UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

INÁCIO ROBSON ALVES DO NASCIMENTO

AGRUPAMENTO DE FORMAS TRIDIMENSIONAIS: UM ESTUDO SOBRE OS MÉTODOS K-MÉDIAS, CLARANS E HILL CLIMBING

Recife

2021

INÁCIO ROBSON ALVES DO NASCIMENTO

AGRUPAMENTO DE FORMAS TRIDIMENSIONAIS: UM ESTUDO SOBRE OS MÉTODOS K-MÉDIAS, CLARANS E HILL CLIMBING

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Prof. Dr. Getúlio José Amorim do Amaral

Recife

2021

Catalogação na fonte
Bibliotecária Fernanda Bernardo Ferreira CRB4-2165

N244a Nascimento, Inácio Robson Alves do
Agrupamento de formas tridimensionais: um estudo sobre os métodos K-médias, Clarans e Hill Climbing/ Inácio Robson Alves do Nascimento– 2021.
75f., il., fig.

Orientador: Getúlio José Amorim do Amaral.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2021.
Inclui referências.

1. Estatística Aplicada. 2. Análise de agrupamento. 3. Análise estatística de formas. 4. Formas tridimensionais. I. Amaral, Getúlio José Amorim do (orientador). II. Título.

310 CDD (22. ed.) UFPE-CCEN 2021-55

INÁCIO ROBSON ALVES DO NASCIMENTO

AGRUPAMENTO DE FORMAS TRIDIMENSIONAIS: UM ESTUDO SOBRE OS MÉTODOS K-MÉDIAS, CLARANS E HILL CLIMBING

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 25 de fevereiro de 2021.

BANCA EXAMINADORA

Profº. Getúlio José Amorim do Amaral (Orientador)
UFPE

Profª. Fernanda De Bastiani (Examinador Interno)
UFPE

Profª. Renata Maria Cardoso Rodrigues de Souza (Examinador Externo)
UFPE

À Deus, minha família, amigos e todos que me apoiam direta e indiretamente.

AGRADECIMENTOS

Primeiramente à Deus, por me conceder discernimento, saúde e determinação para seguir nessa caminhada.

À minha família, ao apoio financeiro no início dessa jornada e ao apoio emocional no decorrer dela. À minha namorada, Rayneide, pelo carinho, companhia e apoio emocional. A presença de vocês foi fundamental para conseguir concluir mais essa etapa da minha vida.

Ao meu orientador, professor Dr. Getúlio Amorim, pela confiança, incentivo, apoio e orientação no desenvolvimento desse trabalho e no meu desenvolvimento acadêmico.

À todos os professores do PPGE, pela competência e pela contribuição na minha formação, aprendi muito com seus ensinamentos.

Aos componentes da minha banca, pelas sugestões e contribuições.

Aos novos e velhos amigos, em especial ao Alexandre e Thalytta.

Aos funcionários da secretaria que de maneira prestativa sempre estavam dispostos a ajudar os alunos da pós-graduação.

À CAPES pelo suporte financeiro.

E a todos que direta ou indiretamente contribuíram e fizeram parte da minha formação, o meu muito obrigado.

"Eu me lembrava de que o mundo real era vasto, e que uma quantidade enorme de esperanças e medos, de sensações e emoções, estava à espera daqueles que ousassem sair por ele afora, buscando, em meio a seus perigos, o verdadeiro conhecimento do que é a vida."

Charlotte Brontë

RESUMO

Os métodos de Análise Estatística de Formas são utilizados para trabalhar com formas geométricas de objetos e forma é toda informação que sobra de um objeto depois que são retirados os efeitos de escala, locação e rotação. As técnicas da Análise de Agrupamento são utilizadas para classificar os objetos de um conjunto de dados em grupos. Em muitos campos de estudos, faz-se necessário agrupar as formas de objetos geométricos. Trabalhos envolvendo a adaptação de algoritmos de agrupamento no contexto de análise de formas foram desenvolvidos por alguns pesquisadores. Descreve-se nesta dissertação os métodos já adaptados no contexto de formas tridimensionais, K-médias, K-médias aparado e K-médias utilizando a média ϕ. Além da apresentação da adaptação de novos métodos: K-médias aparado utilizando a média ϕ, *CLARANS* e *Hill Climbing* no contexto de formas 3D, como propostas do presente trabalho. Foram realizados dois experimentos de simulação em diferentes cenários de isotropia e anisotropia, com diferentes graus de dispersão. Gerou-se amostras da distribuição normal multivariada para a formação dos grupos, com diferentes configurações médias e matrizes de covariâncias. Considerou-se também a aplicação dos algoritmos propostos em três conjuntos de dados reais disponíveis na literatura da área. Os grupos obtidos foram avaliados usando medidas de validação de agrupamento. As medidas utilizadas foram os Índices de Rand Ajustado e Silhueta. De modo geral, avaliando os valores obtidos pelos índices, os algoritmos de agrupamento, tantos os que já foram adaptados quantos os propostos, apresentaram boas eficácias de agrupamento nos cenários com baixa dispersão. Enquanto que para os cenários com alta dispersão, em anisotropia, há indícios de que os métodos propostos possuem bons desempenhos em relação aos métodos implementados por outros autores, na maioria das situações e levando em consideração pelo menos um dos índices medidos.

Palavras-chave: Análise de agrupamento. Análise estatística de formas. Formas tridimensionais. Validação de agrupamento.

ABSTRACT

The methods of Statistical Shape Analysis are used to work with geometric shapes of objects and shape is all the information that remains of an object after the effects of scale, location and rotation are removed. Cluster Analysis techniques are used to classify the objects of a data set into groups. In many fields of study, it is necessary to group the shapes of geometric objects. Studies involving the adaptation of clustering algorithms in the context of shape analysis have been developed by some researchers. This dissertation describes the methods already adapted in the context of three-dimensional forms, K-means, trimmed K-means and K-means using the mean ϕ. In addition to the presentation of the adaptation of new methods: trimmed K-means using the mean ϕ, CLARANS and Hill Climbing in the context of 3D shapes, as proposed in the present work. Two simulation experiments were carried out in different isotropy and anisotropy scenarios, with different degrees of dispersion. Samples of the multivariate normal distribution were generated for the formation of groups, with different mean configurations and covariance matrices. It was also considered the application of the algorithms proposed in three sets of real data available in the literature of the area. The groups obtained were evaluated using cluster validation measures. The measures used were the Adjusted Rand and Silhouette Indexes. In general, evaluating the values obtained by the indexes, the grouping algorithms, both those that have already been adapted and those proposed, showed good grouping efficiencies in scenarios with low dispersion. While for scenarios with high dispersion, in anisotropy, there is evidence that the proposed methods perform well in relation to the methods implemented by other authors, in most cases and taking into account at least one of the measured indices.

Keywords: Cluster analysis. Statistical shape analysis. Three-dimensional shapes. Cluster validation.

LISTA DE FIGURAS

LISTA DE ALGORITMOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Representações geométricas e imagens de objetos são agentes de estudos e pesquisas em diversos campos e aplicações, tais como Biologia, Medicina, Geografia, Neurociência, Arqueologia e Reconhecimento Facial. A Análise Estatística de Formas *(Statistical Shape Analysis)* é um ramo da estatística, utilizado para trabalhar com representações geométricas e formas de objetos. Os primeiros trabalhos na área remetem aos estudos de Kendall (1977) e Bookstein et al. (1986). Forma é toda a informação que resta quando são removidos os efeitos de locação, escala e rotação de um determinado objeto. Algumas aplicações práticas envolvendo o tema envolvem a comparação entre as formas do córtex do cérebro de pacientes com e sem esquizofrenia em um estudo neurocientífico ou a comparação entre as formas de crânios de macacos machos e fêmeas (DRYDEN; MARDIA, 2016). Em diversas áreas do conhecimento faz-se necessário estudar medidas de resumo e comparação entre formas, deste modo é de grande importância desenvolver diferentes técnicas estatísticas para aplicações no contexto de formas.

Em diversas situações do cotidiano é relevante a classificação dos indivíduos de um banco de dados em grupos, seja para auxiliar no entendimento do fenômeno estudado ou simplesmente para organizar os dados. Análise de Agrupamento é um conjunto de técnicas que tem como objetivo criar grupos, de tal maneira que os elementos de cada grupo sejam similares entre si e os grupos sejam diferentes. As técnicas de agrupamento são bastantes disseminadas nas áreas de Análise Multivariada e Aprendizado de Máquina. Os principais métodos de agrupamentos são os hierárquicos e os particionais (EVERITT et al., 2011). Nos métodos hierárquicos, os grupos são formados a partir de uma estrutura hierárquica de acordo com as vizinhanças entres os objetos. Enquanto que os métodos de particionamento formam os grupos a partir de uma partição inicial (XU; WUNSCH, 2005).

Segundo Jain (2010), um dos algoritmos de agrupamento mais conhecido é o K-médias. Esse método foi proposto de diferentes modos e suposições por alguns autores, como referências clássicas pode-se citar Lloyd (1982), MacQueen (1967) e Hartigan e Wong (1979). Porém, o K-médias apresenta algumas desvantagens, dentre elas a de ser influenciado por valores discrepantes dos conjuntos de dados. Com a finalidade de solucionar essa limitação García-Escudero e Gordaliza (1999) propuseram o K-médias aparado, que traz na sua abor-

dagem uma aplicação do K-médias em um subconjunto dos dados sem os *outliers*. Outro algoritmo de agrupamento comumente conhecido é o *PAM* (*Partitioning Around Medoids*) (ROUSSEEUW; KAUFMAN, 1990), o qual utiliza um objeto representativo localizado mais ao centro do grupo, conhecido como medoide. Mas, o PAM não é adequado para trabalhar com grandes conjuntos de dados, neste contexto Ng e Han (2002) desenvolveram o método *CLARANS* (*Clustering Algorithm based on Randomized Search*). Segundo Everitt et al. (2011), os algoritmos mencionados acima são utilizados para o agrupamento de dados pertencentes a espaços euclidianos.

Em várias situações, faz-se necessário o agrupamento de objetos pertenceres ao espaço de formas, os quais são espaços não-euclidianos. Neste sentido, diversos autores adaptaram algoritmos para o agrupamento de formas. Considerando formas bidimensionais, pode-se citar o trabalho de Georgescu (2009) que realiza a adaptação do método K-médias para o agrupamento de formas difusas e o trabalho de Amaral et al. (2010) que realiza a adaptação da versão de Hartigan-Wong do K-médias. No que envolve técnicas de aprendizado de máquina, Assis (2018) realiza a adaptação de métodos de busca, como o *Hill Climbing* (FRIEDMAN; RUBIN, 1967), para a formação de grupos. Porém, no cenário de formas tridimensionais, não existem muitos trabalhos envolvendo adaptações de técnicas para a formação de grupos de dados de formas 3D. Um dos trabalhos mais significativos nesse quesito foi desenvolvido por Vinué, Simó e Alemany (2014), o qual apresenta a adaptação das versões de Lloyd do K-médias e do K-médias aparado. E Teotonio e Amaral (2020) realiza o ajuste da versão de Lloyd do K-médias utilizando a média ϕ. A média ϕ foi desenvolvida por Dryden et al. (2008) com a finalidade de encontra a forma média fechada para formas de objetos com mais de duas dimensões.

Esta dissertação tem como objetivo adaptar e analisar os seguintes algoritmo no contexto de formas tridimensionais: K-médias aparado utilizando a média ϕ, *CLARANS* e *Hill Climbing*. Assim como também apresentar e analisar os métodos já implementados por outros autores, como o K-médias, K-médias aparado e K-médias utilizando a média ϕ. As qualidades dos grupos formados foram avaliadas usando técnicas de validação de agrupamento. As medidas utilizadas foram os índices de Rand Ajustado (HUBERT; ARABIE, 1985) e Silhueta (ROUSSEEUW, 1987).

Desenvolvimentos tecnológicos têm facilitado a obtenção de informações geométricas de dados de formas. No meio computacional, a linguagem de programação R (R Core Team, 2020) dispõe de alguns pacotes que fornecem funções para lidar com esses tipos de dados. Alguns

pacotes são o **shapes** (DRYDEN, 2012), **Anthropometry**, (VINUÉ et al., 2020), **Morpho** (SCHLAGER et al., 2020) e o **geomorph** (ADAMS; OTÁROLA-CASTILLO, 2013). Os desempenhos dos algoritmos foram analisados por intermédio de aplicações em dois experimentos com dados simulados e em três aplicações de conjuntos de dados reais pertencentes as literaturas da área. Nos estudos do presente trabalho utilizou-se um notebook Aspire A315-41, com processador Ryzen 3 2200U, 2,50 GHz, 4GB de RAM, sistema de 64-bit e plataforma Windows 10 Home.

Por fim, é importante desenvolver as técnicas de análise de formas para tornar possível solucionar diversos problemas de diferentes campos de pesquisas, os quais fazem uso de dados geométricos de formas. É de grande relevância adaptar os mais distintos métodos estatísticos na conjuntura de dados de formas e alguns conceitos/definições já implementados no meio computacional pode auxiliar nesse processo.

1.2 OBJETIVOS

- Objetiva-se descrever alguns conceitos e definições da Análise Estatística de Formas e da Análise de Agrupamento.
- Descrever e analisar o algoritmo K-médias (LLOYD, 1982) adaptado por Vinué, Simó e Alemany (2014) para o agrupamento de formas 3D e a adaptação realizada por Teotonio e Amaral (2020) utilizando a média ϕ.
- Descrever e analisar o algoritmo K-médias aparado proposto por García-Escudero e Gordaliza (1999) e adaptado por Vinué, Simó e Alemany (2014) para o agrupamento de formas 3D.
- Adaptar e analisar o K-médias aparado fazendo uso da média ϕ.
- Descrever e analisar os métodos de agrupamentos *CLARANS* (NG; HAN, 2002) e *Hill Cllmbing* (FRIEDMAN; RUBIN, 1967) e suas adaptações no contexto de formas 3D.
- Fazer uso das técnicas de validação de agrupamento para avaliar os desempenho dos algoritmos em diferentes conjuntos de dados.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

A dissertação divide-se em cinco Capítulos, sendo o primeiro um Capítulo introdutório. No Capítulo 2 é apresentada uma breve descrição dos conceitos de interesse relacionados a análise de formas. No Capítulo 3 é apresentado um resumo sobre os métodos de agrupamentos e uma descrição dos algoritmos K-médias e sua variação K-médias aparado, *CLARANS* e *Hill Climbing* adaptados para o contexto de agrupamento de formas 3D. Além disso, é mostrado uma visão geral sobre os índices de validação Rand Ajustado e Silhueta. No Capítulo 4 é avaliado o desempenho dos algoritmos em aplicações nos dados simulados e reais. Por fim, no Capítulo 5 são apresentadas as conclusões acerca dos estudos realizados e algumas sugestões para trabalhos futuros na mesma área.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo, serão apresentadas algumas definições e conceitos de interesse relacionados à Análise Estatística de Formas.

2.1 ANÁLISE ESTATÍSTICA DE FORMAS (*STATISTICAL SHAPE ANALYSIS*)

Um tópico da Estatística, utilizado para trabalhar com formas e informações geométricas relacionadas a um objeto, como particularidades e tomada de decisões sobre o mesmo, é a Análise Estatística de Formas (*Statistical Shape Analysis*). Tendo o seu início em aplicações na Astronomia, Arqueologia e Biologia (KENDALL et al., 1999), atualmente várias áreas de conhecimento fazem uso das técnicas pertencentes a análise de formas, como: Medicina, Genética, Geografia, Geologia, Agricultura e nas últimas décadas vem sendo bastante utilizada em Reconhecimento de Padrões e Visão Computacional. Segundo Kendall (1977), forma é a informação geométrica de um objeto que resta quando são removidos os efeitos de locação, escala e rotação. Ao decorrer deste Capítulo serão apresentados os principais conceitos e definições de interesse relacionados à Análise Estatística de Formas.

2.2 MARCOS ANATÔMICOS (*LANDMARKS*) E MATRIZ DE CONFIGURAÇÃO

2.2.1 Marcos Anatômicos

Tanto no cotidiano quanto em algumas aplicações científicas, a descrição de uma forma pode levar a uma ideia abstrata da mesma, o que dificulta a obtenção de informações geométricas. Assim, um modo de caracterizar uma forma é detectar um conjunto finito de pontos em torno da silhueta de um objeto. A esse conjunto de pontos é dado o nome de marcos e a posição desses pontos relaciona-se com as coordenadas cartesianas.

Segundo Dryden e Mardia (2016), um marco é "um ponto de relação que corresponde entre e dentro das populações", ou seja, todos os indivíduos são homólogos (possuem os mesmos marcos e quantidades). Um ponto pode ser especificado por questões geométricas ou por um especialista e a quantidade de pontos não é limitada por uma restrição teórica (KENDALL et al., 1999).

Os marcos são definidos em três tipos (DRYDEN; MARDIA, 2016):

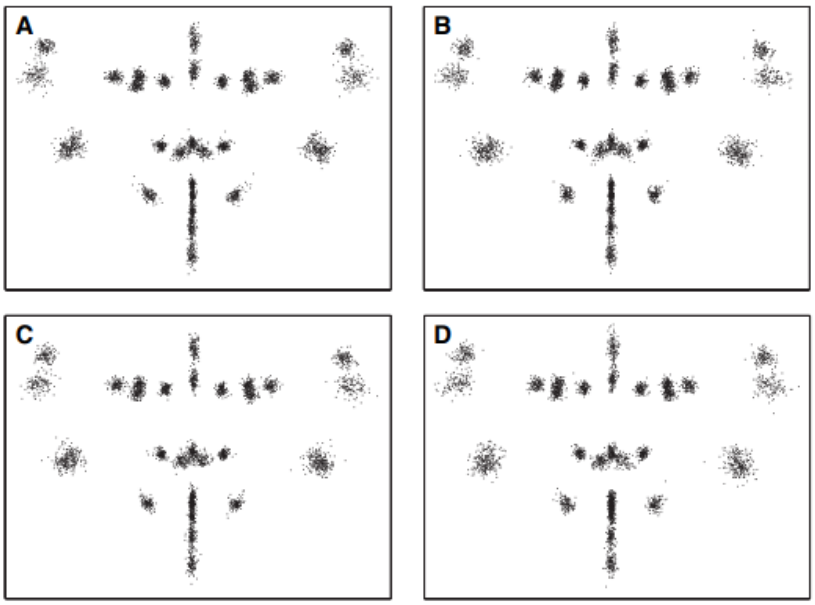
1. Marcos anatômicos: são pontos de um objeto escolhidos por um especialista, os quais possuem alguma característica significante.

2. Marcos matemáticos: são pontos determinados de acordo com alguma propriedade matemática ou geométrica da imagem do objeto.

3. Pseudo-marcos: são pontos atribuídos em um objeto, localizados ao redor do contorno entre os marcos anatômicos e/ou matemáticos. Geralmente são adicionados para descrever a forma do objeto.

