupamento, medidas de distância e critérios de inferência existentes na literatura, foram utilizados diferentes tipos de conjuntos de dados sintéticos, com características e número de grupos conhecidos, conforme descrito na seção **Materiais**. Em seguida, foram aplicados os métodos descritos na última seção deste capítulo.

3.1 Materiais

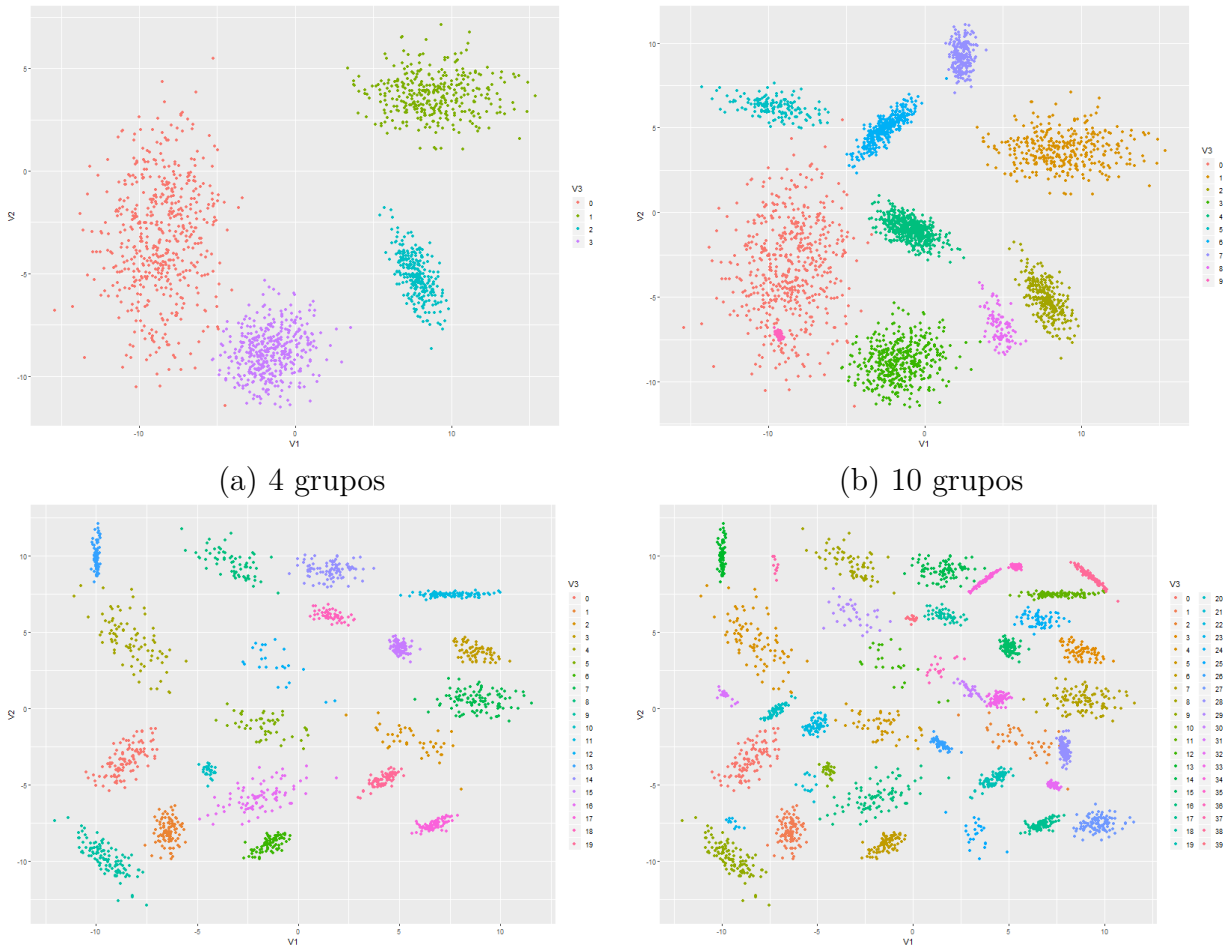
Para avaliar a eficácia dos diferentes critérios de validação de agrupamentos, foram utilizados conjuntos de dados sintéticos disponibilizados por Handl e Knowles (2007) e conjuntos de dados sintéticos gerados utilizando o pacote *KODAMA* (CACCIATORE et al., 2017).

Os conjuntos de dados de Handl e Knowles (2007) foram gerados a partir de modelos de agrupamento usando distribuições normais multivariadas (conjuntos de dimensões $p = 2$ e $p = 10$), dando origem a grupos de forma alongada e de orientação arbitrária. Conjuntos de dimensões maiores ($p = 50$ e $p = 100$) foram gerados por um segundo algoritmo, de forma a garantir formato elipsoidal alongado de orientação arbitrária, mesmo em dimensões elevadas. Para cada uma das 16 combinações de números de grupos e dimensões, Handl e Knowles (2007) disponibilizam 10 conjuntos de dados. Contudo, devido ao alto custo computacional das análises, optou-se por estudar um conjunto de dados para cada combinação, ou seja, 16 conjuntos de dados. As características dos conjuntos de dados estudados são resumidas na Tabela 3. Os conjuntos de dados bidimensionais ($p = 2$) são mostrados na Figura 1.

Tabela 3 – Parâmetros utilizados para geração dos conjuntos de dados sintéticos - gerador gaussiano e gerador elipsoidal.