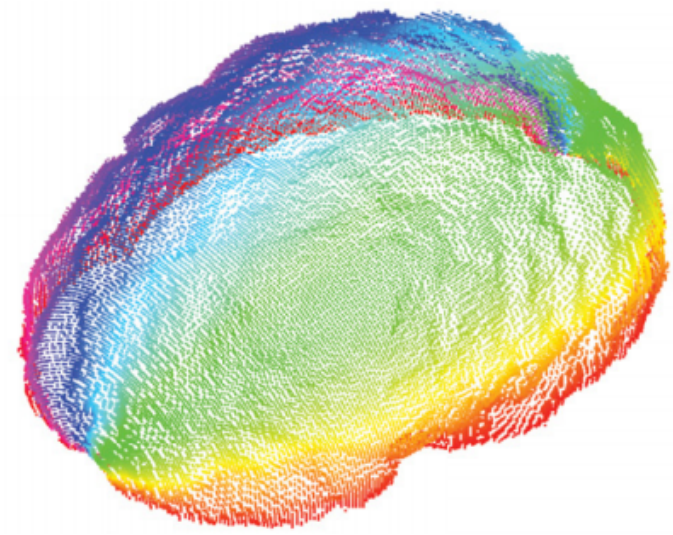
Na Figura 1, pode-se notar a representação de 30 marcos anatômicos de dados de formas do estudo sobre a diferença entre a forma média de rostos humanos em diferentes grupos de idade (EVISON; BRUEGGE, 2008); (PRESTON; WOOD, 2010). Já na Figura 2 tem-se a representação em 3D de 62.501 marcos anatômicos da superfície do córtex cerebral humano, obtidos para um estudo para verificar diferenças na forma do cérebro de pacientes com esquizofrenia e voluntários sem esquizofrenia (BRIGNELL et al., 2010); (DRYDEN; MARDIA, 2016).

Figura 1 – Visualização de uma amostra aleatória de $n = 125$ observações dos dados de formas do rosto humano com $k = 30$ marcos anatômicos para indivíduos com idades entre (A) 15-24, (B) 25-34, (C) 35-44 e (D) $+$ 45 anos



Fonte: (PRESTON; WOOD, 2010).

Figura 2 – Conjunto de k=62.501 marcos anatômicos para representação da superfície do córtex de um cérebro em 3D



Fonte: (DRYDEN; MARDIA, 2016).

2.2.2 Matriz de Configuração

Ao digitalizar os marcos anatômicos por meio de um aparelho de medição, nota-se que os objetos possuem tamanhos, posições e rotações diferentes um do outro. A fim de obter uma padronização desses objetos para realizar uma análise dos dados de modo mais completo, é necessário retirar os efeitos de locação, escala e rotação. Na próxima seção será mostrado o procedimento para a realização da exclusão desses efeitos.

O avanço tecnológico, como a digitalização de objetos, vem tornando-se um aliado bastante relevante na medição de marcos anatômicos e a ideia central da abordagem geométrica da análise de formas é a utilização da representação do próprio objeto (DRYDEN; MARDIA, 2016). Um sistema de coordenadas deve ser especificado com a finalidade de descrever a forma de um objeto. Existem diversos sistemas, como: coordenadas de Bookstein (BOOKSTEIN, 1984), coordenadas QR de Goodall-Mardia (GOODALL; MARDIA, 1993) e as coordenadas polares de Kent (KENT, 1994). As coordenadas que serão utilizadas no presente trabalho são as coordenadas propostas por Kendall (1984) e um dos principais pontos desse trabalho foi a definição dos sistemas de coordenadas. Utiliza-se as coordenadas de Kendall por ser comumente usada na literatura da área. Estes sistemas possuem finalidade de obter a forma de um objeto por meio dos pontos dispostos através dos marcos.

Uma configuração é uma coleção de marcos em um determinado objeto, sendo representada matematicamente por uma matriz $\mathbf{X}$ de dimensão $k\ x\ m$ de coordenadas cartesianas de k

marcos em m dimensões. O espaço de todas as coordenadas possíveis dos marcos é denominado de espaço de configuração. A matriz de configuração $\mathbf{X}$ é definida por:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & \dots & x_{k,m} \end{bmatrix}. \tag{2.1}$$

Nesta dissertação serão estudadas as formas de objetos com três dimensões, ou seja, os casos em que $m = 3$. Dessa maneira, a matriz de configuração $\mathbf{X}$ passa a ser estabelecida como:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ \vdots & \vdots & \vdots \\ x_{k,1} & x_{k,2} & x_{k,3} \end{bmatrix}. \tag{2.2}$$

2.3 ESPAÇO DE FORMA E PRÉ-FORMA

A matriz de configuração $\mathbf{X}$ não descreve de modo adequado um objeto, pois, não é invariável sob transformações de similaridade. Desse modo, tem-se que eliminar os efeitos de locação, escala e rotação por meio de transformações em $\mathbf{X}$, ou seja, $\mathbf{X}$ tem que tornar-se uma matriz invariável. A seguir, uma definição do espaço de forma é apresentada, segundo Dryden e Mardia (2016).

Definição 1 *O **espaço de forma** é o espaço de todas as formas. O espaço de forma é o conjunto de classes de equivalência das matrizes de configuração sob ação de transformações de similaridade euclidiana (translação, rotação e escala).*

O espaço de forma admite uma estrutura de variedade Riemanniana, ou seja, métodos estatísticos padrões não podem ser utilizados. As propriedades geométricas do espaço de forma são discutidas de modo mais profundo pelos seguintes autores: Kendall (1984), Le e Kendall (1993), Kent e Mardia (2001), Kendall et al. (1999) e Small (2012). Segundo Dryden e Mardia (2016), uma estrutura Riemanniana é um espaço que pode ser visualizado como um

espaço euclidiano e que possui um produto interno positivo definido em cada espaço tangente de maneira que a escolha varia sutilmente de ponto a ponto.

A dificuldade de trabalhar com uma variedade Riemanniana depende dos valores de k e m. Para $m \geq 3$ dimensões, o espaço de forma não é uma variedade Riemanniana, mas sim um espaço estratificado (DRYDEN; MARDIA, 2016). Esse espaço possui singularidades e não são espaços familiares. Porém, assume-se que, na prática estamos afastados dessas singularidades e nos limitamos a uma variação do espaço de forma. De modo diferente, quando $m = 2$, o espaço de forma é um espaço projetivo complexo.

Sabe-se que a forma de um dado objeto são as informações geométricas restantes após a remoção dos efeitos de locação, escala e rotação. Inicialmente, para remover o efeito de locação, serão utilizadas as coordenadas de Kendall (1984), as quais fazem uso da submatriz de Helmert $\mathbf{H}$. A matriz de Helmert $\mathbf{H}^F$, é uma matriz ortogonal de dimensão $k\ x\ k$ em que todos os elementos da primeira linha são iguais a $1/\sqrt{k}$ e a *j-ésima* linha possui $(j-1)$ elementos iguais a $-j(j+1)^{\frac{1}{2}}$ seguida por um elemento igual a $(j-1) * (j(j+1))^{\frac{1}{2}}$ e $(k-j)$ zeros. Dado $\mathbf{H}^F$, a submatriz de Helmert é a matriz de Helmert sem a primeira linha. Exemplificando, para $k = 3$, a matriz de Helmert é dada por:

$$\mathbf{H}^F = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \end{bmatrix}, \tag{2.3}$$

e a submatriz de Helmert é:

$$\mathbf{H} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \end{bmatrix}. \tag{2.4}$$

Definida a submatriz $\mathbf{H}$, o efeito de locação é removido multiplicando a matriz de configuração $\mathbf{X}$ pela submatriz de Helmert de dimensão $(k-1)\ x\ k$. Desse modo:

$$\mathbf{X_H} = \mathbf{HX}, \tag{2.5}$$

em que $\mathbf{X_H}$ possui dimensão $(k-1)\ x\ m$. A matriz de centralização $\mathbf{C}$ de dimensão $k\ x\ k$ é definida por:

$$\mathbf{C} = \mathbf{I_k} - \frac{1}{k}\mathbf{1}, \tag{2.6}$$

em que $\mathbf{I_k}$ é a matriz identidade e $\mathbf{1}$ é uma matriz de uns, ambas possuem dimensão $k\ x\ k$.

Como $\mathbf{H}^F$ é uma matriz ortogonal, existe uma relação entre as configurações helmertizadas (2.5) com as configurações centradas (2.6) (DRYDEN; MARDIA, 2016):

$$\mathbf{H}^T\mathbf{H} = \mathbf{C} = \mathbf{I_k} - \frac{1}{k}\mathbf{1}, \tag{2.7}$$

e

$$\mathbf{H}^T\mathbf{X_H} = \mathbf{H}^T\mathbf{HX} = \mathbf{CX}, \tag{2.8}$$

em que $\mathbf{H}^T$ é a transposta da sub-matriz de Helmert.

Para remover o efeito de escala, tem-se que dividir a configuração helmertizada por sua norma, assim:

$$\mathbf{Z} = \frac{\mathbf{X_H}}{||\mathbf{X_H}||} = \frac{\mathbf{HX}}{||\mathbf{HX}||}, \tag{2.9}$$

em que $\mathbf{Z}$ é denominado de pré-forma da matriz de configuração $\mathbf{X}$. É importante notar que na pré-forma os efeitos de rotação permanecem. Uma outra representação da pré-forma pode ser obtida utilizando a matriz $\mathbf{C}$, dessa maneira, obtêm-se as pré-formas centralizadas:

$$\mathbf{Z_C} = \frac{\mathbf{CX}}{||\mathbf{CX}||}, \tag{2.10}$$

de dimensão $k\ x\ m$ e desde que $\mathbf{C} = \mathbf{H}^T\mathbf{H}$. Ainda, $||\mathbf{CX}||$ é definido como tamanho do centroide.

Definição 2 *O espaço de pré-forma S_m^k é o conjunto de todas as pré-formas possíveis.*

Dado que o espaço de pré-forma possui dimensão real $(k-1)(m-1)$, as representações (2.9) e (2.10) são consideradas apropriadas para este espaço. A vantagem em utilizar $\mathbf{Z}$ é que seu número de linhas é menor e a vantagem de utilizar $\mathbf{Z_C}$ é que a sua representação gráfica das coordenadas cartesianas se ajusta com a sua configuração original, embora tanto $\mathbf{Z}$ quanto $\mathbf{Z_C}$ possuam o mesmo *rank* (DRYDEN; MARDIA, 2016).

Para remover o efeito de rotação, tem-se que multiplicar a pré-forma $\mathbf{Z}$ por uma matriz $\mathbf{\Gamma}$ denominada de matriz de rotação de dimensão $m\ x\ m$. Segundo Dryden e Mardia (2016), uma matriz de rotação possui $\frac{1}{2}m(m-1)$ graus de liberdade.

Definição 3 *Uma matriz de rotação Γ, também conhecida como matriz ortogonal especial, satisfaz $\Gamma^T\Gamma = \Gamma\Gamma^T = \mathbf{I}$ e $det(\Gamma) = +1$. O conjunto de todas matrizes de rotação é conhecido como grupo ortogonal especial $SO(m)$.*

Dada a Definição 3, considere:

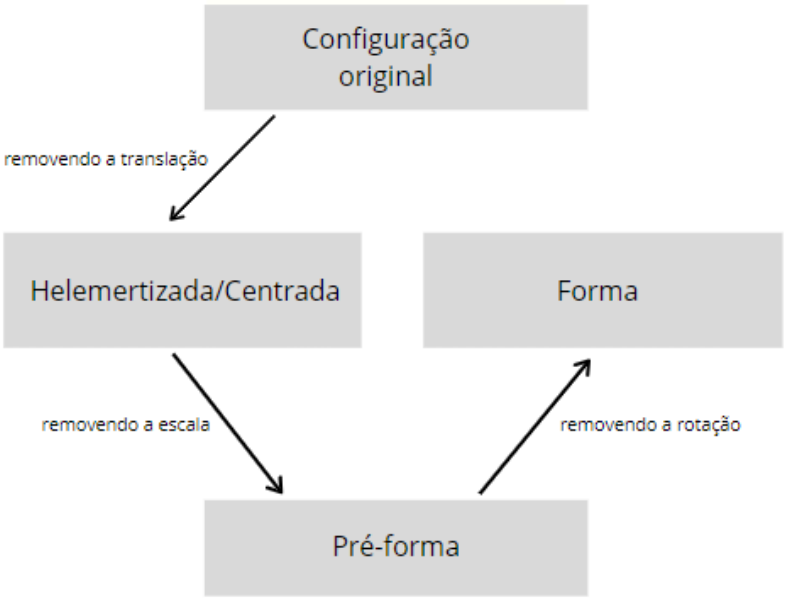
Definição 4 *A forma de uma matriz de configuração é toda informação geométrica que resta após a remoção dos efeitos de locação, escala e rotação. A forma pode ser representada por:*

$$[\mathbf{Z}] = \mathbf{Z}\Gamma \ : \ \Gamma \in SO(m), \tag{2.11}$$

em que Γ *é uma matriz de rotação,* $\mathbf{Z}$ *é a pré-forma e* $SO(m)$ *é o grupo ortogonal de rotações.*

Segundo Vinué, Simó e Alemany (2014), tem-se que S_m^k é uma hiperesfera de raio unitário nos $\mathbb{R}^{k-1}$, ou seja, é uma subvariedade Riemanniana comumente estudada e conhecida. E para $m > 2$, Σ_m^k não é um espaço habitual. Dado que Σ_m^k é considerado um espaço quociente de S_m^k sob rotações, é mais descomplicado e intuitivo trabalhar nesse espaço, por se tratar de uma subvariedade Riemanniana. A Figura 3 ilustra o esquema para obtenção da forma de um objeto.

Figura 3 – Esquema de obtenção da forma



Fonte: Elaborada pelo o autor (2021).

2.4 DISTÂNCIAS PARA O ESPAÇO DE FORMA

Anteriormente foi mencionado que o espaço de forma admite uma estrutura de variedade Riemanniana. Na prática, estamos constantemente interessados em comparar e medir objetos. Nesse contexto faz-se necessário definir conceitos de distâncias voltados para a análise de formas.

2.4.1 Distâncias de Procrustes

Considere duas matrizes de configuração de k marcos em m dimensões $\mathbf{X_1}$ e $\mathbf{X_2}$, com pré-formas iguais a $\mathbf{Z_1}$ e $\mathbf{Z_2}$. A seguir, serão apresentadas as definições da distância parcial de Procrustes e da distância total de Procrustes, respectivamente.

Definição 5 *A distância parcial de Procrustes entre* $\mathbf{X_1}$ *e* $\mathbf{X_2}$ *é:*

$$d_P(\mathbf{X_1}, \mathbf{X_2}) = \inf_{\Gamma \in SO(m)} ||\mathbf{Z_2} - \mathbf{Z_1}\Gamma||,$$

em que $\mathbf{Z_r}$, $r = 1, 2$, *representa a pré-forma de acordo com (2.9),* Γ *é uma matriz de rotação pertencente ao espaço* $SO(m)$ *e* $\inf$ *denota o mínimo.*

Resultado 1 *A distância parcial de Procrustes é:*

$$d_P(\mathbf{X_1}, \mathbf{X_2}) = \sqrt{2}\left(1 - \sum_{i=1}^{m} \lambda_i\right)^{1/2}, \tag{2.12}$$

em que $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_{m-1} \geq |\lambda_m|$ *são as raízes quadradas dos autovalores de* $\mathbf{Z_1}^T\mathbf{Z_2}\mathbf{Z_2}^T\mathbf{Z_1}$ *e o menor valor* λ_m *é o negativo da raiz quadrada se e somente se* $det(\mathbf{Z_1}^T\mathbf{Z_2}) < 0$.

Definição 6 *A distância total de Procrustes entre* $\mathbf{X_1}$ *e* $\mathbf{X_2}$ *é:*

$$d_F(\mathbf{X_1}, \mathbf{X_2}) = \inf_{\Gamma \in SO(m),\ \beta \in \mathbb{R}^+} ||\mathbf{Z_2} - \beta\mathbf{Z_1}\Gamma||,$$

em que β *é um escalar.*

Resultado 2 *A distância total de Procrustes é:*

$$d_F(\mathbf{X_1}, \mathbf{X_2}) = \left(1 - (\sum_{i=1}^{m} \lambda_i)^2\right)^{1/2}. \tag{2.13}$$

O termo Procrustes é utilizado porque as correspondências apresentadas para essas distâncias são semelhantes às técnicas utilizadas na Análise Multivariada, denominadas de Análise de Procrustes. As distâncias acima são consideradas extrínsecas, pois são distâncias em uma incorporação do espaço de forma, para mais detalhes veja (DRYDEN; MARDIA, 2016).

2.4.2 Distância Riemanniana

Na esfera de pré-forma, a forma de uma matriz de configuração é representada por uma fibra. Uma fibra é a pré-forma rotacionada na esfera de pré-forma. A distância Riemanniana

é considerada como o comprimento do caminho geodésico mínimo entre duas formas. Tem-se que no espaço euclidiano, o caminho geodésico é apenas uma linha reta, porém em espaços não-euclidianos o caminho geodésico é uma curva (DRYDEN; MARDIA, 2016).

Resultado 3 *A distância Riemanniana é:*

$$\rho(\mathbf{X_1}, \mathbf{X_2}) = arc\ cos(\sum_{i=1}^{m} \lambda_i), \tag{2.14}$$

em que λ_i, $i = 1, \ldots, m$, *foi definido em (2.12).*

A distância Riemanniana é uma distância intrínseca, ou seja, é definida no próprio espaço de forma Σ_m^k. Já as distâncias de Procrustes são consideradas distâncias extrínsecas, ou seja, são definidas como distâncias em uma incorporação do espaço de forma. As apresentações das provas dos resultados relacionados a este Capítulo podem ser vistas com mais detalhes em Dryden e Mardia (2016). Um resumo sobre as distâncias apresentadas pode ser visualizado na Tabela 1.

As relações entre as distâncias d_P, d_F e ρ são:

$$d_P(\mathbf{X_1}, \mathbf{X_2}) = 2sin\ (\rho/2), \tag{2.15}$$

e

$$d_F(\mathbf{X_1}, \mathbf{X_2}) = sin\ (\rho). \tag{2.16}$$

Vinué, Simó e Alemany (2014) comentam que a distância ρ (uma métrica Riemanniana) e a distância d_F (uma métrica não-Riemanniana) são topologicamente equivalentes.

Tabela 1 – Resumo sobre as distâncias no espaço de formas

Distância	Notação	Fórmula	Suporte
Distância parcial de Procrustes	d_P	$\sqrt{2}\left(1 - \sum_{i=1}^{m} \lambda_i\right)^{1/2}$	$0 \leq d_P \leq \sqrt{2}$
Distância total de Procrustes	d_F	$\left(1 - (\sum_{i=1}^{m} \lambda_i)^2\right)^{1/2}$	$0 \leq d_F \leq 1$
Distância Riemanniana	ρ	$arc\ cos(\sum_{i=1}^{m} \lambda_i)$	$0 \leq \rho \leq \pi/2$

Fonte: (DRYDEN; MARDIA, 2016).

2.5 FORMA MÉDIA

Na Análise Estatística de Formas, a definição do conceito de forma média executa um papel importante para a análise dos dados. Porém, em espaços não-euclidianos, não existe um conceito de média equivalente à média aritmética comumente conhecida.

2.5.1 Forma média de Procrustes

Uma média do tipo Fréchet (FRÉCHET, 1948), é uma média que minimiza a soma das distâncias ao quadrado de qualquer forma. Considere $\mathbf{X_1}, \ldots, \mathbf{X_n}$ um conjunto de matrizes de configuração.

Definição 7 *A forma média total de Procrustes no espaço de forma é dada por:*

$$[\hat{\mu}] = \arg\inf_{\mu} \sum_{i=1}^{n} {d_F}^2(\mathbf{X_i}, \mu), \tag{2.17}$$

em que d_F representa a distância apresentada em (2.13).

Para dados bidimensionais, em que $m = 2$, Kent (1994) apresenta como resultado um autovetor de solução explícita do problema de otimização na Definição 7 de forma média. Porém, para dimensões em que $m \geq 3$, um processo iterativo deve ser utilizado, pois o procedimento de correspondência não pode ser escrito como uma expressão linear. O processo iterativo pode ser visualizado no Algoritmo 1.

Para fazer uso da média de Procrustes no R, basta extrair o objeto *mshape* da função *procGPA* do pacote **shapes**.

2.5.2 Forma média ϕ

Uma outra maneira de calcular a forma média para dados com $m \geq 3$ foi proposta por Dryden et al. (2008). Tal alternativa trabalha com o espaço $P_m(k-1)$ de matrizes simétricas diferentes de zero, de $rank$ no máximo m, traço igual 1 e de dimensão $(k-1)\ x\ (k-1)$. Para uma determinada pré-forma $\mathbf{Z} \in S_m^k$ com uma distribuição arbitrária da população F, escreve-se:

$$\Xi(F) = \mathbb{E}_F(\mathbf{Z}\mathbf{Z}^T) = \int_{\mathbf{Z} \in S_m^k} \mathbf{Z}\mathbf{Z}^T dF(\mathbf{Z}) \tag{2.18}$$

Algoritmo 1: Algoritmo para o cálculo da forma média de Procrustes $[\hat{\mu}]$.

1. Centralize as configurações parar remover os efeitos de locação. Seja $\mathbf{X}$ uma matriz de configuração. Inicialmente, considere:

$$\mathbf{X_i}^P = \mathbf{C}\mathbf{X_i},$$

em que $i = 1, \ldots, n$ e $\mathbf{C}$ é a matriz de centralização apresentada em (2.6).

2. Calcule $G = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}||\mathbf{X_i}^P - \mathbf{X_j}^P||^2$. Para a *i-ésima* iteração faça:

$$\tilde{\mathbf{X}}_{(i)} = \frac{1}{n-1}\sum_{j\neq i}\mathbf{X_i}^P.$$

Otimize $||\tilde{\mathbf{X}}_{(i)} - \mathbf{X_i}^P\Gamma||^2$ por rotações. Seja $\mathbf{X_i}^P = \mathbf{X_i}^P\hat{\Gamma}$, em que $\hat{\Gamma}$ é a matriz de rotação ótima. Repita para todo i. Calcule o novo valor de G. O processo se repete até que G não possa mais ser reduzido.

3. Para a *i-ésima* configuração, calcule:

$$\hat{\beta}_i = \left(\frac{\sum_{k=1}^{n}||\mathbf{X_k}^P||^2}{||\mathbf{X_i}^P||^2}\right)^{\frac{1}{2}}\phi_i,$$

em que ϕ_i é o *i-ésimo* componente do autovetor ϕ correspondente ao maior autovalor da matriz de correlação Φ de $vec\{\mathbf{X_i}^P\}$.
Considere $\mathbf{X_i}^P = \hat{\beta}_i\mathbf{X_i}^P$. Repita para todo i. Calcule o novo valor de G.

4. Repita os passos 2 e 3 até que G não possa mais ser reduzido.
5. Assim, $[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i}^P$ é a forma média.

como a esperança de $\mathbf{Z}\mathbf{Z}^T$ em relação a F. Suponha que $\Xi = \Xi(F)$ possua a seguinte decomposição espectral $\Xi = \mathbf{U}\Lambda\mathbf{U}^T$, em que $\Lambda = diag(\delta_1, \ \ldots \ , \delta_{k-1})$ são os autovalores ordenados $\delta_1 \geq \cdots \geq \delta_{k-1} \geq 0$ e as colunas de $\mathbf{U} = (u_1, \ldots, u_{k-1})$ são os autovetores correspondentes. Dessa maneira, tem-se que:

$$\phi(\Xi) = \frac{1}{\delta_1 + \cdots + \delta_m}\ \sum_{i=1}^{m}\delta_i u_i u_i^T, \tag{2.19}$$

desde que $\delta_m > \delta_{m+1}$.

A Equação (2.19) é chamada de média ϕ de $\mathbf{Z}$. Dada uma amostra $\mathbf{Z_1}$, ..., $\mathbf{Z_n}$, a média ϕ da amostra é denotada por $\phi(\hat{\Xi})$, em que $\hat{\Xi} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z_i}\mathbf{Z_i}^T$ é o análogo da amostra de Ξ.

Em suma, segundo Preston e Wood (2010), essa definição de média é motivada pelos conceitos de escalonamento multidimensional e por conta da relação entre formas e elementos do espaço $P_m(k-1)$, tem-se que $\phi(\Xi)$ pode ser visualizado como uma forma, dado que sua abordagem baseia-se na incorporação do espaço de forma de reflexão (que é o espaço quociente por uma reflexão do espaço de forma) ao espaço $P_m(k-1)$, ou seja, existe um mapeamento um a um entre as possíveis formas de uma matriz de configuração e os elementos do espaço

$P_m(k-1)$. Dryden et al. (2014) mostraram que a média (2.19) faz parte da família de um parâmetro de projeções indexado por $\alpha \geq 2$. Ainda, Dryden e Mardia (2016) comentam que a média ϕ é menos resistente a *outliers* do que a média de Procrustes. Para fazer uso dessa média no R utiliza-se a função *MDSshape* do pacote **shapes**.

3 ANÁLISE DE AGRUPAMENTO NO ESPAÇO DE FORMAS

Neste Capítulo, serão apresentadas algumas definições e conceitos de interesse relacionados à Análise de Agrupamento, assim como também a apresentação e definição dos algoritmos e das técnicas de Validação de Agrupamento, ambos mencionados no Capítulo 1.

3.1 ANÁLISE DE AGRUPAMENTO

A Análise de Agrupamento tem como objetivo classificar os indivíduos de um determinado conjunto de dados em grupos, de tal maneira que os indivíduos de cada grupo sejam semelhantes entre si e os grupos sejam diferentes. A classificação é considerada uma das atividades mais primitiva praticada pelos seres humanos (XU; WUNSCH, 2005). Diversas podem ser as razões para classificar, tais como o auxílio na organização dos dados ou a compreensão do fenômeno estudado. A apresentação dos métodos de agrupamento são bastantes disseminados nas áreas de Análise Multivariada e Aprendizado de Máquina.

No contexto de aprendizado de máquina, James et al. (2013) comentam que os problemas podem ser divididos em duas categorias: supervisionada e não supervisionada. Para os casos supervisionados, geralmente tem-se x_i, $i = 1, \cdots, n$ variáveis explicativas associadas a uma variável resposta e deseja-se ajustar um modelo que relacione as variáveis explicativas à variável resposta, para os fins de predições ou inferências sobres os dados. Como exemplos para a abordagem supervisionada, pode-se citar regressão linear e regressão logística. Em contraposição, a abordagem não supervisionada, para cada objeto $i = 1, \cdots, n$ pertencente ao conjunto de dados, observa-se x_i, $i = 1, \cdots, n$ medidas associadas a esses objetos, mas nenhuma variável resposta associada. Desse modo, não é possível ajustar um modelo e o termo não supervisionado é utilizado pois falta a variável resposta. Umas das metodologias na abordagem não supervisionada é encontrar padrões e relações entre os objetos do conjunto de dados. Para os cenários da presente dissertação, utiliza-se a abordagem não supervisionada desejando a formação de grupos contendo os objetos selecionados. Desse modo, utiliza-se os algoritmos que serão apresentados no decorrer deste Capítulo.

Várias áreas de conhecimento, tais como Ciências Sociais, Biologia, Medicina e Reconhecimento de Padrões fazem uso das metodologias da Análise de Agrupamento. Antes de proceder com a análise de agrupamento, é necessário definir e calcular medidas de similaridade

ou dissimilaridade, isto é, conceitos de distância. Segundo Goshtasby (2012), considere S uma determinada métrica, se o valor de S retornar um valor alto conforme a dependência entre os valores das observações aumenta, então S é dita ser uma medida de similaridade. No entanto, se S retornar um valor baixo conforme a dependência entre os valores das observações diminui, então S é dita ser uma medida de dissimilaridade. Basicamente, comparando dois objetos, tem-se similaridade quando duas observações são mais semelhantes entre si e o valor da distância entre elas é alto e tem-se dissimilaridade quando duas observações são menos semelhantes entre si e o valor da distância entre elas é alto.

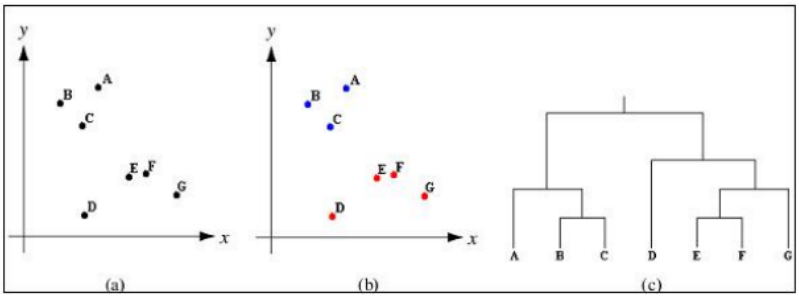
A definição de distância é essencial para proceder com a análise de agrupamento e diferentes distâncias podem produzir diferentes tipos de grupos. As métricas (distâncias) utilizadas nesta dissertação são todas medidas de dissimilaridade. Os métodos de agrupamento diferem de várias maneiras e Everitt et al. (2011) comentam que os dois principais métodos são os hierárquicos e os particionais.

Os métodos hierárquicos são utilizados quando deseja-se determinar/identificar os possíveis grupos a serem gerados de acordo com uma estrutura hierárquica da relação de proximidade entre os indivíduos, ou seja, os números de grupos não é pré-definido. Os algoritmos hierárquicos dividem-se em duas abordagens: divisiva e aglomerativa. Na abordagem divisiva, inicialmente todos os indivíduos estão em um único grupo, em seguida, os grupos são formados de acordo com um processo iterativo, até que cada indivíduo torna-se um único grupo. Nos métodos aglomerativos, de modo contrário, inicialmente tem-se n grupos e a cada passo do processo iterativo, combinações entre os grupos iniciais ocorrem até que todos os indivíduos estejam em um determinado grupo (XU; WUNSCH, 2005). Graficamente, os grupos formados pelos métodos hierárquicos são representados normalmente por dendrogramas, em que no eixo horizontal têm-se os indivíduos e no eixo vertical os valores da distância utilizada.

Por outro lado, os métodos de particionamento formam grupos a partir de uma partição inicial e do número K $(K \leq n)$ de grupos pré-definido. Inicialmente tem-se uma partição inicial e de acordo com o processo do algoritmo, os elementos mudam de grupo em grupo até que os formatos desses grupos cheguem a um agrupamento (partição) final. A partição inicial pode ser definida por diferentes maneiras, como indicação de especialistas, resultado de um método de agrupamento prévio ou obtida de modo arbitrário (EVERITT et al., 2011); (FRIEDMAN; RUBIN, 1967). Segundo Marriott (1982), os resultados dos grupos formados podem ser sensíveis à partição inicial. Os métodos de particionamento dependem de funções objetivos (geralmente associadas a uma distância) e dividem-se em dois tipos: rígidos (*hard*) e difusos (*fuzzy*). Nos

algoritmos do tipo rígido, cada elemento do conjunto de dados pertence exclusivamente a um grupo, ou seja, não existem grupos que contenham o mesmo indivíduo e cada partição representa um grupo (XU; WUNSCH, 2005). Já nos algoritmos do tipo difuso, os indivíduos podem pertencer a mais de um grupo de acordo com graus de pertinência obtidos por meio de uma função. A função de pertinência é capaz de realizar a associação de padrões a cada um dos K-grupos, assumindo valores no intervalo $[0, 1]$ (BEZDEK, 1981); (ZADEH, 1996). Na Figura 4 tem-se uma apresentação da diferença dos métodos hierárquicos dos métodos particionais: (a) Estrutura original de um conjunto de objetos com duas variáveis; (b) Resultado da aplicação de um método de particionamento; (c) Resultado da aplicação de um método hierárquico. Na Figura 5 tem-se uma ilustração de agrupamentos particionais do tipo rígido e difuso. Os retângulos H_1 e H_2 dividem o conjunto de objetos em dois grupos rígidos, enquanto que as elipses F_1 e F_2 dividem em dois grupos difusos.

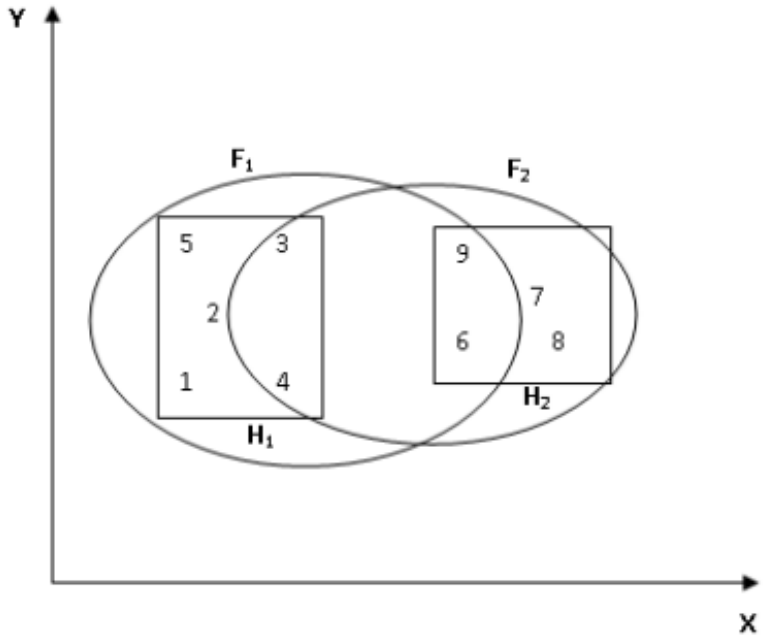
Figura 4 – Exemplo da estrutura de agrupamento dos métodos particionais (b) e hierárquicos (c)



Fonte: (MACIEL, 2013).

Nesta dissertação, trabalha-se exclusivamente com algoritmos de particionamento rígido adaptados para Análise de Formas, sendo eles: K-médias (LLOYD, 1982), K-médias aparado (GARCÍA-ESCUDERO; GORDALIZA, 1999), *CLARANS* (NG; HAN, 2002) e *Hill Climbing* (FRIEDMAN; RUBIN, 1967).

Figura 5 – Exemplo de agrupamento rígido e difuso. Os retângulos representam o agrupamento rígido e as elipses representam o agrupamento difuso



Fonte: (JAIN; MURTY; FLYNN, 1999).

3.2 ALGORITMOS DE AGRUPAMENTO ADAPTADOS PARA DADOS DE FORMAS TRIDIMENSIONAIS

Usualmente, os algoritmos aqui propostos geram os grupos de dados cujas observações são representadas por valores pertencentes ao espaço euclidiano e o quadrado da distância euclidiana é comumente utilizado. Porém, no contexto de análise de formas, as observações serão representadas por matrizes de configuração e as distâncias que podem ser utilizadas são as que foram apresentadas nas Equações (2.12), (2.13) e (2.14). Trabalhos envolvendo a adaptação de algoritmos de agrupamento no contexto de análise de formas foram desenvolvidos por alguns autores. Para dados de formas planares ($m = 2$), Georgescu (2009) faz a adaptação do algoritmo K-médias para o agrupamento de formas *fuzzy*, Amaral et al. (2010) realiza a adaptação da versão de Hartigan-Wong do K-médias, Oliveira (2016) realiza a adaptação da versão de Macquenn do K-médias e Assis (2018) faz a adaptação do algoritmo KI-médias e de algoritmos de busca, como o *Hill Climbing*, aplicando o *Bagging* em ambas metodologias. Para dados de formas tridimensionais ($m = 3$), Vinué, Simó e Alemany (2014) realizaram as adaptações das versões de Lloyd e Hartigan-Wong do K-médias e do K-médias aparado.

A princípio, para as aplicações do presente trabalho, considere $\mathbf{X}^* = \{\mathbf{X_i}, i = 1, \cdots, n\}$

um conjunto de n objetos, em que cada objeto é uma matriz de configuração, definida em (2.2). Tais objetos serão agrupados em um conjunto $G = (G_r, r = 1, \cdots, K)$ de K grupos formados de acordo com o algoritmo utilizado. Os algoritmos adaptados para o agrupamento de formas tridimensionais serão apresentados a seguir.

3.3 MÉTODOS BASEADOS EM CENTROIDES

Nesta seção serão definidos os conceitos do algoritmo baseado em centroides, o *K*-médias e sua versão mais robusta conhecida como *K*-médias aparado, adaptados ao contexto de análise de formas tridimensionais.

3.3.1 Algoritmo K-médias para formas tridimensionais

Sendo um dos algoritmos de particionamento mais conhecido, o K-médias foi proposto com a finalidade de agrupar as observações de um conjunto de dados fazendo o uso do fato de que o valor que minimiza à distância de cada observação ao centroide do grupo ao qual essas observações pertencem é a média amostral. Esse algoritmo originalmente é utilizado em situações que se trabalha com variáveis quantitativas e o quadrado da distância euclidiana é utilizado. Tal algoritmo foi proposto de diferentes modos e suposições, e algumas referências clássicas relacionadas são: o trabalho de Lloyd (1982), originalmente proposto em 1957, que atualiza os centroides (médias dos grupos) somente depois de atribuir todos os elementos a um grupo; o trabalho de MacQueen (1967), no qual o termo K-médias aparece pela primeira vez e o trabalho de Hartigan e Wong (1979), o qual apresenta uma versão mais eficiente do algoritmo K-médias cujas médias são atualizadas após cada nova atribuição. Como foi mencionado, nos algoritmos de particionamento rígido, cada indivíduo pertence a exatamente um dos grupos e o mesmo indivíduo não pode pertencer a mais de um grupo. O algoritmo K-médias trata de um problema de otimização que tenta encontrar uma partição ideal baseando-se na minimização da soma dos quadrados (BOCK, 2007); (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No contexto de agrupamento de formas tridimensionais, Vinué, Simó e Alemany (2014) adaptaram as versões de Lloyd e de Hartigan-Wong do K-médias, constatando que a versão de Lloyd possui um custo benefício computacional melhor em relação a versão de Hartigan-Wong, quando os objetos do conjunto de dados são compostos por um número grande de marcos anatômicos e o tamanho da amostra seja média ou grande. Desse modo, o algoritmo que

será apresentado e utilizado neste trabalho é a versão de Lloyd do K-médias e do K-médias aparado, adaptados por Vinué, Simó e Alemany (2014). E ainda, será utilizada a abordagem do K-médias, adaptada por Teotonio e Amaral (2020) utilizando a média ϕ. Também será apresentada, como uma das contribuições do presente trabalho, uma nova abordagem do K-médias aparado utilizando a média ϕ.

O problema de otimização do K-médias é a minimização do seguinte critério:

$$W = \sum_{r=1}^{K} \sum_{i \in G_r} Dist^2(\mathbf{X_i}, \mu_\mathbf{r}), \tag{3.1}$$

em que $G_r, r = 1, \cdots, K$, representa os grupos formados que contêm cada uma das $\mathbf{X_i}, i = 1, \cdots, n$, observações do conjunto de dados. A função $Dist$ é uma medida de distância, podendo ser umas das Equações (2.12), (2.13) ou (2.14) e $\mu_\mathbf{r}$ representa a forma média, que pode ser obtida pela Definição 7 ou pela Equação (2.19).

É importante salientar que na adaptação de Vinué, Simó e Alemany (2014), utiliza-se a distância Riemanniana e a forma média total de Procrustes apresentada na Definição 7. Nesta dissertação, iremos trabalhar com a versão proposta por Vinué, Simó e Alemany (2014) e também com a abordagem que utiliza a distância Riemanniana e a forma média ϕ (2.19), assim como foi implementado por Teotonio e Amaral (2020), com o intuito de verificar a qualidade dos grupos e o tempo de execução do algoritmo com o seu uso. Para a versão utilizando a média ϕ deu-se o nome de K-médias ϕ. O processo iterativo do algoritmo K-médias encontra-se definido no Algoritmo 2.

Algoritmo 2: Algoritmo K-médias para formas tridimensionais.