Parâmetro	Intervalo de valores
Número de grupos (k)	4, 10, 20, 40
Dimensões (p)	2, 10, 50, 100
Tamanho de cada grupo (n_j)	segundo distribuição $U(50, 500)$ para conjuntos com 4 e 10 grupos e $U(10, 100)$ para conjuntos com 20 e 40 grupos.

Fonte – Adaptado de Handl e Knowles (2007)



(a) 4 grupos (b) 10 grupos

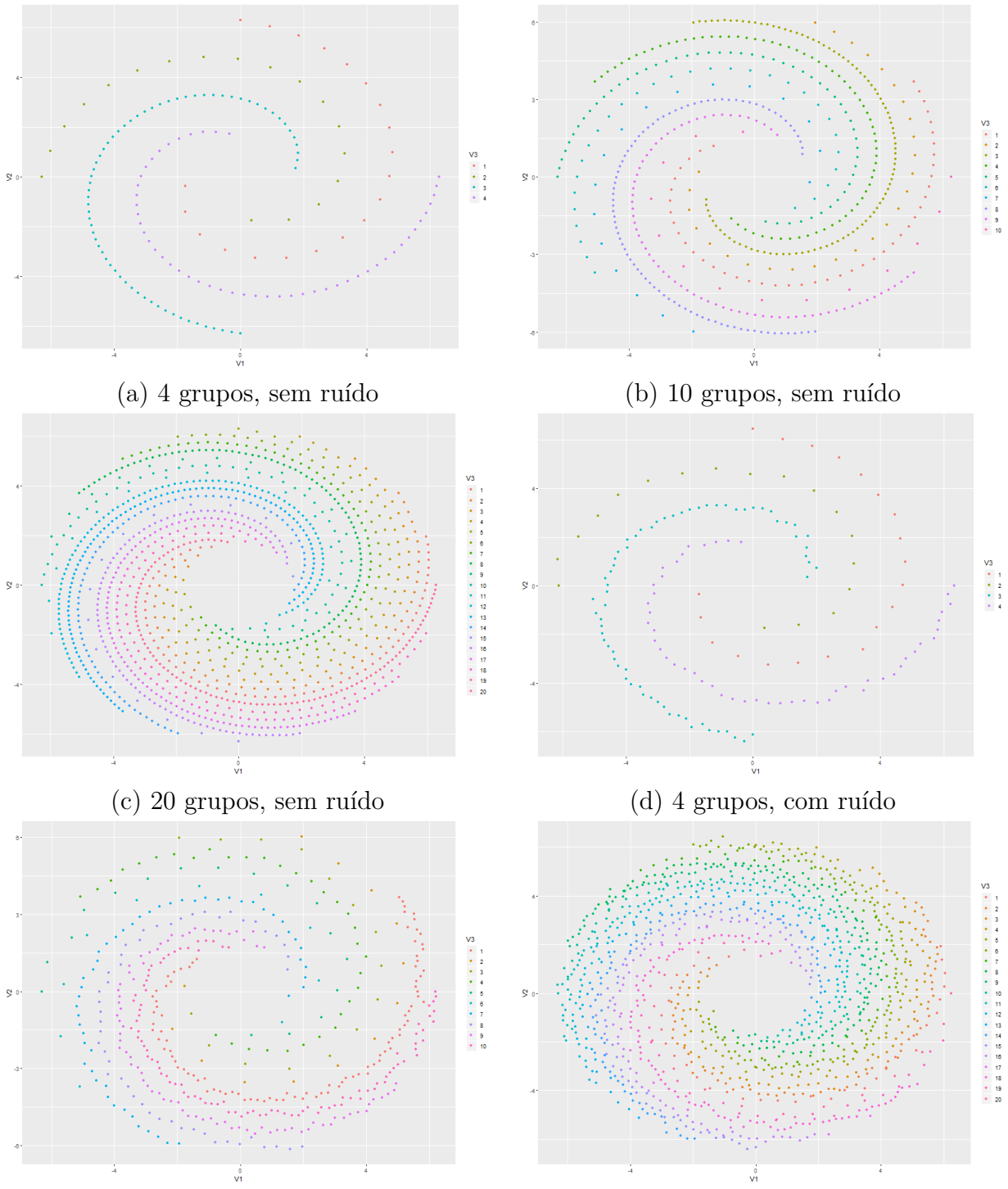
(c) 20 grupos (d) 40 grupos

Figura 1 – Conjuntos de dados sintéticos bidimensionais ($p = 2$) criados por gerador gaussiano com (a) 4 grupos, (b) 10 grupos, (c) 20 grupos e (d) 40 grupos.

Os conjuntos de dados gerados pelo pacote *KODAMA*, proposto por Cacciatore et al. (2017), consistem em conjuntos de dados bidimensionais com grupos em formato espiral, conforme mostrado na Tabela 4. Foram gerados 6 conjuntos de dados: 3 sem ruído no grupo espiral e 3 com ruído no grupo espiral, mostrados na Figura 2.

Tabela 4 – Parâmetros utilizados para geração dos conjuntos de dados sintéticos em formato espiral usando o pacote *KODAMA*.

Parâmetro	Intervalo de valores
Número de grupos (k)	4, 10, 20
Dimensões (p)	2
Tamanho de cada grupo (n_j)	segundo distribuição $U(10, 100)$.