Entrada: Um conjunto $\mathbf{X}^*$ de matrizes de configuração e o número de grupos K.
Saída: Grupos formados $(G = G_r, r = 1, \cdots, K)$.

1. Selecionar arbitrariamente K matrizes de configuração (observações) como os centroides iniciais (partição inicial) dos K grupos.
2. Para cada observação, calcular a distância Riemanniana (2.14) entre as observações e os K-centroides. Em seguida, atribuir cada observação ao grupo com o centroide mais próximo.
3. Para cada um dos K grupos, calcular a forma média pela Definição 7 ou pela Equação (2.19). As formas médias de cada grupo serão consideradas os novos centroides.
4. Repetir as etapas 2 e 3 até as observações não trocarem mais de grupos, isto é, até que o valor do critério (3.1) fique inalterado ou até que um valor mínimo seja atingido.

Os grupos formados pelo Algoritmo 2 dependem de uma partição inicial que representa os centroides, podendo convergir para um ótimo local, dependendo de quais observações foram

definidas como os centroides dos K grupos. Dessa maneira, é necessário executar o algoritmo várias vezes com diferentes partições iniciais e selecionar a execução cuja partição inicial apresentou o melhor valor do critério de agrupamento. A principal limitação do K-médias é a influência de valores discrepantes das observações, uma vez que é a média amostral que minimiza à distância de cada observação ao centroide do seu grupo. Com a finalidade de solucionar essa limitação, foi proposto uma versão considerada mais robusta, o algoritmo K-médias aparado.

3.3.2 Algoritmo K-médias aparado para formas tridimensionais

O algoritmo K-médias pode ser influenciado por observações discrepantes (*outliers*) nos conjuntos de dados. Com o intuito de lidar com esse problema, García-Escudero e Gordaliza (1999) propuseram uma versão mais robusta desse método. A estratégia adotada pelos autores é a remoção das observações consideradas atípicas (tais observações são as que apresentarem maiores valores para as distâncias). Tem-se que uma proporção α (entre 0 e 1), escolhida pelo o usuário, das observações é aparada (corte) e a proporção de observações que apresentaram os maiores valores para as distâncias são cortadas. Basicamente, o procedimento se resume na aplicação do método K-médias em um subconjunto dos dados sem os *outliers*.

No contexto de formas tridimensionais, Vinué, Simó e Alemany (2014), adaptaram essa versão utilizando a distância Riemanniana e a forma média total de Procrustes apresentada na Definição 7. Nesta dissertação, iremos trabalhar com a versão proposta por Vinué, Simó e Alemany (2014) e também com uma nova abordagem que utiliza a distância Riemanniana e a forma média ϕ (2.19), como uma das propostas do presente trabalho, com o intuito de verificar a qualidade dos grupos e o tempo de execução do algoritmo com o seu uso. Para a versão utilizando a média ϕ deu-se o nome de K-médias ϕ aparado. O processo iterativo do algoritmo K-médias aparado encontra-se definido no Algoritmo 3.

Apesar de ser considerado mais robusto, o método K-médias aparado pode apresentar algumas desvantagens. A principal delas é a perda de informação. Dependendo do tipo de fenômeno estudado, a remoção de observações pode levar a uma compreensão errônea sobre os fatos, principalmente em pequenos conjuntos de dados. Com a finalidade de solucionar as limitações dos métodos K-médias e K-médias aparado, implementou-se, no presente trabalho, as versões adaptadas dos algoritmos *CLARANS* e *Hill Climbing* ao espaço de formas.

Algoritmo 3: Algoritmo K-médias aparado para formas tridimensionais.

Entrada: Um conjunto $\mathbf{X}^*$ de matrizes de configuração, número de grupos K, proporção do corte α.

Saída: Grupos formados $(G = G_r, r = 1, \cdots, K)$.

1. Selecionar arbitrariamente K matrizes de configuração (observações) como os centroides iniciais (partição inicial) dos K grupos.
2. Para cada observação, calcular a distância Riemanniana (2.14) entre as observações e os K-centroides.
3. As $n\alpha$ observações com os maiores valores para a distância são removidas. Em seguida, as $n(1-\alpha)$ observações restantes são atribuídos ao grupo com o centroide mais próximo.
4. Para cada um dos K grupos, calcular a forma média pela Definição 7 ou pela Equação (2.19). As formas médias de cada grupo serão consideradas os novos centroides.
5. Repetir as etapas 2 e 3 até as observações não trocarem mais de grupos, isto é, até que o valor do critério (3.1) fique inalterado ou até que um valor mínimo seja atingido.

3.4 MÉTODOS BASEADOS EM MEDOIDES

Nesta seção serão definidas as novas abordagens dos algoritmos baseados em medoides, *PAM* e *CLARANS*, adaptados ao contexto de análise de formas tridimensionais.

3.4.1 Algoritmo *PAM* para formas tridimensionais

O algoritmo *PAM* (*Partitioning Around Medoids*) foi proposto por Rousseeuw e Kaufman (1990) e sua abordagem para obtenção de grupos faz uso de um objeto representativo, conhecido como medoide, que deve ser localizado mais ao centro do grupo. Diferentemente do K-médias, que utiliza um centroide, que é a média de todos os pontos pertencentes aos grupos. Basicamente o algoritmo funciona da seguinte maneira, seleciona-se k observações aleatoriamente para serem os medoides de cada grupo e os objetos que não forem selecionados como medoides (não-medoides) são alocados ao grupo do medoide aos quais cada um desses objetos é mais semelhante. Essa alocação é feita de acordo com uma medida de dissimilaridade (distância). Em seguida, utilizando um processo iterativo, troca-se um dos medoides selecionados por um não-medoide e verifica se há uma melhora na qualidade do agrupamento. Essa qualidade é calculada utilizando a dissimilaridade entre um objeto e o medoide do seu grupo (ROUSSEEUW; KAUFMAN, 1990).

Existem quatro casos a serem examinados para determinar se a troca de um medoide

selecionado por um não selecionado produz um bom agrupamento e Ng e Han (2002) resumem esse processo da seguinte maneira. Considere O_s um medoide selecionado, O_h um medoide não selecionado como um substituto de O_s, O_j como os objetos restantes não selecionados (não-medoides) que podem ou não precisam ser movidos, O_i como um medoide que está mais próximo de O_j e d uma medida de dissimilaridade. Os quatro casos são os seguintes:

- Caso 1: Suponha que O_j está alocado ao grupo representado por O_s. Ainda, considere que O_j seja mais próximo de O_i do que de O_h, ou seja, $d(O_j, O_h) \geq d(O_j, O_i)$. Desse modo, se O_s for substituído por O_h, O_j pertencerá ao grupo representado por O_i. Logo, o custo de troca em relação a O_j é:

$$C_{jsh} = d(O_j, O_i) - d(O_j, O_s). \tag{3.2}$$

  O custo calculado por esta equação sempre será positivo, apontando que o custo da troca de O_s por O_h nessa situação é positivo.

- Caso 2: Atualmente O_j está alocado ao grupo representado por O_s. Porém, agora considere O_j mais próximo de O_h do que de O_i, ou seja, $d(O_j, O_h) < d(O_j, O_i)$. Desse modo, se O_s for trocado por O_h, então O_j pertencerá ao grupo representado por O_h. Logo o custo relacionado a O_j é:

$$C_{jsh} = d(O_j, O_h) - d(O_j, O_s). \tag{3.3}$$

  O custo calculado por (3.3) depende de O_j ser mais próximo de O_s do que de O_h, levando ao custo ser positivo ou negativo.

- Caso 3: Agora suponha que O_j atualmente esteja alocado a um grupo que seja representado por outro medoide selecionado, exceto O_s e seja O_i o objeto mais representativo desse grupo. Considere O_j mais próximo de O_i do que de O_h, então mesmo que O_s seja trocado por O_h, O_j continuará alocado no grupo representado por O_i. Logo, o custo é:

$$C_{jsh} = 0. \tag{3.4}$$

- Caso 4: Considere que O_j atualmente esteja alocado ao grupo representado por O_i, mas O_j é menos próximo de O_i do que de O_h. Desse modo, trocar O_s por O_h faria com que O_j fosse alocado para o grupo representado por O_h. Logo, o custo é:

$$C_{jsh} = d(O_j, O_h) - d(O_j, O_i). \tag{3.5}$$

  O custo calculado por (3.5) sempre será negativo.

Combinando os quatro casos que foram apresentados, o custo total da substituição de O_s por O_h é dado por:

$$TC_{sh} = \sum_j C_{jsh}. \tag{3.6}$$

Desse modo, a Equação (3.6) é a soma de valores positivos e negativos. O Algoritmo 4 apresenta o passo a passo do *PAM* e como os valores de (3.6) são utilizados.

Algoritmo 4: Algoritmo *PAM* para formas tridimensionais.

Entrada: Um conjunto $\mathbf{X}^*$ de matrizes de configuração e o número de grupos K.
Saída: Grupos formados $(G = G_r, r = 1, \cdots, K)$.

1. Selecionar arbitrariamente K matrizes de configuração (observações) como os medoides iniciais (partição inicial) dos K grupos.
2. Em seguida, atribuir cada observação ao grupo com o medoide mais semelhante usando uma das distâncias apresentadas nas Equações (2.12), (2.13) e (2.14).
3. Calcular TC_{sh} para todos os pares de objetos O_s, O_h, em que O_s está atualmente selecionado e O_h não.
4. Selecionar o par O_s, O_h que corresponde ao mínimo de TC_{sh}. Se o TC_{sh} mínimo for negativo, substituir O_s por O_h; vá para o passo 2.
5. Repetir os passos 2-4 até que não haja mais trocas dos medoides.

Com o intuito de superar uma das limitações do *PAM*, ou seja, a de não ser apropriado para trabalhar com grandes conjuntos de dados, Rousseeuw e Kaufman (1990) propuseram o algoritmo *CLARA* (*Clustering Large Applications*). O *CLARA* é uma extensão do *PAM* baseada em amostragem e funciona da seguinte maneira: o algoritmo seleciona 5 amostras do conjunto de dados de tamanho $40+2K$ e executa o *PAM* em cada uma dessas amostras. Dessa maneira, seleciona-se os medoides que apresentam o melhor agrupamento e se as amostras forem bem representativas, então os medoides que foram selecionados são equivalentes aos medoides que seriam selecionados com a aplicação do *PAM* em todo o conjunto de dados (ROUSSEEUW; KAUFMAN, 1990). Basicamente, enquanto o *PAM* encontra os melhores medoides com base em todo o conjunto de dados, o *CLARA* quer encontrar os melhores medoides entre as amostras.

3.4.2 Algoritmo *CLARANS* para formas tridimensionais

Agora que os métodos *PAM* e *CLARA* foram citados, pode-se definir o processo de agrupamento do *CLARANS*. O algoritmo *CLARANS* (*Clustering Algorithm based on Randomized Search*) foi proposto por Ng e Han (2002), cuja formação dos grupos é baseada em uma pesquisa aleatória. Os grupos formados pelo CLARANS são baseados por buscas em todo o conjunto de dados, ou seja, buscas em um grafo. Considere $G_{n,K}$ um grafo, define-se como

um nó um conjunto de medoides obtidos desse grafo e cada nó resulta em grupos diferentes. Cada nó possui $K(n-K)$ vizinhos, em que K representa o número de grupos e n o número de indivíduos do conjunto de dados. Dois nós são considerados vizinhos se seus conjuntos diferem em apenas um objeto. Exemplificando, seja $S_1 = (A, B)$ e $S_2 = (C, D)$ dois nós, em que A, B, C e D são medoides, assim $V_1 = (A, C)$, $V_2 = (A, D)$, $V_3 = (B, C)$ e $V_4 = (B, D)$ diferem de S_1 e S_2 por apenas um objeto, logo S_1 e S_2 são considerados vizinhos de $V_i,\ i = 1, 2, 3, 4$. Um custo é definido para cada nó com o objetivo de ser a dissimilaridade total entre cada objeto e o medoide de seu grupo. O diferencial de custo entre dois vizinhos é calculado pela Equação (3.6) (NG; HAN, 2002). De modo geral, o *CLARANS* busca pelo nó ótimo em todo o grafo aplicando o *PAM* em uma amostra dos vizinhos.

O *CLARANS* possui dois parâmetros principais: *maxneighbor*, que representa o número máximo de vizinhos a serem examinados e *numlocal*, que representa o número de iterações de busca por um nó mínimo. Se o valor de *maxneighbor* for muito próximo ou igual a $K(n-K)$, mais a qualidade dos grupos formados pelo *CLARANS* se iguala aos grupos gerados pelo *PAM* e mais longa é a busca por um nó mínimo (NG; HAN, 2002). O Algoritmo 5 apresenta os passos da adaptação do *CLARANS*, uma das propostas do presente trabalho, ao contexto de formas 3D.

Algoritmo 5: Algoritmo *CLARANS* para formas tridimensionais.

Entrada: Um conjunto $\mathbf{X}^*$ de matrizes de configuração, número de grupos K, *numlocal* e *maxneighbor*.
Saída: Grupos formados $(G = G_r, r = 1, \cdots, K)$.
1. Inicializar $i = 1$ (primeira iteração de busca) e *mincost* como um número grande.
2. Definir *current* como um nó (conjunto de medoides) selecionado aleatoriamente de $G_{n,K}$.
3. Inicializar $j = 1$ (primeiro vizinho analisado).
4. Seja S um vizinho aleatório de *current*. Calcular o custo diferencial entre S e *current* de acordo com a Equação (3.6). Cada C_{jsh} será calculado a partir de uma das Equações (2.12), (2.13) ou (2.14).
5. Se S possuir um custo menor, então S é definido como *current*; vá para o passo 3.
6. Caso contrário, incrementar j em 1. Se $j \leq maxneighbor$, vá para o passo 4.
7. Se $j > maxneighbor$, compare o custo de *current* com *mincost*. Se o custo de *current* for menor que *mincost*, então *mincost* é definido como custo de *current* e *current* é definido como o melhor nó.
8. Incrementar i em 1. Se $i > numlocal$, o agrupamento formado pelo melhor nó é retornado como resultado. Senão, retorne ao passo 2.

Nos experimentos realizados por Ng e Han (2002), definiu-se $maxneighbor = 0,0125 * K(n-K)$ e $numlocal = 2$ como valores razoáveis para esses parâmetros. Os autores co-

mentam que a qualidade do agrupamento produzido é efetivamente a mesma do agrupamento gerado pelo *PAM*, mas são obtidos com um custo computacional menor e possui uma qualidade melhor em relação ao *CLARA*. O algoritmo *CLARANS* é baseado em pesquisa aleatória, sendo voltado para o agrupamento de grandes conjuntos de dados.

3.5 MÉTODO BASEADO EM BUSCA

Nesta seção serão definidos os conceitos do algoritmo baseado em busca *Hill Climbing*, adaptado ao contexto de análise de formas tridimensionais.

3.5.1 Algoritmo *Hill Climbing* para formas tridimensionais

O algoritmo *Hill Climbing* foi proposto por Friedman e Rubin (1967) com o objetivo de encontrar, por meio de um processo iterativo, o agrupamento formado pela partição que possua o valor ótimo de um determinado critério de agrupamento ou função objetivo. O método *Hill Climbing* é considerado um algoritmo de busca. Um algoritmo de busca tem como finalidade encontrar a solução de um determinado problema por meio de uma série de caminhos possíveis. Resumidamente, tais algoritmos buscam pela partição ótima descartando as partições analisadas recentemente cujo valor do critério não seja o ótimo (ASSIS, 2018); (RUSSELL; NORVIG, 2014). O algoritmo *Hill Climbing* é um laço que se repete de modo contínuo em busca de um valor ótimo.

O critério de agrupamento adotado foi proposto por Rousseeuw (1987). Rousseeuw e Kaufman (1990) comentam que a seguinte equação pode ser utilizada como critério para medir a qualidade dos grupos formados pelo *PAM*:

$$TD = \sum_{r=1}^{K} \sum_{i \in G_r} d(\mathbf{X_i}, m_r) \tag{3.7}$$

em que $G_r, r = 1, \cdots, K$, representa os grupos formados contendo cada uma das $\mathbf{X_i}, i = 1, \cdots, n$, observações do conjunto de dados. A função d é uma medida de distância, podendo ser umas das Equações (2.12), (2.13) ou (2.14) e m_r representa a partição de cada grupo. Quanto menor o valor de (3.7), melhor o agrupamento. No contexto de formas, o algoritmo *Hill Climbing* foi adaptado por Assis (2018) para o agrupamento de formas 2D. Como um dos objetivos da presente dissertação, adaptou-se esse algoritmo no contexto de formas 3D. Os passos do *Hill Climbing* para formas tridimensionais encontra-se apresentado no Algoritmo 6.

Algoritmo 6: Algoritmo *Hill Climbing* para formas tridimensionais.

Entrada: Um conjunto $\mathbf{X}^*$ de matrizes de configuração e o número de grupos K.
Saída: Grupos formados $(G = G_r, r = 1, \cdots, K)$.
1. Selecionar arbitrariamente K matrizes de configuração (observações) para ser a partição inicial dos K grupos e calcule o valor do critério.
2. Selecione um vizinho do nó atual e calcule o valor do critério.
3. Se o valor do critério (3.7) melhorar, realize a alteração. Caso contrário, examine outro nó vizinho .
4. Repetir os passos 2 e 3 até não haver melhorias no valor do critério de agrupamento.

Basicamente o algoritmo funciona da seguinte maneira: o nó selecionado (partição inicial) é substituído por um nó vizinho até que o valor ótimo do critério seja encontrado ou até o algoritmo atingir um número de iterações definido previamente. O conceito de nó é o mesmo definido na descrição do método *CLARANS*. Esses algoritmos examinam as noções de proximidades entre as partições examinadas durante o processo iterativo.

3.6 VALIDAÇÃO DE AGRUPAMENTO

Para avaliar os resultados obtidos pelos algoritmos propostos, fez-se uso das técnicas de validação de agrupamento, com o intuito de verificar se os grupos obtidos são realmente bons ou se foram gerados ao acaso. Tais técnicas consistem em procedimentos que analisam de maneira objetiva e quantitativa, por meio de índices, os grupos formados. Halkidi, Batistakis e Vazirgiannis (2002) comentam que as técnicas de validação de agrupamento são divididas em três metodologias: validação externa, validação interna e validação relativa.

Os índices de validação externa consistem na avaliação do grau de correspondência entre os grupos gerados pelos algoritmos e uma estrutura de agrupamento conhecida ou tida como verdadeira. Por outro lado, nos índices de validação interna os grupos obtidos pelos algoritmos são avaliados com base no próprio conjunto de dados, ou seja, não faz uso de informação adicional. Já na validação relativa, tem-se como objetivo encontrar a melhor solução comparando diversos agrupamentos (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002); (JAIN; DUBES, 1988). Entre as várias abordagens, pode-se citar os índices de Rand (RAND, 1971), Jaccard (JACCARD, 1908) e Rand Ajustado (HUBERT; ARABIE, 1985) de validação externa e os índices de Dunn (DUNN, 1974), Davies-Boudin (DAVIES; BOULDIN, 1979) e Silhueta (ROUSSEEUW, 1987) de validação interna e relativa. No presente trabalho, fez-se uso dos índices de Rand Ajustado e de Silhueta para avaliar os grupos formados.

3.6.1 Índice de Rand Ajustado

O Índice de Rand (IR), proposto por Rand (1971), é bastante conhecido e utilizado em diversos estudos. Considere R o agrupamento do conjunto de dados conhecido a priori, Q o agrupamento formado por um dos algoritmos e Ω o conjunto de dados e considere as seguintes variáveis:

- a: representa o número de pares de objetos de Ω que pertencem ao mesmo grupo em R e em Q.
- b: representa o número de pares de objetos de Ω que pertencem ao mesmo grupo em R e a diferentes grupos em Q.
- c: representa o número de pares de objetos de Ω que pertencem a diferentes grupos em R e ao mesmo grupo em Q.
- d: representa o número de pares de objetos de Ω que pertencem a grupos diferentes em R e em Q.