(a) 4 grupos, sem ruído

(b) 10 grupos, sem ruído

(c) 20 grupos, sem ruído

(d) 4 grupos, com ruído

(e) 10 grupos, com ruído

(f) 20 grupos, com ruído

Figura 2 – Conjuntos de dados sintéticos em espiral: sem ruído e com (a) 4 grupos, (b) 10 grupos e (c) 20 grupos; com ruído e com (d) 4 grupos, (e) 10 grupos e (f) 20 grupos.

Assim, um total de 22 conjuntos de dados foram utilizados no estudo. As características dos conjuntos de dados são mostradas na Tabela 5.

Tabela 5 – Características dos conjuntos de dados do estudo.

Tipo	Dimensões	Objetos	Grupos
Gaussiano	2	1572	4
		2972	10
		1517	20
		2563	40
	10	1289	4
		2729	10
		1013	20
		1937	40
Elipsoidal	50	1064	4
		2698	10
		1254	20
		2334	40
	100	1286	4
		2892	10
		1338	20
		2211	40
Espiral	2	154	4
		505	10
		1159	20
Espiral Ruído	2	154	4
		451	10
		1188	20

3.2 Métodos

O trabalho foi desenvolvido em linguagem R, utilizando principalmente o pacote *NbClust*, proposto por Charrad et al. (2014).

Devido ao alto custo computacional na execução de algumas análises, o estudo se restringiu a analisar agrupamentos resultantes da aplicação dos algoritmos não-hierárquico k-médias e hierárquicos aglomerativos Ward, ligação completa e centroide, todos utilizando distância euclidiana.

A partir dos dez critérios mais citados, conforme Tabela 2, foram excluídos os critérios Gamma de Hubert (método gráfico de alto custo computacional), Pseudo t^2 (aplicável apenas em métodos hierárquicos) e Duda (implementação do critério no pacote *NbClust* apresentou erro de execução).

Portanto, os critérios Hartigan, Silhueta, DB, CH, Tracew, Trcovw e Gap foram aplicados aos agrupamentos realizados nos 22 conjuntos de dados sintéticos descritos anteriormente.

Em cada análise, foi avaliada a eficácia dos critérios na inferência do número correto de grupos em cada conjunto de dados, comparando o número de grupos identificados e o número real de grupos, por meio do erro relativo dado por

$$ER = \frac{|k_{estimado} - k_{real}|}{k_{real}}. \tag{3.1}$$

Foi avaliada também a qualidade do agrupamento por meio do índice de Rand ajustado (ARI), implementado no pacote *mclust*.

O índice de Rand compara duas partições diferentes de um mesmo conjunto de dados por meio da proporção de pares concordantes de objetos (pares de objetos alocados em um mesmo grupo ou em grupos diferentes em ambas as partições) no total de pares (concordantes e discordantes) de objetos do conjunto de dados. O índice de Rand varia no intervalo [0,1], porém o seu valor esperado quando duas partições aleatórias são comparadas não é constante (por exemplo, zero).

O índice de Rand ajustado propõe justamente uma modificação de tal forma que seu valor esperado seja constante e igual a zero (partições tomadas ao acaso) e tenha valor máximo igual a 1 (partições comparadas totalmente concordantes).

4 Resultados

Os métodos de agrupamento foram aplicados aos 22 conjuntos de dados sintéticos, conforme detalhado na metodologia. Em todos os casos, foi usada distância euclidiana como medida de dissimilaridade.

As Tabelas 8, 11, 14 e 17 mostram o número de grupos estimado para cada um dos conjuntos de dados utilizando os métodos k-médias, Ward, ligação completa e centroide, respectivamente.

As Tabelas 9, 12, 15 e 18 mostram o erro relativo em cada estimativa do número correto de grupos para cada um dos conjuntos de dados utilizando os métodos k-médias, Ward, ligação completa e centroide, respectivamente.

As Tabelas 10, 13, 16 e 19 mostram o índice de Rand ajustado (ARI) para cada partição obtida em cada um dos conjuntos de dados utilizando os métodos k-médias, Ward, ligação completa e centroide, respectivamente.

Inicialmente, nota-se que o critério Gap não apresentou resultado satisfatório no estudo realizado: em quase todos os casos, o critério apresentou baixa eficácia na estimação do número correto de grupos, resultando em erros relativos elevados, e baixa qualidade no resultado do agrupamento, com ARI baixos, mesmo nos casos mais simples (grupos gaussianos bidimensionais).

Para o conjunto de dados bidimensionais com 4 grupos gaussianos, os demais critérios conseguiram estimar de forma satisfatória o número correto de grupos, identificando partições bastante semelhantes às originais (ARI próximo de 1).

No entanto, à medida que o número de grupos aumentou, os critérios Hartigan, Tracew e Trcovw apresentaram queda na eficácia da estimação do número de grupos, bem como na qualidade do agrupamento encontrado.

Já os critérios CH, Silhueta e DB, no que diz respeito aos conjuntos de dados com grupos gaussianos, apresentaram bons resultados tanto para os conjuntos de dados bidimensionais como para os conjuntos de dados com $p = 10$. Os erros relativos foram baixos e a qualidade do agrupamento alta em todos os métodos de agrupamento, exceto para o método centroide, especificamente no conjunto de dados com 10 grupos gaussianos com $p = 10$, no qual o resultado não foi satisfatório.

Para os conjuntos de dados com grupos elipsoidais de alta dimensão, os resultados não foram tão bons quanto para os dados com grupos gaussianos. Os critérios conseguiram identificar corretamente o número de grupos quando este era pequeno ($k = 4$), porém a qualidade do agrupamento foi no máximo regular ($ARI \leq 0,56$), indicando que os grupos

formados não correspondiam aos grupos originais.

Quando o número de dimensões e de grupos aumentou, os critérios CH, Silhueta e DB ainda conseguiram em alguns casos identificar corretamente o número de grupos, em especial o critério CH quando usado o método centroide de agrupamento. Porém, os baixos ARI indicam novamente a baixa qualidade da partição encontrada.

Quanto aos dados em espiral, os critérios apresentaram baixos erros relativos para os casos com 4 grupos, tanto nos dados com ruído quanto nos dados sem ruído. Porém, os baixos ARI indicaram que as partições identificadas não correspondiam às partições originais, como pode ser visto na Figura 3. O resultado mostrado corresponde ao agrupamento usando k-médias, critério CH, dos dados em espiral sem ruído com quatro grupos: o erro relativo zero indica que o número de grupos foi identificado corretamente, porém o ARI próximo de zero indica que a partição encontrada é espúria.

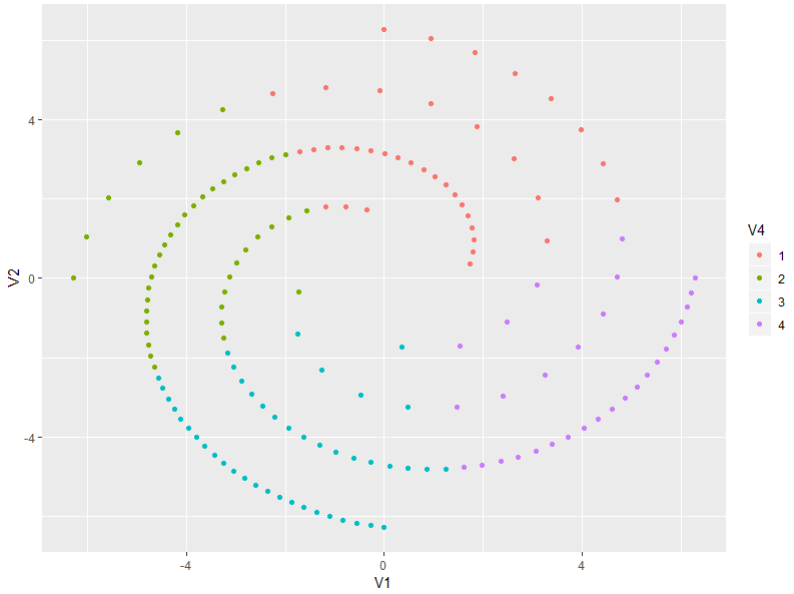


Figura 3 – Resultado do agrupamento de dados em espiral sem ruído com quatro grupos usando k-médias e critério CH.

Para todos os outros casos de conjuntos de dados com grupos em espiral, os resultados não foram satisfatórios, com alto erro relativo e baixo ARI.

As Tabelas 6 e 7 resumem de forma agregada por método de agrupamento os erros relativos e ARI encontrados, respectivamente.

Tabela 6 – Resultados dos agrupamentos - Erro relativo - Total por método de agrupamento

Método	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
k-médias	6,05	6,55	6,65	8,18	11,78	13,50	16,75
Ward	8,22	5,48	5,38	10,50	13,10	14,50	17,00
Ligação Completa	7,85	8,78	6,43	9,50	12,78	14,30	16,40
Centroide	7,58	11,35	10,65	8,80	11,50	12,38	16,45
Total	29,70	32,15	29,10	36,98	49,15	54,68	66,60

Tabela 7 – Resultados dos agrupamentos - Índice de Rand Ajustado - Média por método de agrupamento e média geral

Método	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
k-médias	0,46	0,46	0,44	0,37	0,24	0,23	0,13
Ward	0,51	0,52	0,52	0,31	0,25	0,21	0,13
Ligação Completa	0,45	0,37	0,37	0,34	0,21	0,18	0,1
Centroide	0,35	0,32	0,34	0,21	0,14	0,14	0,07
Média	0,44	0,42	0,42	0,31	0,21	0,19	0,11

De forma geral, os critérios CH, Silhueta e DB mostraram melhores resultados, com erros relativos mais baixos e ARI mais elevados, especialmente nos conjuntos de dados com grupos gaussianos.

Porém, ao aumentar a complexidade da geometria dos grupos e a dimensão dos dados, os resultados obtidos já não foram tão satisfatórios, o que pode ser explicado pela simplicidade relativa dos métodos de agrupamento e medida de distância utilizados.

Tabela 8 – Resultados dos agrupamentos - Número de grupos estimado - k-médias

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	5	4	4	4	4	4	2
		10	10	9	9	7	3	3	2
		20	30	21	20	28	3	3	2
		40	48	38	37	4	3	3	2
	10	4	3	3	3	3	3	3	2
		10	4	11	5	4	4	4	2
		20	11	11	11	11	4	4	2
		40	44	44	28	28	3	5	2
Elipsoidal	50	4	3	3	3	3	3	3	3
		10	15	7	7	10	3	3	2
		20	26	24	18	24	18	6	2
		40	46	36	47	37	23	5	2
	100	4	2	2	2	3	3	3	2
		10	13	13	13	7	5	5	2
		20	30	24	24	8	8	4	2
		40	60	60	54	28	28	5	2
Espiral	2	4	4	4	4	4	3	3	2
		10	13	3	4	4	3	3	2
		20	19	3	4	3	3	4	2
Espiral Ruído	2	4	4	4	4	4	3	3	2
		10	7	3	4	3	3	3	2
		20	19	3	4	3	3	3	2