O IR é definido da seguinte maneira:

$$IR(R,Q) = \frac{a+d}{a+b+c+d}. \tag{3.8}$$

Os valores do IR variam de $[0,1]$ e valores próximos de 1 indicam que o agrupamento gerado por um algoritmo e o agrupamento conhecido a priori são ditos semelhantes. O principal problema do IR é que os termos a e d possuem a mesma importância, isto é, na avaliação do grau de correspondência não é feita uma divisão entre os pares ligados e pares separados (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002); (JAIN; DUBES, 1988).

Com o objetivo de avaliar o desempenho dos grupos formados pelos algoritmos apresentados nesta dissertação, utilizou-se o Índice de Rand Ajustado (IRA). O IRA é um método de validação externa proposto por Hubert e Arabie (1985), isto é, mede a similaridade entre um agrupamento formado por um algoritmo e um agrupamento tido como verdadeiro. Considere R o agrupamento do conjunto de dados conhecido a priori, Q o agrupamento formado por um dos algoritmos e Ω o conjunto de dados. O índice de Rand Ajustado é dado por:

$$IRA(R,Q) = \frac{a - \dfrac{(a+c)(a+b)}{a+b+c+d}}{\dfrac{(a+c)+(a+b)}{2} - \dfrac{(a+c)(a+b)}{a+b+c+d}}. \tag{3.9}$$

Os valores do IRA variam de $[-1, 1]$ e as interpretações que podem-se obter são as seguintes:

- o agrupamento gerado por um algoritmo e o agrupamento conhecido a priori são ditos semelhantes se o valor do IRA for próximo de 1.
- o agrupamento gerado por um algoritmo e o agrupamento conhecido a priori são ditos diferentes se o valor do IRA for próximo de -1.

3.6.2 Índice de Silhueta

O Índice de Silhueta foi proposto por Rousseeuw (1987) como uma nova exibição gráfica para técnicas de particionamento em que cada grupo é representado por uma silhueta. Sendo um método de validação interna, é bastante utilizado para estimar o número de grupos de um determinado conjunto de dados. Ainda, Flach (2012) comenta que também podemos utilizá-lo para medir a qualidade de grupos formados. Considere Ω a representação de um conjunto de dados e $G_r, r = 1, \cdots, K$ os K grupos formados pelos algoritmos. O Índice de Silhueta atribui uma medida de qualidade, conhecida como largura da silhueta, para cada elemento de Ω e o seu valor é utilizado para indicar o grau de associação do *i-ésimo* elemento no grupo G_r. A largura da silhueta é definida como:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}. \tag{3.10}$$

Em que $a(i)$ representa a distância média do *i-ésimo* elemento em relação a todos os outros elementos pertencentes ao seu grupo e $b(i)$ representa a distância mínima entre o *i-ésimo* elemento e todos os outros elementos não pertencentes ao seu grupo. No contexto de análise de formas, as distâncias que podem ser utilizadas são as apresentadas nas Equações (2.12), (2.13) e (2.14).

Dessa maneira, o Índice de Silhueta é definido por:

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i). \tag{3.11}$$

Isto é, a média de todos as silhuetas dos elementos de Ω. E quanto mais próximo de 1, melhor o agrupamento formado.

4 EXPERIMENTOS NUMÉRICOS E RESULTADOS

Com o objetivo de avaliar o desempenho dos algoritmos propostos para o agrupamento de formas tridimensionais, foram realizados experimentos em conjuntos de dados simulados e aplicações em conjuntos de dados reais. Nos experimentos e nas aplicações utilizou-se a distância Riemanniana (2.14) como medida de dissimilaridade para substituir as distâncias comumente usadas nos algoritmos originais. Em todos os experimentos, a escolha da partição inicial foi feita de modo aleatório. Com a finalidade de apresentar a eficiência dos algoritmos adaptados, avaliou-se o desempenho destes em diferentes conjuntos de dados. Utilizou-se o Índice de Rand Ajustado (IRA) (3.9) e o Índice de Silhueta (3.11) para verificar a qualidade dos grupos formados. Quanto mais próximo os valores desses índices for de 1, melhores os desempenhos medidos para os métodos. Os algoritmos propostos foram implementados na linguagem de programação R (R Core Team, 2020) com o auxílio das funções pertencentes aos pacotes voltados para análise de formas, **shapes** (DRYDEN, 2012) e **Anthropometry** (VINUÉ et al., 2020). Para ilustrar o desempenho dos algoritmos, simulou-se dois experimentos no espaço de marcos baseados nos dados simulados por Vinué, Simó e Alemany (2014) e Preston e Wood (2010), respectivamente. Ambos os trabalhos simulam formas tridimensionais.

4.1 CONJUNTOS DE DADOS SIMULADOS

4.1.1 Dados simulados de cubos e paralelepípedos aleatórios

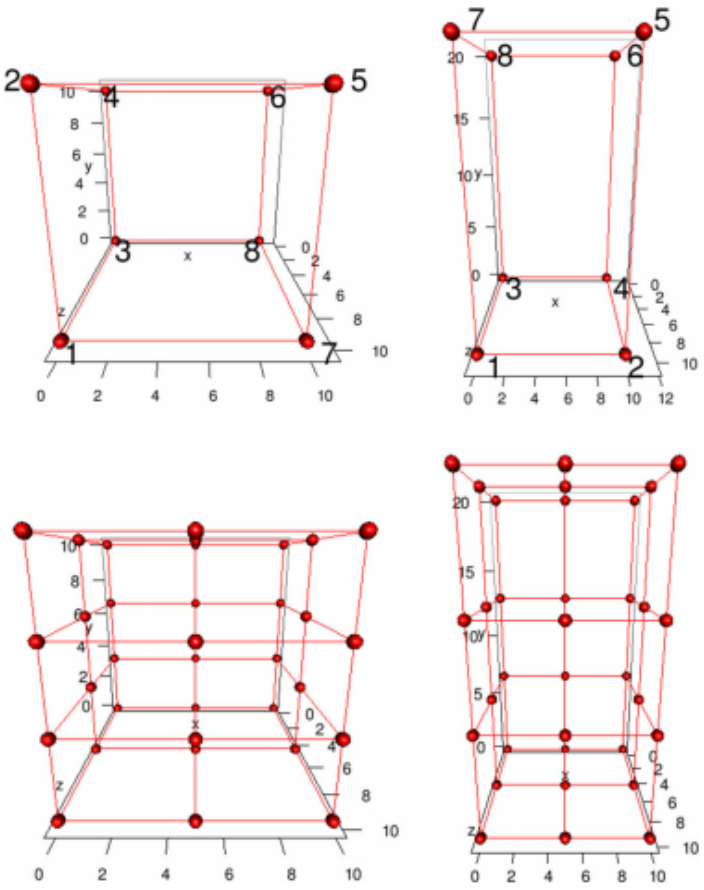
No primeiro experimento, baseado nos dados simulados por Vinué, Simó e Alemany (2014), foram consideradas duas figuras geométricas predefinidas, um cubo e um paralelepípedo para representar as configurações médias, com números diferentes de marcos anatômicos (k $=$ 8 e k $=$ 34). Em seguida, simulou-se n_1 objetos correspondentes a um grupo e n_2 objetos correspondentes a outro grupo. O grupo 1 é definido por uma distribuição normal multivariada com vetor médio tridimensional de dimensão $3k$ representado pelo cubo gerado e uma matriz de covariância $\mathbf{\Sigma_1}$ de dimensão $3k\ x\ 3k$. Da mesma maneira, o grupo 2 é definido por uma distribuição normal multivariada com vetor médio tridimensional de dimensão $3k$ representado pelo paralelepípedo gerado e uma matriz de covariância $\mathbf{\Sigma_2}$ de dimensão $3k\ x\ 3k$. A orientação dos cubos e paralelepípedos foi definida de modo arbitrário. Uma rotação é aplicada sobre o eixo z de acordo com um ângulo aleatório gerado pela função *rvm* do pacote **CircStats** do R (LUND;

AGOSTINELLI, 2012), a função *rvm* gera números pseudoaleatórios de uma distribuição de von Mises. A Figura 6 mostra os cubos e paralelepípedos gerados para os diferentes tamanhos de marcos anatômicos e a Figura 7 apresenta o paralelepípedo com $k = 8$ marcos anatômicos rotacionado de acordo com um ângulo aleatório. Foram considerados tamanhos de amostras iguais e dois cenários foram estudados, isotrópico e anisotrópico. Selecionou-se os mesmos valores para n_1 e n_2 para o tamanho de amostra de $N = 100$ $(n_1 = n_2 = 50)$. Cada conjunto de dados simulados possui N objetos divididos em dois grupos. Os marcos anatômicos gerados seguem distribuição normal multivariada e são transformados em matrizes de configuração para formação dos objetos de cada grupo. Para o cenário de isotropia, os marcos anatômicos possuem aproximadamente a mesma variabilidade, a matriz de covariância é um múltiplo da matriz identidade e considerou-se $\mathbf{\Sigma_i} = \sigma_i^2\mathbf{I_{3k\ x\ 3k}}$ com diferentes valores para $\sigma_i,\ i = 1, 2$. Os valores para σ_1 e σ_2 foram escolhidos de modo que os dados tenham dispersão variada (0,1, 1,5 e 6) e sejam iguais em cada caso. Já para o cenário de anisotropia, os marcos anatômicos não possuem aproximadamente a mesma variabilidade, considerou-se matrizes de covariâncias geradas de modo arbitrário utilizando a função *genPositiveDefMat* do pacote **clusterGeneration** do R (QIU; JOE, 2013). Essa função gera valores próprios para a matriz de covariância aleatoriamente. O nível de variabilidade pode ser controlada usando os argumentos *lambdaLow* e *ratioLambda*. Utilizou-se para o argumento *covMethod* o método "*eigen*" e foram consideradas duas variações para os valores de (*lambdaLow, ratioLambda*). Na Variação 1 todos os valores da matriz de covariância estão no intervalo $(1, 10)$ e na Variação 2 todos os valores estão no intervalo $(10, 30)$.

4.1.2 Dados simulados de representações geométricas aleatórias

No segundo experimento, baseado nos dados simulados por Preston e Wood (2010), foram consideradas duas configurações médias aleatórias formadas por valores obtidos de uma distribuição uniforme, com números diferentes de marcos anatômicos (k = 6 e k = 36). Em seguida, simulou-se n_1 objetos correspondentes a um grupo e n_2 objetos correspondentes a outro grupo. O grupo 1 é definido por uma distribuição normal multivariada com vetor médio tridimensional de dimensão $3k$ representado por valores de uma distribuição $U(-1, 1)$ e uma matriz de covariância $\mathbf{\Sigma_1}$ de dimensão $3k\ x\ 3k$. Da mesma maneira, o grupo 2 é definido por uma distribuição normal multivariada com vetor médio tridimensional de dimensão $3k$ representado por valores de uma distribuição $U(-5, 5)$ e uma matriz de covariância $\mathbf{\Sigma_2}$ de dimensão $3k\ x\ 3k$. Os inter-

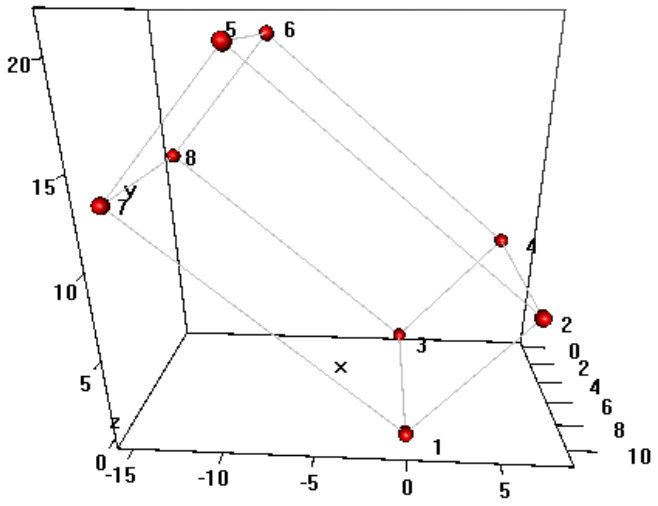
Figura 6 – Cubos e paralelepípedos formados por 8 marcos anatômicos (primeira linha) e 34 marcos anatômicos (segunda linha)



Fonte: (VINUÉ; SIMÓ; ALEMANY, 2014).

valos para a distribuição uniforme foram definidos de modo que se tenha duas configurações médias distintas para a formação dos grupos. Foram considerados tamanhos de amostras iguais e dois cenários foram estudados, isotrópico e anisotrópico. Selecionou-se os mesmos valores para n_1 e n_2 para o tamanho de amostra de $N = 100$ $(n_1 = n_2 = 50)$. Os marcos anatômicos gerados seguem distribuição normal multivariada e são transformados em matrizes de configuração para formação dos objetos de cada grupo. Para o cenário de isotropia, os marcos anatômicos possuem aproximadamente a mesma variabilidade, a matriz de covariância é um múltiplo da matriz identidade e considerou-se $\mathbf{\Sigma_i} = \sigma_i^2 \mathbf{I_{3k\ x\ 3k}}$ com diferentes valores para $\sigma_i,\ i = 1, 2$. Os valores para σ_1 e σ_2 foram escolhidos de modo que os dados tenham dispersão variada (0,1, 1,5 e 6) e sejam iguais em cada caso. Já para o cenário de anisotropia, os marcos anatômicos não possuem aproximadamente a mesma variabilidade, a matriz de covariância é representada pelo resultado do produto de Kronecker entre duas operações envolvendo a matriz identidade, isto é, $\mathbf{\Sigma_i} = \sigma_i^2 \left(\mathbf{1_k 1_k}^T + (\gamma - 1)\mathbf{I_{k\ x\ k}}\right) \otimes \left(\mathbf{1_m 1_m}^T + (\gamma - 1)\mathbf{I_{m\ x\ m}}\right) / \gamma^2$,

Figura 7 – Paralelepípedo formado por 8 marcos anatômicos rotacionado por um ângulo aleatório



Fonte: Elaborada pelo o autor (2021).

$i = 1, 2$. Os valores para σ_1 e σ_2 seguem os mesmos para o cenário de isotropia e o valor de γ é igual a 2.

4.1.3 Avaliação numérica

Sobre as aplicações dos algoritmos nos dados simulados, as Tabelas 2, 3, 4 e 5 apresentam os resultados obtidos referentes aos experimentos baseados nos dados simulados por Vinué, Simó e Alemany (2014) e as Tabelas 6, 7, 8 e 9 apresentam os resultados obtidos referentes aos experimentos baseados nos dados de formas 3D baseados nas simulações de Preston e Wood (2010). No cotidiano, as representações das formas podem ser simples ou complexas, isto é, bem definidas ou não. Considerou-se esses dois experimentos com o intuito de verificar a eficácia dos algoritmos em agrupar diferentes tipos de formas, isto é, no primeiro experimento tem-se formas bem estruturas (cubos e paralelepípedos) e no segundo tem-se formas definidas ao acaso por configurações médias definidas por pontos de uma distribuição uniforme. Para cada aplicação dos algoritmos, todos os resultados apresentados foram obtidos para a melhor inicialização aleatória, considerando diferentes graus de dispersão/variação e os diferentes números de marcos anatômicos.

Quanto aos valores dos parâmetros utilizados pelos algoritmos adaptados, optou-se por

usar, para cada aplicação, os valores propostos/usados pelos autores em suas aplicações, isto é, para o K-médias e o K-médias aparado considerou-se o critério de parada igual a $0,0001$ (as observações não trocarão mais de grupos se o critério (3.1) atingir o valor $0,0001$) e o número de inicializações aleatórias assim como o número de etapas por inicialização foram fixados em 10. Ainda para o K-médias aparado, fez se um corte de $\alpha = 10\%$ de N. O mesmo se repete para os métodos K-médias ϕ e K-médias ϕ aparado. Para o *CLARANS*, considerou-se *numlocal* igual 2 e *maxneighbor* igual $1,25\%$ de $K(n-K)$. E para o *Hill Climbing*, o processo de busca do algoritmo finaliza quando o menor valor do critério (3.7) é encontrado ou quando o tamanho do laço é alcançado.

4.1.3.1 Resultados para os dados simulados de cubos e paralelepípedos

As Tabelas 2 e 3 expõem os resultados das medidas de validação para as aplicações dos algoritmos aos dados simulados de cubos e paralelepípedos, e os tempos de execução para o cenário de isotropia, considerando os diferentes graus de dispersão e os números de marcos anatômicos iguais a $k = 8$ e $k = 34$, respectivamente.

Tabela 2 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de cubos e paralelepípedos representados por 8 marcos anatômicos, sob isotropia

ALGORITMOS	**Dados isotrópicos (k=8)**								
	$\sigma_1 = \sigma_2 = 0,5$			$\sigma_1 = \sigma_2 = 1,5$			$\sigma_1 = \sigma_2 = 6$		
	IRA	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**
K-médias	1,0000	0,9049	23,57 s	1,0000	0,7177	22,59 s	0,7369	0,1429	32,46 s
K-médias ϕ	1,0000	0,9049	20,34 s	1,0000	0,7177	21,45 s	0,7369	0,1429	30,56 s
K-médias aparado	1,0000	0,9046	44,03 s	1,0000	0,7133	53,06 s	0,8696	0,1617	48,35 s
K-médias ϕ aparado	1,0000	0,9067	16,13 s	1,0000	0,7283	25,34 s	0,8696	0,1621	21,31 s
CLARANS	1,0000	0,9049	6,06 m	1,0000	0,7177	6,01 m	0,5137	0,1204	5,15 m
Hill Climbing	1,0000	0,9049	2,25 m	1,0000	0,7177	2,54 m	0,6691	0,1391	2,01 m

Fonte: Elaborada pelo o autor (2021).

A partir dos resultados exibidos na Tabela 2, para as situações de isotropia com $k = 8$, observa-se um comportamento semelhante para os casos com dispersões iguais a $\sigma = 0,5$ e $\sigma = 1,5$, ou seja, os desempenhos medidos pelo índice de Rand Ajustado atingiram o valor 1 em todas as aplicações dos algoritmos, indicando que nestes casos os algoritmos aparentemente possuem uma eficácia de agrupamento boa. Quanto aos valores do índice de Silhueta, os

valores mais significativos obtidos foram para o método K-médias ϕ aparado, $0,9067$ para o caso de dispersão de $\sigma = 0,5$ e $0,7283$ para o caso de $\sigma = 1,5$. Já para o caso com alta dispersão $(\sigma = 6)$, observa-se um comportamento variado para os desempenhos medidos pelos índices para todos os métodos. Os métodos K-médias e K-médias ϕ apresentaram as mesmas medições para o índice de Rand Ajustado, $0,7369$, e para o índice de Silhueta, $0,1429$. Os métodos K-médias aparado e K-médias ϕ aparado apresentaram as mesmas medições para o índice de Rand Ajustado, $0,8696$, mas medições diferentes para o índice de Silhueta. Os métodos *CLARANS* e *Hill Climbing* apresentaram valores menores dos índices, mas próximos aos dos demais métodos. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

Tabela 3 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de cubos e paralelepípedos representados por 34 marcos anatômicos, sob isotropia

ALGORITMOS	**Dados isotrópicos (k=34)**								
	$\sigma_1 = \sigma_2 = 0,5$			$\sigma_1 = \sigma_2 = 1,5$			$\sigma_1 = \sigma_2 = 6$		
	IRA	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**
K-médias	1,0000	0,6085	28,59 s	1,0000	0,2069	45,24 s	0,3304	0,0405	49,14 s
K-médias ϕ	1,0000	0,6085	28,37 s	1,0000	0,2069	35,02 s	0,4305	0,0479	47,31 s
K-médias aparado	1,0000	0,6033	29,36 s	1,0000	0,2221	35,04 s	0,0867	0,0159	34,33 s
K-médias ϕ aparado	1,0000	0,6197	27,27 s	1,0000	0,2261	34,58 s	0,2099	0,0606	30,08 s
CLARANS	1,0000	0,6085	6,53 m	0,8081	0,1911	8,01 m	0,1854	0,0217	5,44 m
Hill Climbing	1,0000	0,6085	4,13 m	0,8824	0,1949	6,25 m	0,0158	0,0079	3,54 m

Fonte: Elaborada pelo o autor (2021).