Tabela 9 – Resultados dos agrupamentos - Erro relativo - k-médias

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,25	0,00	0,00	0,00	0,00	0,00	0,50
		10	0,00	0,10	0,10	0,30	0,70	0,70	0,80
		20	0,50	0,05	0,00	0,40	0,85	0,85	0,90
		40	0,20	0,05	0,08	0,90	0,93	0,93	0,95
	10	4	0,25	0,25	0,25	0,25	0,25	0,25	0,50
		10	0,60	0,10	0,50	0,60	0,60	0,60	0,80
		20	0,45	0,45	0,45	0,45	0,80	0,80	0,90
		40	0,10	0,10	0,30	0,30	0,93	0,88	0,95
Elipsoidal	50	4	0,25	0,25	0,25	0,25	0,25	0,25	0,25
		10	0,50	0,30	0,30	0,00	0,70	0,70	0,80
		20	0,30	0,20	0,10	0,20	0,10	0,70	0,90
		40	0,15	0,10	0,18	0,08	0,43	0,88	0,95
	100	4	0,50	0,50	0,50	0,25	0,25	0,25	0,50
		10	0,30	0,30	0,30	0,30	0,50	0,50	0,80
		20	0,50	0,20	0,20	0,60	0,60	0,80	0,90
		40	0,50	0,50	0,35	0,30	0,30	0,88	0,95
Espiral	2	4	0,00	0,00	0,00	0,00	0,25	0,25	0,50
		10	0,30	0,70	0,60	0,60	0,70	0,70	0,80
		20	0,05	0,85	0,80	0,85	0,85	0,80	0,90
Espiral Ruído	2	4	0,00	0,00	0,00	0,00	0,25	0,25	0,50
		10	0,30	0,70	0,60	0,70	0,70	0,70	0,80
		20	0,05	0,85	0,80	0,85	0,85	0,85	0,90

Tabela 10 – Resultados dos agrupamentos - Índice de Rand Ajustado - k-médias

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,83	0,96	0,96	0,96	0,96	0,96	0,49
		10	0,83	0,84	0,84	0,59	0,33	0,33	0,21
		20	0,79	0,82	0,79	0,68	0,20	0,20	0,11
		40	0,83	0,88	0,85	0,15	0,11	0,11	0,06
	10	4	0,97	0,97	0,97	0,97	0,97	0,97	0,69
		10	0,57	0,88	0,67	0,57	0,57	0,57	0,15
		20	0,74	0,74	0,74	0,74	0,30	0,30	0,12
		40	0,94	0,94	0,88	0,88	0,11	0,18	0,04
Elipsoidal	50	4	0,41	0,41	0,41	0,41	0,41	0,41	0,41
		10	0,44	0,31	0,31	0,42	0,09	0,09	0,04
		20	0,41	0,36	0,25	0,36	0,25	0,08	0,01
		40	0,29	0,20	0,27	0,26	0,08	0,02	0,02
	100	4	0,16	0,16	0,16	0,29	0,29	0,29	0,16
		10	0,51	0,51	0,51	0,30	0,20	0,20	0,03
		20	0,45	0,37	0,37	0,08	0,08	0,03	0,04
		40	0,36	0,36	0,31	0,14	0,14	0,03	0,00
Espiral	2	4	0,10	0,10	0,10	0,10	0,08	0,08	0,07
		10	0,10	0,05	0,05	0,05	0,05	0,05	0,02
		20	0,07	0,02	0,03	0,02	0,02	0,03	0,02
Espiral Ruído	2	4	0,10	0,10	0,10	0,10	0,08	0,08	0,07
		10	0,04	0,04	0,04	0,04	0,04	0,04	0,03
		20	0,07	0,02	0,03	0,02	0,02	0,02	0,02

Tabela 11 – Resultados dos agrupamentos - Número de grupos estimado - Método Ward

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	4	4	4	4	4	3	2
		10	9	8	8	9	3	3	2
		20	30	20	20	4	4	3	2
		40	60	36	36	4	4	3	2
	10	4	3	4	3	3	3	3	2
		10	4	9	9	4	4	3	2
		20	20	19	20	3	3	3	2
		40	39	40	40	7	7	3	2
Elipsoidal	50	4	6	3	3	3	3	3	2
		10	15	8	8	8	3	3	2
		20	29	22	23	14	3	3	2
		40	60	46	46	36	3	3	2
	100	4	2	2	2	6	3	3	2
		10	15	13	13	5	5	3	2
		20	30	27	28	5	5	3	2
		40	60	43	43	21	6	3	2
Espiral	2	4	6	4	4	4	4	3	2
		10	15	3	15	3	3	3	2
		20	28	3	4	3	3	3	2
Espiral Ruído	2	4	4	4	4	4	4	3	2
		10	14	3	4	3	3	3	2
		20	30	3	4	3	3	3	2

Tabela 12 – Resultados dos agrupamentos - Erro relativo - Método Ward

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,00	0,00	0,00	0,00	0,00	0,25	0,50
		10	0,10	0,20	0,20	0,10	0,70	0,70	0,80
		20	0,50	0,00	0,00	0,80	0,80	0,85	0,90
		40	0,50	0,10	0,10	0,90	0,90	0,93	0,95
	10	4	0,25	0,00	0,25	0,25	0,25	0,25	0,50
		10	0,60	0,10	0,10	0,60	0,60	0,70	0,80
		20	0,00	0,05	0,00	0,85	0,85	0,85	0,90
		40	0,03	0,00	0,00	0,83	0,83	0,93	0,95
Elipsoidal	50	4	0,50	0,25	0,25	0,25	0,25	0,25	0,50
		10	0,50	0,20	0,20	0,20	0,70	0,70	0,80
		20	0,45	0,10	0,15	0,30	0,85	0,85	0,90
		40	0,50	0,15	0,15	0,10	0,93	0,93	0,95
	100	4	0,50	0,50	0,50	0,50	0,25	0,25	0,50
		10	0,50	0,30	0,30	0,50	0,50	0,70	0,80
		20	0,50	0,35	0,40	0,75	0,75	0,85	0,90
		40	0,50	0,08	0,08	0,48	0,85	0,93	0,95
Espiral	2	4	0,50	0,00	0,00	0,00	0,00	0,25	0,50
		10	0,50	0,70	0,50	0,70	0,70	0,70	0,80
		20	0,40	0,85	0,80	0,85	0,85	0,85	0,90
Espiral Ruído	2	4	0,00	0,00	0,00	0,00	0,00	0,25	0,50
		10	0,40	0,70	0,60	0,70	0,70	0,70	0,80
		20	0,50	0,85	0,80	0,85	0,85	0,85	0,90

Tabela 13 – Resultados dos agrupamentos - Índice de Rand Ajustado - Método Ward

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,98	0,98	0,98	0,98	0,98	0,81	0,49
		10	0,86	0,91	0,91	0,86	0,35	0,35	0,12
		20	0,83	0,99	0,99	0,29	0,29	0,17	0,11
		40	0,77	0,93	0,93	0,14	0,14	0,11	0,06
	10	4	0,97	1,00	0,97	0,97	0,97	0,97	0,69
		10	0,59	0,97	0,97	0,59	0,59	0,45	0,26
		20	1,00	0,99	1,00	0,18	0,18	0,18	0,04
		40	1,00	1,00	1,00	0,30	0,30	0,12	0,06
Elipsoidal	50	4	0,49	0,43	0,43	0,43	0,43	0,43	0,27
		10	0,45	0,33	0,33	0,33	0,10	0,10	0,07
		20	0,43	0,37	0,40	0,26	0,05	0,05	0,01
		40	0,49	0,33	0,33	0,29	0,01	0,01	0,00
	100	4	0,21	0,21	0,21	0,41	0,37	0,37	0,21
		10	0,56	0,64	0,64	0,20	0,20	0,06	0,02
		20	0,61	0,60	0,58	0,04	0,04	0,02	0,01
		40	0,41	0,32	0,32	0,08	0,02	0,01	0,01
Espiral	2	4	0,11	0,15	0,15	0,15	0,15	0,07	0,15
		10	0,10	0,05	0,10	0,05	0,05	0,05	0,04
		20	0,08	0,03	0,03	0,03	0,03	0,03	0,02
Espiral Ruído	2	4	0,09	0,09	0,09	0,09	0,09	0,10	0,08
		10	0,08	0,05	0,04	0,05	0,05	0,05	0,02
		20	0,08	0,02	0,03	0,02	0,02	0,02	0,02

Tabela 14 – Resultados dos agrupamentos - Número de grupos estimado - Ligação Completa

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	5	2	3	4	3	3	3
		10	12	12	8	11	3	3	2
		20	30	24	10	24	3	3	2
		40	47	47	35	4	4	3	2
	10	4	4	4	5	4	4	3	2
		10	4	9	9	4	4	3	2
		20	20	19	21	13	3	3	2
		40	40	40	41	19	4	3	2
Elipsoidal	50	4	6	2	3	6	3	3	2
		10	11	6	7	5	4	4	2
		20	28	28	15	15	15	3	2
		40	59	50	41	35	9	3	2
	100	4	6	2	2	4	4	3	2
		10	13	4	7	4	4	4	2
		20	28	2	18	4	4	3	2
		40	60	60	60	58	6	3	2
Espiral	2	4	6	3	4	3	3	3	3
		10	5	3	4	3	3	3	3
		20	5	3	5	3	3	3	2
Espiral Ruído	2	4	5	3	4	5	3	3	2
		10	7	4	4	3	3	3	2
		20	7	3	5	3	3	3	2

Tabela 15 – Resultados dos agrupamentos - Erro relativo - Ligação Completa

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,25	0,50	0,25	0,00	0,25	0,25	0,25
		10	0,20	0,20	0,20	0,10	0,70	0,70	0,80
		20	0,50	0,20	0,50	0,20	0,85	0,85	0,90
		40	0,18	0,18	0,13	0,90	0,90	0,93	0,95
	10	4	0,00	0,00	0,25	0,00	0,00	0,25	0,50
		10	0,60	0,10	0,10	0,60	0,60	0,70	0,80
		20	0,00	0,05	0,05	0,35	0,85	0,85	0,90
		40	0,00	0,00	0,03	0,53	0,90	0,93	0,95
Elipsoidal	50	4	0,50	0,50	0,25	0,50	0,25	0,25	0,50
		10	0,10	0,40	0,30	0,50	0,60	0,60	0,80
		20	0,40	0,40	0,25	0,25	0,25	0,85	0,90
		40	0,48	0,25	0,03	0,13	0,78	0,93	0,95
	100	4	0,50	0,50	0,50	0,00	0,00	0,25	0,50
		10	0,30	0,60	0,30	0,60	0,60	0,60	0,80
		20	0,40	0,90	0,10	0,80	0,80	0,85	0,90
		40	0,50	0,50	0,50	0,45	0,85	0,93	0,95
Espiral	2	4	0,50	0,25	0,00	0,25	0,25	0,25	0,25
		10	0,50	0,70	0,60	0,70	0,70	0,70	0,70
		20	0,75	0,85	0,75	0,85	0,85	0,85	0,90
Espiral Ruído	2	4	0,25	0,25	0,00	0,25	0,25	0,25	0,50
		10	0,30	0,60	0,60	0,70	0,70	0,70	0,80
		20	0,65	0,85	0,75	0,85	0,85	0,85	0,90