A partir dos resultados exibidos na Tabela 3, para as situações de isotropia com $k = 34$, observa-se um comportamento semelhante para os casos com dispersões iguais a $\sigma = 0,5$ e $\sigma = 1,5$, ou seja, os desempenhos medidos pelo índice de Rand Ajustado atingiram o valor 1 na maioria das aplicações dos algoritmos. Com exceções dos métodos *CLARANS* e *Hill Climbing* para o caso de $\sigma = 1,5$, que apresentaram valores do IRA de $0,8081$ e $0,8824$, respectivamente. Para todos esses casos, há indícios de que os algoritmos aparentemente possuem uma eficácia de agrupamento boa. Quanto aos valores do índice de Silhueta, os valores mais significativos obtidos foram para o método K-médias ϕ aparado, $0,6197$ para o caso de dispersão igual a $\sigma = 0,5$ e $0,2261$ para o caso de $\sigma = 1,5$. Já para o caso com

alta dispersão $(\sigma = 6)$, observa-se um comportamento variado para os desempenhos medidos pelos índices para todos os métodos. As versões dos métodos K-médias e K-médias aparado utilizando a média ϕ apresentaram melhores valores para ambos os índices em relação as suas versões utilizando a média de Procrustes. E o *CLARANS* apresentou desempenho melhor em relação ao método K-médias aparado, considerando os valores para ambos os índices. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

As Tabelas 4 e 5 expõem os resultados das medidas de validação para as aplicações dos algoritmos aos dados simulados de cubos e paralelepípedos, e os tempos de execução para o cenário de anisotropia, considerando os diferentes graus de variação e os números de marcos anatômicos iguais a $k = 8$ e $k = 34$, respectivamente. Lembrando que para os casos anisotrópicos, a matriz de covariância foi gerada de modo aleatório. Variação 1 significa que todos os valores da matriz de covariância estão no intervalo $(1, 10)$ e Variação 2 significa que todos os valores estão no intervalo $(10, 30)$.

A partir dos resultados exibidos na Tabela 4, para as situações de anisotropia com $k = 8$, observa-se um comportamento semelhante para os casos considerando a Variação 1, ou seja, os desempenhos medidos pelo índice de Rand Ajustado atingiram o valor 1 em todas as aplicações dos algoritmos, indicando que nestes casos os algoritmos aparentemente possuem uma eficácia de agrupamento boa. Quanto aos valores do índice de Silhueta, o valor mais significativo obtido foi para o método K-médias ϕ aparado, $0,5742$. Já para os casos considerando a Variação 2, observa-se um comportamento variado para os desempenhos medidos pelos índices. Os valores positivos obtidos pelo IRA foram para os métodos K-médias ϕ aparado $(0, 0067)$, *CLARANS* $(0, 0105)$ e *Hill Climbing* $(0, 0303)$, indicando que para esses casos os demais algoritmos não apresentam aparentemente uma eficácia de agrupamento boa, levando em consideração esse valores. Enquanto que os valores mais significativos obtidos pelo Silhueta foram para os métodos K-médias ϕ aparado $(0, 0538)$ e *CLARANS* $(0, 0487)$. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores, na maioria das situações, em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

A partir dos resultados exibidos na Tabela 5, para as situações de anisotropia com $k =$

Tabela 4 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de cubos e paralelepípedos representados por 8 marcos anatômicos, sob anisotropia

ALGORITMOS	Dados anisotrópicos (k=8)					
	Variação 1			Variação 2		
	IRA	Silhueta	Tempo E.	IRA	Silhueta	Tempo E.
K-médias	1,0000	0,5484	19,01 s	-0,0037	0,0453	1,22 m
K-médias ϕ	1,0000	0,5484	16,02 s	-0,0037	0,0453	1,23 m
K-médias aparado	1,0000	0,5712	51,19 s	-0,0066	0,0401	1,39 m
K-médias ϕ aparado	1,0000	0,5742	16,59 s	0,0067	0,0538	47,42 s
CLARANS	1,0000	0,5484	1,09 m	0,0105	0,0487	2,25 m
Hill Climbing	1,0000	0,5484	1,54 m	0,0303	0,0428	2,12 m

Fonte: Elaborada pelo o autor (2021).

34, observa-se um comportamento semelhante para os casos considerando a Variação 1, ou seja, os desempenhos medidos pelo índice de Rand Ajustado atingiram o valor 1 na maioria das aplicações dos algoritmos. Com exceções dos métodos *CLARANS* e *Hill Climbing*, que apresentaram valores do IRA iguais a $0,5735$ e $0,6364$, respectivamente. Já para os casos considerando a Variação 2, observa-se que as versões dos métodos K-médias e K-médias aparado utilizando a média ϕ apresentaram melhores valores para ambos os índices em relação as suas versões utilizando a média de Procrustes. E o *CLARANS* apresentou desempenho melhor em relação aos métodos K-médias e *Hill Climbing*, considerando os valores para o IRA. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

Para o cenário de isotropia, os marcos anatômicos possuem aproximadamente a mesma variabilidade. Para as situações com $k = 8$ e $k = 34$, as considerações sobre os desempenhos dos algoritmos medidos pelos índices são as seguintes. Para os casos de dispersão iguais a $\sigma = 0,5$ e $\sigma = 1,5$, aparentemente os valores obtidos pelo IRA indicam que os métodos possuem boas eficácias de agrupamento. Enquanto que para os valores obtidos pelo Silhueta, os valores obtidos por uma das nossas propostas, o K-médias ϕ aparado, foram mais significativos em relação aos demais métodos. Para o caso com alta dispersão, $\sigma = 6$, as versões do métodos K-médias e K-médias aparado utilizando a média ϕ apresentaram melhores desempenhos considerando ambos os índices em relação as suas versões utilizando a média de

Tabela 5 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de cubos e paralelepípedos representados por 34 marcos anatômicos, sob anisotropia

ALGORITMOS	Dados anisotrópicos (k=34)					
	Variação 1			Variação 2		
	IRA	Silhueta	Tempo E.	IRA	Silhueta	Tempo E.
K-médias	1,0000	0,1101	51,35 s	0,0303	0,0139	1,14 m
K-médias ϕ	1,0000	0,1101	42,39 s	0,0690	0,0154	58,47 s
K-médias aparado	1,0000	0,1195	32,18 s	-0,0032	0,0084	45,33 s
K-médias ϕ aparado	1,0000	0,1205	31,15 s	0,0738	0,0112	36,11 s
CLARANS	0,5735	0,0988	5,25 m	0,0388	0,0063	4,21 m
Hill Climbing	0,6364	0,0997	3,54 m	0,0045	0,0096	2,59 m

Fonte: Elaborada pelo o autor (2021).

Procrustes. Ainda, o método *CLARANS*, uma das propostas do presente trabalho, apresentou maiores valores para ambos índices em relação ao K-médias aparado.

Para o cenário de anisotropia, os marcos anatômicos não possuem aproximadamente a mesma variabilidade. Para as situações com $k = 8$ e $k = 34$, as considerações sobre os desempenhos dos algoritmos medidos pelos índices são as seguintes. Considerando a Variação 1, aparentemente os valores obtidos pelo IRA indicam que os métodos possuem boas eficácias de agrupamento, mas há indícios de que os métodos *CLARANS* e *Hill Climbing* apresentam uma eficácia menor em relação aos demais métodos para o caso com $k = 34$ marcos anatômicos. Já para a Variação 2, as versões dos métodos K-médias e K-médias aparado utilizando a média ϕ apresentaram valores superiores ou iguais para os índices em relação as suas versões que utilizam a média de Procrutes. E ainda, há casos em que algumas das propostas do presente trabalho, *CLARANS* e *Hill Climbing*, apresentaram desempenhos superiores em relação aos métodos K-médias e K-médias aparado, considerando pelo menos um dos valores dos índices.

De modo geral, quanto aos tempos de execução, era esperado que os algoritmos *CLARANS* e *Hill Climbing* apresentassem tempos maiores do que os demais métodos, uma vez que o processo iterativo desses algoritmos são baseados em busca. Quanto aos grupos formados pelos métodos K-médias e K-médias aparado utilizando a média ϕ, os tempos de execução são menores, na maioria dos casos, do que a adaptação utilizando a média de Procrustes e os grupos formados são aparentemente melhores. Era esperado que os métodos *CLARANS* e *Hill Climbing* apresentassem melhores desempenhos em relação ao K-médias (por ser influenciado por valores discrepantes), mas os desempenhos apresentados só são melhores em alguns casos,

considerando pelo menos um dos índices medidos, expressivamente no cenário de anisotropia. De modo geral, há indícios de que os resultados obtidos pelos os algoritmos apresentados, tanto os propostos como os já presentes na literatura, para o agrupamento de formas 3D foram eficientes para esses conjuntos de dados analisados. Quanto ao desempenho em específico das propostas do presente trabalho, os métodos K-médias ϕ aparado, *CLARANS* e *Hill Climbing* apresentaram bons desempenhos quanto aos grupos formados, se sobressaindo em muitas situações em relação aos outros métodos já implementados na literatura.

4.1.3.2 Resultados para os dados simulados de representações geométricas aleatórias

As Tabelas 6 e 7 expõem os resultados das medidas de validação para as aplicações dos algoritmos em dados de formas simuladas com configurações médias definidas por pontos de uma distribuição uniforme com diferentes intervalos e os tempos de execução para o cenário de isotropia, considerando os diferentes graus de dispersão e os números de marcos anatômicos iguais a $k = 6$ e $k = 36$, respectivamente.

Tabela 6 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de formas com configurações médias definidas por representações geométricas aleatórias representados por 6 marcos anatômicos, sob isotropia

ALGORITMOS	**Dados isotrópicos (k=6)**								
	$\sigma_1 = \sigma_2 = 0,5$			$\sigma_1 = \sigma_2 = 1,5$			$\sigma_1 = \sigma_2 = 6$		
	IRA	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**
K-médias	0,7721	0,4026	36,12 s	0,2059	0,2254	56,39 s	-0,0085	0,0668	1,11 m
K-médias ϕ	0,7721	0,4026	33,45 s	0,2647	0,2142	41,01 s	-0,0098	0,0618	1,01 m
K-médias aparado	0,9555	0,4733	45,38 s	0,2088	0,1062	34,28 s	-0,0033	0,0625	1,35 m
K-médias ϕ aparado	0,9555	0,4782	43,43 s	0,3006	0,2473	27,51 s	-0,0093	0,0605	46,36 s
CLARANS	0,4853	0,3612	1,02 m	0,2859	0,1906	7,54 m	0,0225	0,0475	3,39 m
Hill Climbing	0,5735	0,3801	1,59 m	0,2244	0,1768	3,12 m	-0,0097	0,0678	2,49 m

Fonte: Elaborada pelo o autor (2021).

A partir dos resultados exibidos na Tabela 6, para as situações de isotropia com $k = 6$, observa-se um comportamento variado para os casos com dispersões iguais a $\sigma = 0,5$ e $\sigma = 1,5$. Os desempenhos medidos para as versões do K-médias e K-médias aparado utilizando a média ϕ pelo índice de Rand Ajustado foram melhores ou iguais ao desempenhos em relação as suas versões utilizando a média de Procrustes. Quanto aos valores do índice

de Silhueta, os valores mais significativos obtidos foram para o método K-médias ϕ aparado, $0,4782$ para o caso de dispersão de $\sigma = 0,5$ e $0,2473$ para o caso de $\sigma = 1,5$. Ainda, para o caso de $\sigma = 1,5$, os métodos *CLARANS* e *Hill Climbing*, propostas da dissertação, apresentaram melhores desempenhos em relação aos métodos K-médias e K-médias aparado, métodos já implementados na literatura, considerandos os valores do IRA. Já para o caso com alta dispersão $(\sigma = 6)$, observa-se um comportamento variado para os desempenhos medidos pelos índices para todos os métodos. O único valor positivo obtido pelo o IRA foi para o método *CLARANS*. E os valores mais significativos obtidos pelo Silhueta foram para os métodos K-médias e *Hill Climbing*. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

Tabela 7 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de formas com configurações médias definidas por representações geométricas aleatórias representados por 36 marcos anatômicos, sob isotropia

ALGORITMOS	**Dados isotrópicos (k=36)**								
	$\sigma_1 = \sigma_2 = 0,5$			$\sigma_1 = \sigma_2 = 1,5$			$\sigma_1 = \sigma_2 = 6$		
	IRA	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**
K-médias	1,0000	0,5733	44,27 s	0,8824	0,2580	50,23 s	0,2848	0,0196	3,23 m
K-médias ϕ	1,0000	0,5733	33,09 s	0,8824	0,2580	48,14 s	0,4851	0,0189	2,43 m
K-médias aparado	1,0000	0,6116	1,11 m	0,9555	0,2937	55,31 s	0,3008	0,0169	1,15 m
K-médias ϕ aparado	1,0000	0,6119	36,28 s	1,0000	0,3001	41,21 s	0,6004	0,0239	1,18 m
CLARANS	1,0000	0,5733	5,31 m	0,5735	0,2259	3,35 m	0,1071	0,0087	6,13 m
Hill Climbing	1,0000	0,5733	3,42 m	0,4853	0,2205	5,15 m	0,0227	0,0092	5,05 m

Fonte: Elaborada pelo o autor (2021).

A partir dos resultados exibidos na Tabela 7, para as situações de isotropia com $k = 36$, observa-se um comportamento variado para os casos com dispersões iguais a $\sigma = 0,5$ e $\sigma = 1,5$. Os desempenhos medidos para as versões do K-médias e K-médias aparado utilizando a média ϕ pelo índice de Rand Ajustado foram melhores ou iguais aos desempenhos em relação as suas versões utilizando a média de Procrustes. Quanto aos valores do índice de Silhueta, os valores mais significativos obtidos foram para o método K-médias ϕ aparado, $0,6119$ para o caso de dispersão de $\sigma = 0,5$ e $0,3001$ para o caso de $\sigma = 1,5$. Ainda, os métodos *CLARANS* e *Hill Climbing*, propostas da dissertação, apresentaram melhores desempenhos em

relação aos métodos K-médias e K-médias aparado, métodos já implementados na literatura, considerandos os valores do IRA. Já para o caso com alta dispersão ($\sigma = 6$), observa-se um comportamento variado para os desempenhos medidos pelos índices para todos os métodos. As versões do K-médias e do K-médias aparado que utilizam a média ϕ apresentaram maiores valores, considerando ambos os índices, em relação as suas versões que utilizam a média de Procrustes. E ainda, os métodos propostos nesta dissertação, *CLARANS* e *Hill Climbing*, não apresentaram bons desempenhos em relação aos demais métodos para esse cenário. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

As Tabelas 8 e 9 expõem os resultados das medidas de validação para as aplicações dos algoritmos em dados de formas simuladas com configurações médias definidas por pontos de uma distribuição uniforme com diferentes intervalos e os tempos de execução para o cenário de anisotropia, considerando os diferentes graus de dispersão, $\gamma = 2$ e os números de marcos anatômicos iguais a $k = 6$ e $k = 36$, respectivamente.

Tabela 8 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de formas com configurações médias definidas por representações geométricas aleatórias representados por 6 marcos anatômicos, sob anisotropia

ALGORITMOS	**Dados anisotrópicos (k=6)**								
	$\sigma_1 = \sigma_2 = 0,5$ e $\gamma = 2$			$\sigma_1 = \sigma_2 = 1,5$ e $\gamma = 2$			$\sigma_1 = \sigma_2 = 6$ e $\gamma = 2$		
	IRA	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**
K-médias	0,9599	0,5409	29,59 s	0,3308	0,2692	50,41 s	0,0388	0,0721	1,54 m
K-médias ϕ	0,9599	0,5409	24,04 s	0,3308	0,2692	40,34 s	0,0388	0,0721	59,32 s
K-médias aparado	0,9555	0,5932	31,58 s	0,3801	0,2689	1,04 m	0,0012	0,0471	37,38 s
K-médias ϕ aparado	1,0000	0,5997	25,14 s	0,4381	0,3094	39,02 s	0,1333	0,0663	37,02 s
CLARANS	0,8081	0,4957	2,14 m	0,2442	0,2664	2,55 m	0,0826	0,0812	5,37 m
Hill Climbing	0,9599	0,5409	3,06 m	0,2646	0,2584	4,01 m	0,0159	0,0746	1,59 m

Fonte: Elaborada pelo o autor (2021).

A partir dos resultados exibidos na Tabela 8, para as situações de anisotropia com $k = 6$, observa-se um comportamento variado para os casos com dispersões iguais a $\sigma = 0,5$ e $\sigma = 1,5$. Os desempenhos medidos para as versões do K-médias e K-médias aparado utilizando a média ϕ pelo índice de Rand Ajustado foram melhores ou iguais ao desempenhos em relação as suas versões utilizando a média de Procrustes. Quanto aos valores do índice de Silhueta,

os valores mais significativos obtidos foram para o método K-médias ϕ aparado, $0,5997$ para o caso de $\sigma = 0,5$ e $0,3094$ para o caso de $\sigma = 1,5$. Ainda, os métodos *CLARANS* e *Hill Climbing*, propostas da dissertação, apresentaram bons desempenhos quando $\sigma = 0,5$. Já para o caso com alta dispersão $(\sigma = 6)$, observa-se um comportamento variado para os desempenhos medidos pelos índices para todos os métodos. As versões do K-médias e do K-médias aparado utilizando a média ϕ apresentaram valores melhores ou iguais para ambos os índices em relação as suas versões utilizando a média de Procrustes. Ainda, o método *CLARANS* apresentou maior valor para o IRA, em relação aos métodos já implementados na literatura (K-médias, K-médias ϕ e K-médias aparado). E o *Hill Climbing* apresentou maior valor em relação aos métodos já implementados somente no que diz respeito ao valor do Silhueta. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

Tabela 9 – Validação de agrupamento e tempo de execução dos algoritmos aplicados aos dados simulados de formas com configurações médias definidas por representações geométricas aleatórias representados por 36 marcos anatômicos, sob anisotropia

ALGORITMOS	**Dados anisotrópicos (k=36)**								
	$\sigma_1 = \sigma_2 = 0,5$ e $\gamma = 2$			$\sigma_1 = \sigma_2 = 1,5$ e $\gamma = 2$			$\sigma_1 = \sigma_2 = 6$ e $\gamma = 2$		
	IRA	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**	**IRA**	**Silhueta**	**Tempo E.**
K-médias	1,0000	0,6628	36,11 s	1,0000	0,3521	43,28 s	0,4853	0,0571	1,23 m
K-médias ϕ	1,0000	0,6628	32,34 s	1,0000	0,3521	35,18 s	0,4853	0,0571	1,58 m
K-médias aparado	1,0000	0,6935	46,05 s	1,0000	0,3923	50,18 s	0,2303	0,0368	1,23 m
K-médias ϕ aparado	1,0000	0,6953	30,02 s	1,0000	0,3924	43,25 s	0,6724	0,0417	1,15 m
CLARANS	1,0000	0,6628	5,58 m	0,7721	0,3171	6,40 m	0,3541	0,0401	6,37 m
Hill Climbing	1,0000	0,6628	6,23 m	0,8824	0,3354	4,19 m	0,0096	0,0086	7,39 m

Fonte: Elaborada pelo o autor (2021).