Tabela 16 – Resultados dos agrupamentos - Índice de Rand Ajustado - Ligação Completa

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,86	0,49	0,81	0,68	0,81	0,81	0,81
		10	0,81	0,81	0,75	0,67	0,39	0,39	0,16
		20	0,81	0,86	0,54	0,86	0,15	0,15	0,10
		40	0,80	0,80	0,78	0,13	0,13	0,08	0,06
	10	4	1,00	1,00	1,00	1,00	1,00	0,86	0,19
		10	0,52	0,67	0,67	0,52	0,52	0,42	0,23
		20	1,00	0,98	1,00	0,81	0,16	0,16	0,03
		40	1,00	1,00	1,00	0,67	0,13	0,08	0,07
Elipsoidal	50	4	0,54	0,07	0,10	0,54	0,10	0,10	0,07
		10	0,32	0,26	0,27	0,21	0,21	0,21	0,06
		20	0,29	0,29	0,13	0,13	0,13	0,02	0,01
		40	0,28	0,24	0,15	0,16	0,03	0,00	0,00
	100	4	0,40	0,00	0,00	0,35	0,35	0,26	0,00
		10	0,41	0,08	0,12	0,08	0,08	0,08	0,05
		20	0,34	0,01	0,11	0,02	0,02	0,02	0,01
		40	0,26	0,26	0,26	0,25	0,02	0,01	0,00
Espiral	2	4	0,07	0,06	0,10	0,06	0,06	0,06	0,06
		10	0,07	0,07	0,06	0,07	0,07	0,07	0,07
		20	0,04	0,03	0,04	0,03	0,03	0,03	0,01
Espiral Ruído	2	4	0,06	0,08	0,08	0,06	0,08	0,08	0,07
		10	0,05	0,05	0,05	0,07	0,07	0,07	0,06
		20	0,03	0,03	0,03	0,03	0,03	0,03	0,01

Tabela 17 – Resultados dos agrupamentos - Número de grupos estimado - Método Centroide

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	4	4	3	4	4	4	2
		10	8	8	8	6	3	3	2
		20	26	26	28	3	3	4	3
		40	40	40	57	4	3	4	2
	10	4	2	2	2	3	3	5	2
		10	2	2	2	10	10	15	2
		20	23	23	30	15	3	3	2
		40	54	54	59	26	5	9	2
Elipsoidal	50	4	3	3	6	3	3	3	2
		10	5	2	13	5	5	5	2
		20	22	2	26	22	10	8	2
		40	38	2	2	36	11	11	2
	100	4	2	2	4	4	4	4	2
		10	9	6	14	6	6	6	2
		20	5	2	2	5	5	3	2
		40	35	2	2	26	21	13	2
Espiral	2	4	5	3	6	3	3	3	3
		10	14	4	8	4	4	3	2
		20	4	4	4	4	4	3	2
Espiral	2	4	4	3	5	3	3	3	3
		10	3	3	15	3	3	3	2
		20	5	4	9	5	3	3	2

Tabela 18 – Resultados dos agrupamentos - Erro relativo - Método Centroide

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,00	0,00	0,25	0,00	0,00	0,00	0,50
		10	0,20	0,20	0,20	0,40	0,70	0,70	0,80
		20	0,30	0,30	0,40	0,85	0,85	0,80	0,85
		40	0,00	0,00	0,43	0,90	0,93	0,90	0,95
	10	4	0,50	0,50	0,50	0,25	0,25	0,25	0,50
		10	0,80	0,80	0,80	0,00	0,00	0,50	0,80
		20	0,15	0,15	0,50	0,25	0,85	0,85	0,90
		40	0,35	0,35	0,48	0,35	0,88	0,78	0,95
Elipsoidal	50	4	0,25	0,25	0,50	0,25	0,25	0,25	0,50
		10	0,50	0,80	0,30	0,50	0,50	0,50	0,80
		20	0,10	0,90	0,30	0,10	0,50	0,60	0,90
		40	0,05	0,95	0,95	0,10	0,73	0,73	0,95
	100	4	0,50	0,50	0,00	0,00	0,00	0,00	0,50
		10	0,10	0,40	0,40	0,40	0,40	0,40	0,80
		20	0,75	0,90	0,90	0,75	0,75	0,85	0,90
		40	0,13	0,95	0,95	0,35	0,48	0,68	0,95
Espiral	2	4	0,25	0,25	0,50	0,25	0,25	0,25	0,25
		10	0,40	0,60	0,20	0,60	0,60	0,70	0,80
		20	0,80	0,80	0,80	0,80	0,80	0,85	0,90
Espiral Ruído	2	4	0,00	0,25	0,25	0,25	0,25	0,25	0,25
		10	0,70	0,70	0,50	0,70	0,70	0,70	0,80
		20	0,75	0,80	0,55	0,75	0,85	0,85	0,90