A partir dos resultados exibidos na Tabela 9, para as situações de anisotropia com $k = 36$, observa-se um comportamento semelhante para os casos com dispersões iguais a $\sigma = 0,5$ e $\sigma = 1,5$, ou seja, os desempenhos medidos pelo índice de Rand Ajustado atingiram o valor 1 na maioria das aplicações dos algoritmos. Com exceções dos métodos *CLARANS* e *Hill Climbing* para o caso de $\sigma = 1,5$, que apresentaram valores do IRA de $0,7721$ e $0,8824$, respectivamente. Para todos esses casos, há indícios de que os algoritmos aparentemente

possuem uma eficácia de agrupamento boa. Quanto aos valores do índice de Silhueta, os valores mais significativos obtidos foram para o método K-médias ϕ aparado, $0,6953$ quando $\sigma = 0,5$ e $0,3924$ para o caso de $\sigma = 1,5$. Já para o caso com alta dispersão $(\sigma = 6)$, observa-se um comportamento variado para os desempenhos medidos pelos índices para todos os métodos. As versões do K-médias e do K-médias aparado utilizando a média ϕ apresentaram valores maiores ou iguais para ambos os índices em relação as suas versões utilizando a média de Procrustes. O *CLARANS* apresentou desempenho melhor em relação ao método K-médias aparado, considerando os valores para ambos os índices. Os valores dos tempos de execução para os métodos K-médias e K-médias aparado utilizando a média ϕ foram menores em relação as suas versões utilizando a média de Procrustes e os tempos de execução foram maiores paras os métodos *CLARANS* e *Hill Climbing*, em todos os cenários.

Para o cenário de isotropia, para as situações com $k = 6$, as considerações sobre os desempenhos dos algoritmos medidos pelos índices são as seguintes. Para o caso de dispersão igual a $\sigma = 0,5$, aparentemente os valores obtidos pelo IRA indicam que a maioria dos métodos possuem boas eficácias de agrupamento. Para o caso de dispersão igual a $\sigma = 1,5$. Há indícios de que todos os métodos apresentam eficácias de agrupamento razoáveis, mas é possível observar que as versões dos métodos K-médias e K-médias aparado utilizando a média ϕ possuem maiores valores que suas versões utilizando a média de Procrustes e ainda, o método *CLARANS* apresentou maiores valores para o IRA em relação aos métodos já implementados na literatura (K-médas, K-médias ϕ e K-médias aparado). E o *Hill Climbing* apresentou maiores valores em relação aos métodos K-médias e K-médias aparado. Enquanto que para os valores obtidos pelo Silhueta, os valores obtidos por uma das nossas propostas, o K-médias ϕ aparado, foram mais significativos em relação aos demais métodos, para ambos os cenários de dispersão iguais a $\sigma = 0,5$ e $\sigma = 1,5$. Para os casos com alta dispersão, os valores apresentados pelos índices foram baixos para todos os métodos. Em relação às situações considerando $k = 36$, para os casos com dispersão iguais a $\sigma = 0,5$ e $\sigma = 1,5$, aparentemente os valores obtidos pelo IRA indicam que a maioria dos métodos possuem boas eficácias de agrupamento, com exceções do *CLARANS* e *Hill Climbing*, que obtiveram valores um pouco menores para o caso de dispersão igual a $\sigma = 1,5$. Há indícios de que todos os métodos apresentam eficácias de agrupamento boas. Quanto aos valores do índice de Silhueta, os valores mais significativos obtidos foram para o método K-médias ϕ aparado. Para o caso com $\sigma = 6$ é possível observar que as versões dos métodos K-médias e K-médias aparado utilizando a média ϕ possuem maiores valores que suas versões utilizando a média de Procrustes, analisando ambos os índices. Para o cenário de

anisotropia, de modo geral, os resultados obtidos seguem um panorama bem semelhante para os casos de isotropia, isto é, há casos em que as propostas do trabalho apresentam melhores valores em relação aos métodos implementados por outros autores.

De modo geral, quanto aos tempos de execução, os algoritmos *CLARANS* e *Hill Climbing* apresentaram tempos de execução maiores do que os demais métodos. Quanto aos grupos formados pelos métodos K-médias e K-médias aparado utilizando a média ϕ, os tempos de execução são menores, na maioria dos casos, do que a adaptação utilizando a média de Procrustes e os grupos formados são aparentemente melhores, na maioria dos casos. Ainda, os métodos *CLARANS* e *Hill Climbing* apresentaram desempenhos melhores em relação à alguns dos demais métodos, considerando pelo menos um dos índices analisados, expressivamente nos casos com um número pequeno de marcos anatômicos. É importante detalhar que para esse experimento, a configuração média das formas foi definida de modo aleatório por pontos de uma distribuição uniforme e no cenário em que os marcos não possuem aproximadamente a mesma variabilidade, todos os algoritmos apresentaram eficácia de agrupamento baixa. De modo geral, os resultados mostraram que os algoritmos apresentados para o agrupamento de formas 3D foram relativamente eficientes para esses conjuntos de dados analisados. Quanto ao desempenho em específico das propostas do presente trabalho, os métodos K-médias ϕ aparado, *CLARANS* e *Hill Climbing* apresentaram bons desempenhos quanto aos grupos formados, se sobressaindo (em muitas situações) em relação aos outros métodos já implementados na literatura.

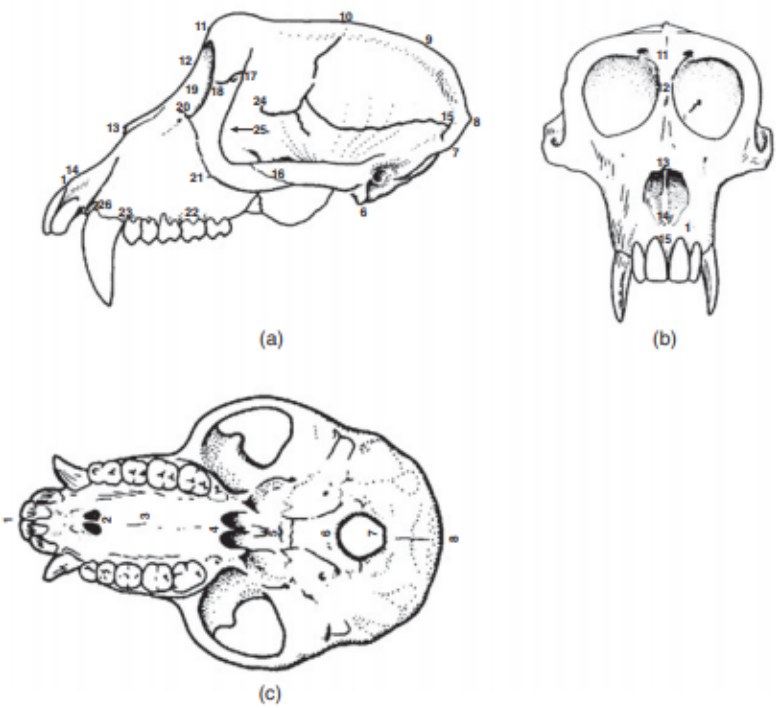
4.2 CONJUNTOS DE DADOS REAIS

Para aplicações de dados reais nesta pesquisa, foram considerados três conjuntos de dados de formas de objetos tridimensionais. Os conjuntos de dados de Crânios de Macacos e de Cérebros de Adultos Saudáveis são conjuntos disponíveis na literatura de análise estatística de formas e podem ser encontrados no pacote **shapes** (DRYDEN, 2012) do R. Enquanto que o conjunto de dados de Ossos do Nariz Humano pode ser encontrado no pacote **Morpho** (SCHLAGER et al., 2020). Quanto aos valores dos parâmetros utilizados pelos algoritmos adaptados, fez-se uso dos mesmos valores utilizados nas aplicações dos dados simulados, com exceções dos métodos K-médias aparado e K-médias ϕ aparado. Agora o tamanho do corte foi de $\alpha = 5\%$ de N, pois os valores de N para os dados reais são menores que 100.

4.2.1 Crânios de Macacos

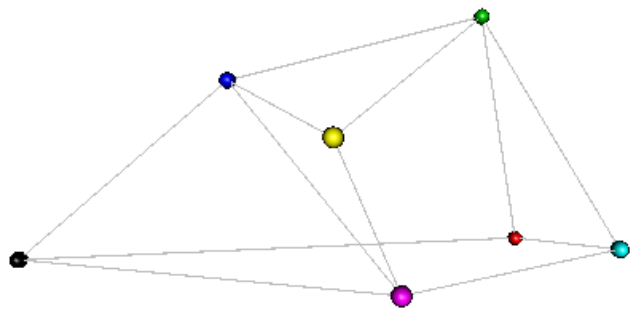
Em um estudo para avaliar a existência de diferenças de tamanho e forma entre os crânios de macacos machos e fêmeas da espécie *Macaca fasciculares*, uma amostra de 18 indivíduos (9 machos e 9 fêmeas) foi obtida por Paul O'Higgins *(Hall-York Medical Scholl)*. Um subconjunto de $k = 7$ marcos anatômicos foram selecionados de um total de $k = 26$ que representam cada crânio. Os nomes dos marcos selecionados são: *1-prosthion*, *2-nasion*, *3-bregma*, *4-opisthion*, *5-asterion*, *6-interfrontomalare* e *7-midpoint of zyg/temp suture* (DRYDEN; MARDIA, 1993); (DRYDEN; MARDIA, 2016). A impressão de um artista da representação de um crânio em 3D com as projeções dos marcos anatômicos pode ser visualizada na Figura 8. Já a Figura 9 representa a configuração da forma do crânio de um macaco fêmea com os sete marcos anatômicos obtidos, com cada marco anatômico exibido por uma cor diferente. A Figura 10 ilustra a forma média de Procrustes $[\hat{\mu}]$ para o conjunto de dados de crânios de macacos da espécie *Macaca fascicularis*. Para acessar o conjunto de dados pelo R, faz-se uso da função *data(macaques)* com o pacote **shapes** carregado. Mais detalhes sobre os dados podem ser encontrados em Dryden e Mardia (1993) e Dryden e Mardia (2016).

Figura 8 – Representação 3D de um crânio de macaco da espécie *Macaca fascicularis*: (a) vista lateral; (b) vista frontal; e (c) vista inferior



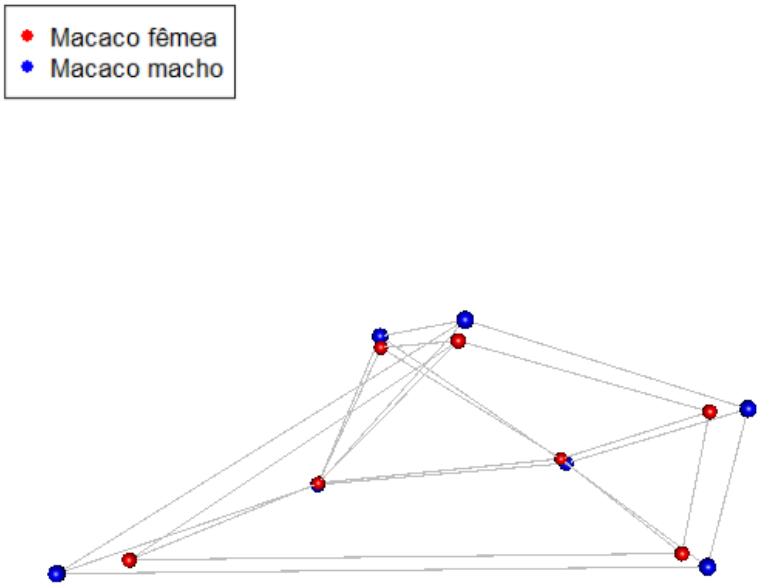
Fonte: (DRYDEN; MARDIA, 2016).

Figura 9 – Configurações do crânio de um macaco fêmea da espécie *Macaca fascicularis*



Fonte: Elaborada pelo o autor (2021).

Figura 10 – Forma média de Procrustes $[\hat{\mu}]$ de crânios de macacos da espécie *Macaca fascicularis*



Fonte: Elaborada pelo o autor (2021).

4.2.1.1 Resultados para os dados de Crânios de Macacos

A Tabela 10 exibe os resultados obtidos pelos índices de Rand Ajustado e Silhueta para os agrupamentos realizados pelos métodos K-médias, K-médias ϕ, K-médias aparado, K-médias

ϕ aparado, *CLARANS* e *Hill Climbing*, para o conjunto de dados de Crânios de Macacos.

Tabela 10 – Índice de Rand Ajustado e Silhueta para os dados de Crânios de Macacos

ALGORITMOS	**Índices**		**Tempo E.**
	IRA	**Silhueta**	
K-médias	0,0267	0,3688	3,53 s
K-médias ϕ	0,0267	0,3688	2,01 s
K-médias aparado	0,0033	0,3780	4,29 s
K-médias ϕ aparado	0,0033	0,3780	3,53 s
CLARANS	0,0555	0,0689	5,32 s
Hill Climbing	0,2676	0,0705	28,05 s

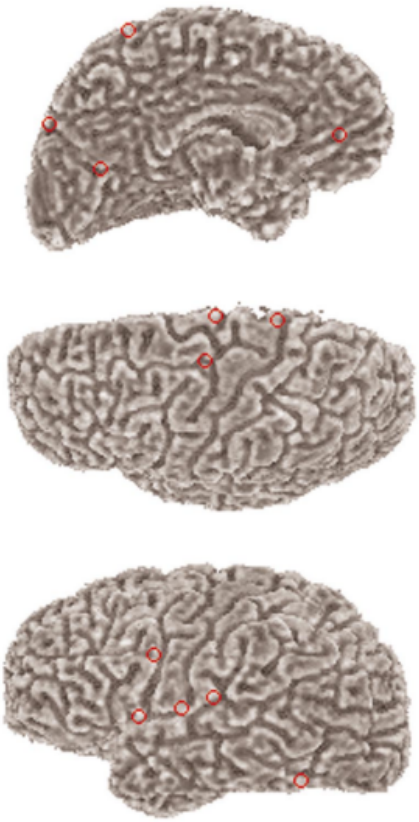
Fonte: Elaborada pelo o autor (2021).

Pode-se perceber nos resultados apresentados na Tabela 10 que, para os métodos K-médias e K-médias ϕ, o valor obtido pelo IRA foi de $0,0267$ e para o Shilhoueta foi de $0,3688$. Em relação aos métodos K-médias aparado e K-médias ϕ aparado, os valores obtidos foram de $0,0033$ para o IRA e $0,3780$ para o Silhueta. Para o método *CLARANS* o valor obtido pelo IRA foi de $0,0555$ e do Silhueta foi de $0,0689$. Já em relação ao *Hill Climbing*, o valor obtido pelo IRA foi de $0,2676$ e pelo Silhueta foi de $0,0705$. Nota-se que os valores obtidos pelo o IRA, para os métodos *CLARANS* e *Hill Climbing* foram maiores em relação aos demais métodos. Enquanto que segundo os valores do Silhueta, os métodos K-médias e K-médias aparado e suas versões utilizando a média ϕ apresentaram os maiores valores. É possível observar ainda que as versões dos métodos K-médias e K-médias aparado utilizando a média ϕ apresentaram menor tempo de execução em relação às suas versões que utilizam a média de Procrustes. Percebe-se que os métodos propostos no presente trabalho, K-médias ϕ aparado, *CLARANS* e *Hill Climbing*, apresentaram desempenhos bons em relação aos métodos já implementados na literatura, considerando pelo menos um dos índices analisados, considerado os valores obtidos.

4.2.2 Cérebros de Adultos Saudáveis

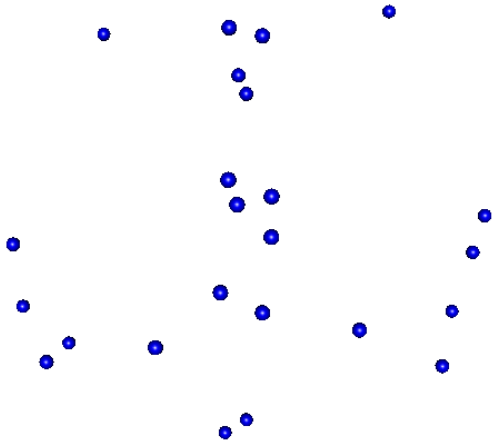
Em uma investigação para verificar a diferença entre as formas de cérebros humanos adultos, coletou-se marcos anatômicos distribuídos pela superfície do córtex do cérebro de 58 adultos saudáveis, 27 são do sexo masculino e 31 são do sexo feminino. O conjunto de dados se divide em quatro grupos distintos: 24 homens destros, 7 homens canhotos, 19 mulheres destras e 8 mulheres canhotas. Foram identificados $k = 12$ marcos anatômicos em cada hemisfério do cérebro, contabilizando um total de $k = 24$ marcos por indivíduo. A Figura 11 apresenta três visualizações do hemisfério esquerdo de um indivíduo indicando as localizações aproximadas dos marcos anatômicos. A Figura 12 exibe a representação em 3D gerada pelo R dos $k = 24$ marcos anatômicos localizados no cérebro de um indivíduo. Para acessar o conjunto de dados pelo R, faz-se uso da função *data(brains)* com o pacote **shapes** carregado. Mais detalhes sobre os dados podem ser encontrados em Free et al. (2001).

Figura 11 – Visualização das localizações dos k=12 marcos anatômicos do hemisfério esquerdo de um indivíduo



Fonte: (FREE et al., 2001).

Figura 12 – Representação em 3D gerada pelo R dos k=24 marcos anatômicos do cérebro de um indivíduo



Fonte: Elaborada pelo o autor (2021).

4.2.2.1 Resultados para os dados de Cérebros de Adultos Saudáveis

A Tabela 11 exibe os resultados obtidos pelos índices de Rand Ajustado e Silhueta para os agrupamentos realizados pelos métodos K-médias, K-médias ϕ, K-médias aparado, K-médias ϕ aparado, *CLARANS* e *Hill Climbing*, para o conjunto de dados de Cérebros de Adultos Saudáveis.

Tabela 11 – Índice de Rand Ajustado e Silhueta para os dados de Cérebros de Adultos Saudáveis

ALGORITMOS	**Índices**		**Tempo E.**
	IRA	**Silhueta**	
K-médias	-0,0262	0,0383	28,00 s
K-médias ϕ	0,0340	0,0427	27,50 s
K-médias aparado	0,0318	0,0156	21,25 s
K-médias ϕ aparado	0,0442	0,0362	18,43 s
CLARANS	0,0395	0,0002	36,27 s
Hill Climbing	0,0503	0,0303	2,52 m

Fonte: Elaborada pelo o autor (2021).

Pode-se perceber nos resultados apresentados na Tabela 11 que, para o métodos K-médias o valor obtido pelo IRA foi de $-0,0262$ e para o Shilhouette foi de $0,0383$, enquanto que para o método K-médias ϕ os valores obtidos foram de $0,0340$ para o IRA e $0,0427$ para o Silhueta. Em relação ao método K-médias aparado os valores obtidos foram de $0,0318$ para o IRA e $0,0156$ para o Silhueta, enquanto que para o método K-médias ϕ aparado os valores foram $0,0442$ e $0,0362$ para o IRA e o Silhueta, respectivamente. Para o método *CLARANS* o valor obtido pelo IRA foi de $0,0395$ e do Silhueta foi de $0,0002$. Já em relação ao *Hill Climbing*, o valor obtido pelo IRA foi de $0,0503$ e pelo Silhueta foi de $0,0303$. As versões dos métodos K-médias e K-médias aparado utilizando a média ϕ apresentaram maiores valores para ambos os índices e menores tempos de execução em relação as suas versões utilizando a média de Procrustes. Nota-se que pelos valores do IRA, os métodos K-médias ϕ aparado, *CLARANS* e *Hill Climbing*, todas propostas do presente trabalho, obtiveram maiores valores em relação aos métodos já implementados na literatura. Desse modo, se considerarmos esses valores, os métodos propostos na presente dissertação apresentaram um bom desempenho em relação aos demais métodos.