Tabela 19 – Resultados dos agrupamentos - Índice de Rand Ajustado - Método Centroide

Tipo	Dimensões	Grupos	CH	Silhueta	DB	Hartigan	Tracew	Trcovw	Gap
Gaussiano	2	4	0,97	0,97	0,62	0,97	0,97	0,97	0,49
		10	0,90	0,90	0,90	0,49	0,23	0,23	0,12
		20	0,98	0,98	0,98	0,15	0,15	0,22	0,15
		40	0,87	0,87	0,89	0,13	0,07	0,13	0,01
	10	4	0,00	0,00	0,00	0,00	0,00	0,00	0,00
		10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
		20	0,96	0,96	0,96	0,65	0,04	0,04	0,01
		40	0,98	0,98	0,97	0,56	0,02	0,04	0,00
Elipsoidal	50	4	0,57	0,57	0,68	0,57	0,57	0,57	0,28
		10	0,16	0,01	0,32	0,16	0,16	0,16	0,01
		20	0,12	0,00	0,12	0,12	0,04	0,02	0,00
		40	0,07	0,00	0,00	0,06	0,01	0,01	0,00
	100	4	0,23	0,23	0,23	0,23	0,23	0,23	0,23
		10	0,31	0,19	0,34	0,19	0,19	0,19	0,03
		20	0,01	0,00	0,00	0,01	0,01	0,00	0,00
		40	0,07	0,00	0,00	0,05	0,03	0,01	0,00
Espiral	2	4	0,08	0,09	0,09	0,09	0,09	0,09	0,09
		10	0,11	0,06	0,06	0,06	0,06	0,03	0,01
		20	0,03	0,03	0,03	0,03	0,03	0,03	0,01
Espiral Ruído	2	4	0,12	0,11	0,13	0,11	0,11	0,11	0,11
		10	0,03	0,03	0,06	0,03	0,03	0,03	0,01
		20	0,03	0,03	0,04	0,03	0,02	0,02	0,02

5 Conclusão

Neste trabalho, foram analisados critérios de inferência do número correto de grupos em conjuntos de dados, considerando diferentes métodos de agrupamento aplicados a diversas configurações de conjuntos de dados.

Por uma limitação de capacidade computacional e tempo, foram selecionados sete dos trinta critérios disponíveis no pacote *NbClust* (linguagem R), aplicados conjuntamente com quatro métodos de agrupamento (k-médias, Ward, ligação completa e centroide), utilizando distância euclidiana.

Após aplicar os métodos e critérios nos 22 conjuntos de dados sintéticos, verificou-se que os critérios CH, Silhueta e DB apresentaram melhores resultados, identificando de forma satisfatória o número correto de grupos e resultando em partições próximas às originais, em especial nos conjuntos de dados com grupos gaussianos. Nestes conjuntos de dados, os critérios Hartigan, Tracew e Trcovw também apresentaram bons resultados, principalmente nos conjuntos com poucos grupos. Porém, à medida que o número de grupos e de dimensões aumentaram, o resultado obtido por estes três critérios se deteriorou.

Para dados de geometria mais complexa ou de dimensões mais elevadas, o desempenho dos critérios não foi satisfatória de forma geral. Em alguns casos, os critérios CH, Silhueta e DB ainda conseguiram se aproximar do número correto de grupos, porém a partição encontrada não correspondia ao agrupamento original dos dados.

Quanto ao critério Gap, em nenhuma das análises realizadas os resultados foram satisfatórios. Considerando ainda que o custo computacional para aplicar o critério é alto, o critério Gap não se mostrou como bom critério para inferência do número de grupos nos casos estudados. Vale ressaltar que o método disponível no pacote *NbClust* permite apenas uma quantidade limitada de configuração de parâmetros para execução do agrupamento. Talvez em outras implementações do critério Gap seja possível ajustar de forma mais refinada os parâmetros de execução, como por exemplo a possibilidade de escolha da distribuição de referência, podendo levar a resultados melhores.

Os resultados deste trabalho indicam que a utilização de métodos de agrupamento mais sofisticados poderiam melhorar o desempenho dos critérios analisados.

Caso se dispusesse de máquina de alto poder computacional, uma análise mais abrangente, incluindo outros métodos de agrupamento, mais conjuntos de dados para cada configuração de geometria/dimensões, poderia gerar resultados mais satisfatórios para os critérios analisados, mesmo nos casos mais complexos.

Referências

CACCIATORE, S. et al. KODAMA: An R package for knowledge discovery and data mining. *Bioinformatics*, v. 33, n. 4, p. 621–623, 2017. Citado 2 vezes nas páginas 23 e 24.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Citado na página 20.

CHARRAD, M. et al. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, v. 61, 2014. Citado 10 vezes nas páginas 11, 15, 16, 17, 18, 19, 20, 21, 22 e 27.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, n. 2, p. 224–227, 1979. Citado na página 20.

HANDL, J.; KNOWLES, J. An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, v. 11, p. 56–76, 2007. Citado 2 vezes nas páginas 11 e 23.

HARTIGAN, J. A. *Clustering Algorithms*. New York: John Willey and Sons, 1975. Citado na página 21.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. Upper Saddle River: Pearson Prentice Hall, 2007. Citado 4 vezes nas páginas 11, 14, 16 e 17.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, Springer, v. 50, n. 2, p. 159–179, 1985. Citado na página 21.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987. Citado na página 21.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 63, n. 2, p. 411–423, 2001. Citado na página 22.