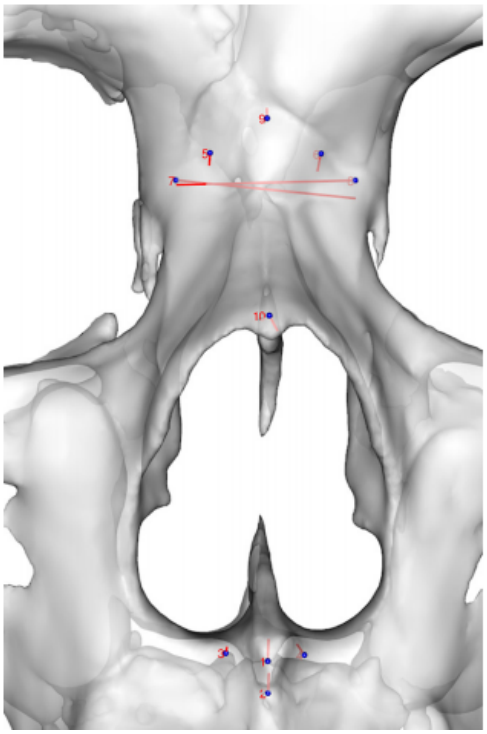
4.2.3 Ossos do Nariz Humano

O conjunto de dados de Ossos do Nariz Humano encontra-se disponível no pacote **Morpho** do R. Coletou-se manualmente $k = 10$ marcos anatômicos na superfície óssea de narizes humanos em um estudo para reconstrução de tecidos moles do nariz (SCHLAGER, 2013). Tem-se que quatro marcos anatômicos são colocados ao redor da coluna nasal e seis são colocados nos ossos nasais. O conjunto de dados possui um total de 80 indivíduos divididos em dois grupos, 40 homens e 40 mulheres. A Figura 13 apresenta a distribuição dos marcos ao redor dos ossos do nariz. A Figura 14 exibe a representação em 3D gerada pelo R dos $k = 10$ marcos anatômicos localizados ao redor dos ossos do nariz de um indivíduo. Para acessar o conjunto de dados pelo R, faz-se uso da função *data(boneData)* com o pacote **Morpho** carregado.

4.2.3.1 Resultados para os dados de Ossos do Nariz Humano

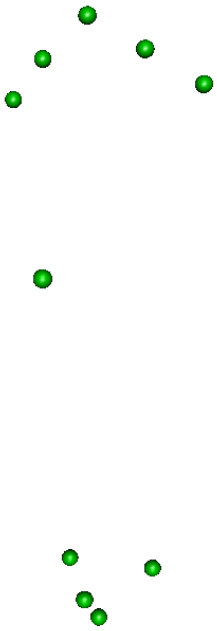
A Tabela 12 exibe os resultados obtidos pelos índices de Rand Ajustado e Silhueta para os agrupamentos realizados pelos métodos K-médias, K-médias ϕ, K-médias aparado, K-médias ϕ aparado, *CLARANS* e *Hill Climbing*, para o conjunto de dados de Ossos do Nariz Humano.

Figura 13 – Visualização das localizações dos k=10 marcos anatômicos ao redor da superfície óssea do nariz de um indivíduo



Fonte: (ZHENG; LI; SZEKELY, 2017).

Figura 14 – Representação em 3D gerada pelo R dos k=10 marcos anatômicos ao redor da superfície óssea de um nariz



Fonte: Elaborada pelo o autor (2021).

Pode-se perceber nos resultados apresentados na Tabela 12 que, para os métodos K-médias e K-médias ϕ, o valor obtido pelo IRA foi de $-0,0103$ e para o Silhueta foi de $0,2433$. Em

Tabela 12 – Índice de Rand Ajustado e Silhueta para os dados de Ossos do Nariz Humano

ALGORITMOS	**Índices**		**Tempo E.**
	IRA	**Silhueta**	
K-médias	-0,0103	0,2433	15,53 s
K-médias ϕ	-0,0103	0,2433	15,07 s
K-médias aparado	-0,0106	0,2439	18,42 s
K-médias ϕ aparado	-0,0106	0,2439	16,08 s
CLARANS	0,0119	0,1145	1,04 m
Hill Climbing	-0,0101	0,2325	1,38 m

Fonte: Elaborada pelo o autor (2021).

relação aos métodos K-médias aparado e K-médias ϕ aparado, os valores obtidos foram de $-0,0106$ para o IRA e $0,2439$ para o Silhueta. Para o método *CLARANS* o valor obtido pelo IRA foi de $0,0119$ e do Silhueta foi de $0,1145$. Já em relação ao *Hill Climbing*, o valor obtido pelo IRA foi de $-0,0101$ e pelo Silhueta foi de $0,2325$. Nota-se que pelos valores do IRA, o único método que apresentou valor positivo foi o *CLARANS*. Enquanto que segundo os valores do Silhueta, os métodos K-médias aparado e K-médias ϕ aparado apresentaram um desempenho melhor. É possível observar ainda que as versões dos métodos K-médias e K-médias aparado utilizando a média ϕ apresentaram menor tempo de execução em relação às versões que utilizam a média de Procrustes.

4.3 CONSIDERAÇÕES FINAIS

Em relação aos desempenhos dos algoritmos nos experimentos realizados em dados simulados, há indícios de que todos os métodos apresentaram bons desempenhos para os casos de isotropia e anisotropia com baixa dispersão/variabilidade. Para a maioria das aplicações, o método K-médias ϕ aparado apresentou performances superiores aos demais métodos. No que se refere aos casos de isotropia e anisotropia com dispersão/variabilidade alta, os métodos *CLARANS* e *Hill Climbing* apresentaram uma performance superior para os casos com um

número pequeno de marcos anatômicos, considerando pelo menos um dos índices analisados, especialmente para o experimento em que as configurações médias foram baseadas em uma forma não predefinida (geradas a partir de pontos da distribuição uniforme). Ainda, os métodos K-médias ϕ e K-médias ϕ aparado apresentaram performances superiores ou iguais aos métodos K-médias e K-médias aparado (os quais utilizam a média de Procrustes), tanto em tempo de execução, quanto em qualidade dos grupos, levando em conta os valores de pelo menos um dos índices, na maioria dos casos.

Observando os resultados relacionados aos dados reais, as performances dos algoritmos são influenciadas pelas singularidades de cada conjunto de dados. De modo geral, observou-se um panorama muito semelhante aos resultados obtidos para os conjuntos de dados simulados, isto é, existem situações em que as propostas da dissertação, o K-médias ϕ aparado, *CLARANS* e *Hill Climbing* apresentaram desempenhos superiores em relação aos métodos já implementados, considerandos os valores de pelo menos um dos índices analisados. Além do que, os menores tempos de execução pertencem às variações do K-médias e K-médias aparado utilizando a média ϕ. No entanto essas interpretações, tanto para os resultados obtidos para os conjuntos de dados simulados quanto para os conjuntos de dados reais, são baseadas nos valores obtidos pelos índices, porém esse valores em algumas situações são muitos baixos, fazendo-se necessário aplicar técnicas que otimizem a eficácia de agrupamento dos métodos.

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 CONCLUSÕES

A produção científica tem como objetivo melhorar a análise da realidade e produzir transformações por intermédio desse objetivo. A discussão sobre novos métodos a serem adaptados pela Análise Estatística de Formas é um assunto de grande importância para o meio científico. Nesta dissertação, foram enunciados métodos para o agrupamento de formas tridimensionais. Foram apresentados métodos já abordados assim como novos, com o intuito de enriquecer a literatura da área de análise de agrupamento no contexto de formas. O principal objetivo era avaliar as performances dos algoritmos por meio de técnicas de validação de agrupamento em relação à qualidade dos grupos formados.

Para avaliar o desempenho dos métodos propostos, foram realizados dois experimentos com dados simulados baseados em formas tridimensionais, considerandos dois cenários distintos: isotropia e anisotropia. Também foram considerados diferentes graus de dispersão e números de marcos anatômicos. As análises foram interpretadas com base nos valores obtidos pelos índices de Rand Ajustado e Silhueta. Há indícios de que os resultados mostraram que para os cenários de isotropia e anisotropia com baixa dispersão, todos algoritmos, tantos os proposto quanto os já implementados, apresentaram desempenhos bons ou razoáveis. Enquanto que para os casos com alta dispersão, as propostas do presente trabalho, o K-médias ϕ aparado, *CLARANS* e *Hill Climbing* apresentaram bons desempenhos, na maioria das situações, em relação aos métodos já implementados (K-médias e K-médias aparado), levando em consideração pelo menos um dos índices medidos. Porém, era esperado que os desempenhos dos métodos *CLARANS* e *Hill Climbing* tivessem ainda mais destaque, por ser tratarem de métodos que possuem eficácia superior em relação ao K-médias na literária clássica. No entanto isso não ocorreu de modo significativo quando adaptou-se esses métodos ao espaço de formas tridimensionais, ou seja, existem situações em que esses métodos se destacam.

Nas aplicações em dados reais, foram considerados três conjuntos de dados de formas tridimensionais: Crânios de Macacos, Cérebros de Adultos Saudáveis e Ossos do Nariz Humano. Quanto aos resultados obtidos, há indícios de que o desempenho dos métodos apresentaram um panorama bem semelhante com os resultados obtidos para os dados simulados, isto é, o K-médias ϕ aparado, *CLARANS* e *Hill Climbing*, propostas do trabalho, apresentaram performances superiores na maioria dos casos.

É importante destacar ainda que o método K-médias ϕ possui um desempenho superior em relação ao método K-médias, tanto nos conjunto de dados simulados quanto no conjunto de dados reais.

Dessa maneira, a presente dissertação colaborou com a área de análise de agrupamento de formas com a investigação e desenvolvimento de métodos já abordados, assim como a implementação de novos métodos, para o agrupamento de formas tridimensionais.

5.2 TRABALHOS FUTUROS

Como sugestões de continuidade para esta pesquisa, tem-se: adaptar o método *fastCLARANS* proposto por Schubert e Rousseeuw (2019) para o agrupamento de formas 3D, com a finalidade de melhorar o tempo de execução do método *CLARANS*; Utilizar outros critérios de agrupamento para o método *Hill Climbing*, com o intuito de encontrar melhores grupos e adaptar outros algoritmos de busca; Adaptar os métodos de agrupamentos *fuzzy* ao contexto de formas 3D; Fazer uso dos métodos de *clustering ensembles*, como o *Bagging* e o *Boosting*, conjuntamente com os métodos propostos, com o objetivo de melhorar o desempenho dos algoritmos na formação dos grupos.

REFERÊNCIAS

ADAMS, D. C.; OTÁROLA-CASTILLO, E. geomorph: an r package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, Wiley Online Library, v. 4, n. 4, p. 393–399, 2013.

AMARAL, G. J. A.; DORE, L. H.; LESSA, R. P.; STOSIC, B. K-means algorithm in statistical shape analysis. *Communications in Statistics—Simulation and Computation®*, Taylor & Francis, v. 39, n. 5, p. 1016–1026, 2010.

ASSIS, E. C. d. *Algoritmos de particionamento aplicados à análise estatística de formas*. Tese (Doutorado), 2018.

BEZDEK, J. C. *Pattern recognition with fuzzy objective function algorithms*. 1. ed. New York: Plenum Press, 1981.

BOCK, H. H. Clustering methods: a history of k-means algorithms. In: BERLIM. *Selected contributions in data analysis and classification*. [S.l.]: Springer, 2007. p. 161–172.

BOOKSTEIN, F. L. A statistical method for biological shape comparisons. *Journal of Theoretical Biology*, Elsevier, v. 107, n. 3, p. 475–520, 1984.

BOOKSTEIN, F. L. et al. Size and shape spaces for landmark data in two dimensions. *Statistical science*, v. 1, n. 2, p. 181–222, 1986.

BRIGNELL, C. J.; DRYDEN, I. L.; GATTONE, S. A.; PARK, B.; LEASK, S.; BROWNE, W. J.; FLYNN, S. Surface shape analysis with an application to brain surface asymmetry in schizophrenia. *Biostatistics*, v. 11, n. 4, p. 609–630, 2010.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-1, n. 2, p. 224–227, 1979.

DRYDEN, I. L. Shapes package. *R package version 1.2.5*, R Foundation for Statistical Computing, 2012.

DRYDEN, I. L.; KUME, A.; LE, H.; WOOD, A. T. A multi-dimensional scaling approach to shape analysis. *Biometrika*, v. 95, n. 4, p. 779–798, 2008.

DRYDEN, I. L.; LE, H.; PRESTON, S. P.; WOOD, A. T. Mean shapes, projections and intrinsic limiting distributions. *Journal of Statistical Planning and Inference*, v. 145, p. 25–32, 2014.

DRYDEN, I. L.; MARDIA, K. V. Multivariate shape analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, JSTOR, v. 95, n. 3, p. 460–480, 1993.

DRYDEN, I. L.; MARDIA, K. V. *Statistical shape analysis: with applications in R*. 2. ed. [S.l.]: John Wiley & Sons, 2016.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, Taylor & Francis, v. 4, n. 1, p. 95–104, 1974.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. *Cluster Analysis*. 5. ed. [S.l.]: John Wiley, 2011.

EVISON, M.; BRUEGGE, R. V. The magna database: A database of three-dimensional facial images for research in human identification and recognition. *Forensic Science Communications*, v. 10, n. 2, p. 1528–8005, 2008.

FLACH, P. *Machine Learning: the art and science of algorithms that make sense of data*. [S.l.]: Cambridge University Press, 2012.

FRÉCHET, M. Les éléments aléatoires de nature quelconque dans un espace distancié. In: *Annales de l'institut Henri Poincaré*. [S.l.: s.n.], 1948. v. 10, n. 4, p. 215–310.

FREE, S. L.; O'HIGGINS, P.; MAUDGIL, D. D.; DRYDEN, I. L.; LEMIEUX, L.; FISH, D. R.; SHORVON, S. D. Landmark-based morphometrics of the normal adult brain using mri. *NeuroImage*, Elsevier, v. 13, n. 5, p. 801–813, 2001.

FRIEDMAN, H. P.; RUBIN, J. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, v. 62, n. 320, p. 1159–1178, 1967.

GARCÍA-ESCUDERO, L. Á.; GORDALIZA, A. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, v. 94, n. 447, p. 956–969, 1999.

GEORGESCU, V. Clustering of fuzzy shapes by integrating procrustean metrics and full mean shape estimation into k-means algorithm. IFSA-EUSFLAT Conference, p. 1679–1684, 2009.

GOODALL, C. R.; MARDIA, K. V. Multivariate aspects of shape theory. *The Annals of Statistics*, v. 21, n. 3, p. 848–866, 1993.

GOSHTASBY, A. A. *Image registration: Principles, tools and methods*. [S.l.]: Springer, 2012.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: part i. *ACM Sigmod Record*, Association for Computing Machinery, v. 31, n. 2, p. 40–45, 2002.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, JSTOR, v. 28, n. 1, p. 100–108, 1979.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. 2. ed. [S.l.]: Springer, 2009.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, n. 1, p. 193–218, 1985.

JACCARD, P. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, v. 44, p. 223–270, 1908.

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

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. [S.l.]: Prentice-Hall, Inc., 1988.

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

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning: With Applications in R*. [S.l.]: Springer, 2013.

KENDALL, D. G. The diffusion of shape. *Advances in applied probability*, v. 9, n. 3, p. 428–430, 1977.

KENDALL, D. G. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London mathematical society*, v. 16, n. 2, p. 81–121, 1984.

KENDALL, D. G.; BARDEN, D.; CARNE, T. K.; LE, H. *Shape and Shape Theory*. 1. ed. [S.l.]: John Wiley & Sons, 1999.

KENT, J. T. The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 56, n. 2, p. 285–299, 1994.

KENT, J. T.; MARDIA, K. V. Shape, procrustes tangent projections and bilateral symmetry. *Biometrika*, v. 88, n. 2, p. 469–485, 2001.

LE, H.; KENDALL, D. G. The riemannian structure of euclidean shape spaces: a novel environment for statistics. *The Annals of Statistics*, v. 21, n. 3, p. 1225–1271, 1993.

LLOYD, S. Least squares quantization in pcm. *IEEE transactions on information theory*, v. 28, n. 2, p. 129–137, 1982.

LUND, S. plus original by U.; AGOSTINELLI, R. port by C. Circstats: Circular statistics, from "topics in circular statistics" (2001). *R package version 0,2–6*, 2012.

MACIEL, D. B. d. *Um novo método difuso multivariado para análise de agrupamento na presença de variáveis qualitativas*. Dissertação (Mestrado), 2013.

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

MARRIOTT, F. H. C. Optimization methods of cluster analysis. *Biometrika*, v. 69, n. 2, p. 417–421, 1982.

NG, R. T.; HAN, J. Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, IEEE, v. 14, n. 5, p. 1003–1016, 2002.

OLIVEIRA, R. A. d. *Algoritmos para determinação do número de grupos em estudos de formas planas*. Dissertação (Mestrado), 2016.

PRESTON, S. P.; WOOD, A. T. A. Two-sample bootstrap hypothesis tests for three-dimensional labelled landmark data. *Scandinavian journal of statistics*, v. 37, n. 4, p. 568–587, 2010.

QIU, W.; JOE, H. clustergeneration: random cluster generation (with specified degree of separation). *R package version 1.3.7*, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <http://www.R-project.org/>.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, v. 66, n. 336, p. 846–850, 1971.

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.

ROUSSEEUW, P. J.; KAUFMAN, L. *Finding Groups in Data*. [S.l.]: John Wiley & Sons, 1990.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Pearson, 2014.

SCHLAGER, S. *Soft-tissue reconstruction of the human nose: population differences and sexual dimorphism*. Tese (Doutorado) — Universität, 2013.

SCHLAGER, S.; JEFFERIS, G.; IAN, D.; SCHLAGER. Package 'morpho'. *R package version 2.8*, 2020.

SCHUBERT, E.; ROUSSEEUW, P. J. Faster k-medoids clustering: improving the pam, clara, and clarans algorithms. In: SPRINGER. *International Conference on Similarity Search and Applications*. [S.l.], 2019. p. 171–187.

SMALL, C. G. *The Statistical Theory of Shape*. [S.l.]: Springer Science & Business Media, 2012.

TEOTONIO, G. H. F.; AMARAL, G. J. A. *Optimized k-Means Algorithm for Three-dimensional Shapes (Artigo em preparação)*. 2020.

VINUÉ, G. et al. Anthropometry: An r package for analysis of anthropometric data. *R package version 1.14*, Foundation for Open Access Statistics, 2020.

VINUÉ, G.; SIMÓ, A.; ALEMANY, S. The k-means algorithm for 3d shapes with an application to apparel design. *Advances in Data Analysis and Classification*, Springer, v. 10, n. 1, p. 103–132, 2014.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on neural networks*, Ieee, v. 16, n. 3, p. 645–678, 2005.

ZADEH, L. A. Fuzzy sets. In: *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*. [S.l.]: World Scientific, 1996. p. 394–432.

ZHENG, G.; LI, S.; SZEKELY, G. *Statistical Shape and Deformation Analysis: Methods, Implementation and Applications*. 1. ed. [S.l.]: Academic Press, 2017.