# Variantes do Índice Silhueta para Validação de Agrupamentos

**Victória Vargas[1], Eduardo Rodrigues Amorim[2], José André de Moura Brito[1], Gustavo Silva Semaan[3]**

[1]Escola Nacional de Ciências Estatísticas (ENCE-IBGE), Rio de Janeiro, RJ, Brasil.

[2]Universidade Anhanguera (Campus Niterói), Niterói, RJ, Brasil.

[3]Universidade Federal Fluminense (INFES-UFF), Santo Antônio de Pádua, RJ, Brasil.

victoriavargasestudo@gmail.com, jose.m.brito@ibge.gov.br

***Abstract.*** *This paper aims to evaluate four variants of the silhouette index for their ability to detect good quality solutions to clustering problems. Five computational experiments were carried out, covering 51 diversified databases (natural and artificial). As dissimilarity measures, Euclidean and Manhattan distances were used, and for clustering algorithms PAM, DBSCAN, and Bisecting k-means. Besides, computational experiments with the Hopkins statistic were applied to measure the clustering tendency on real datasets where k is unknown. The results obtained indicate that the median-based variant is a good alternative to detect quality solutions.*

***Resumo.*** *O presente artigo traz a proposta de avaliação de quatro variantes do índice de silhueta quanto `a sua capacidade de detectar soluções de boa qualidade para problemas de agrupamento. Neste sentido, foram realizados cinco experimentos computacionais, contemplando 51 instâncias da literatura diversificadas (dados reais e artificiais). Como medidas de dissimilaridade foram utilizadas as distâncias euclidiana e de Manhattan, além de três algoritmos clássicos de agrupamento, a saber: PAM, DBSCAN e Bisecting k-means. De modo adicional, experimentos com a Estatística de Hopkins foram realizados com o intuito de verificar a existência de tendência de agrupamentos nas instâncias reais, em que o número de grupos k não é conhecido a priori. Os resultados obtidos indicam que a variante baseada na mediana constitui-se como boa alternativa para detectar soluções de qualidade.*

## 1.Introdução

A análise de agrupamentos corresponde uma técnica de mineração de dados que abarca uma coleção de algoritmos para a resolução de Problemas de Agrupamento (PA). De acordo com [Han et al. 2012], dado um conjunto $\boldsymbol{X}$ constituído por $\boldsymbol{n}$ objetos, de forma que $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$ e cada objeto $\boldsymbol{x_i}$ possui $\boldsymbol{p}$ atributos, resolver um problema de agrupamento consiste em construir, a partir de $\boldsymbol{X}$, $\boldsymbol{k}$ grupos $\boldsymbol{G_r}$, $\boldsymbol{r} = \boldsymbol{1}, \dots, \boldsymbol{k}$ que definem uma solução $\prod = \{\boldsymbol{G_1}, \dots, \boldsymbol{G_k}\}$, também denominada partição. Como pressuposto de PA,

*Cadernos do IME - Série Informática*
*e-ISSN: 2317-2193 (online)*
*DOI: 10.12957/cadinf.2021.68558*

os objetos alocados a um mesmo grupo devem ter baixo grau de dissimilaridade entre si com base em seus $\boldsymbol{p}$ atributos, sendo tal dissimilaridade definida a partir de uma métrica. Adicionalmente, a estrutura de grupos produzida (alocação dos objetos aos grupos) depende da medida de dissimilaridade e do algoritmo escolhido [Bussab et al. 1990].

Assim, dada uma solução, é de suma importância verificar se ela corresponde a uma boa estrutura de agrupamento [Kaufman e Rousseeuw 1989]. Nesse sentido, neste trabalho foram utilizadas cinco versões do Índice de Validação Silhueta, e realizadas análises das soluções produzidas a partir da aplicação de três algoritmos clássicos da literatura, sendo eles: PAM (*Partitioning Around Medoids*), DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) e *Bisecting k-means*(BK) [Han et al. 2012].

Além da introdução, este trabalho traz, na seção dois, a descrição da metodologia considerada. Na seção três, como principal alvo desse estudo de pesquisa, é apresentado o índice de silhueta clássico e quatro variantes, sendo a última variante uma nova proposta de cálculo desse índice. Na seção quatro são descritas as bases de dados utilizadas nos experimentos computacionais, reportados na seção cinco. Por fim, na seção seis, são apresentadas as conclusões e os trabalhos futuros.

## 2. Metodologia

A metodologia utilizada neste trabalho contempla: (i) algoritmos para construção de soluções; (ii) medidas de distância que permitem quantificar a dissimilaridade entre pares de objetos; (iii) índices para mensurar a qualidade das soluções; (iv) instâncias (bases de dados) da literatura com características diversificadas para a realização dos experimentos.

Para a construção de soluções, este estudo utilizou os algoritmos PAM, DBSCAN e BK. O PAM e o BK possuem, como parâmetro de entrada, a quantidade ($\boldsymbol{k}$) de grupos a ser formada. Já o DBSCAN tem seus parâmetros relacionados ao conceito de densidade (quantidade de objetos em uma dada região). Neste caso, uma abordagem para calibrar os parâmetros de entrada para o DBSCAN, intitulada *DistK*, foi considerada [Semaan 2013]. Em relação à dissimilaridade entre objetos, tendo em vista que todas as bases de dados utilizadas neste trabalho são constituídas por variáveis quantitativas, foram utilizadas as Distâncias Euclidiana (DE) e de Manhattan (DM).

## 3. Índices Silhueta

Proposto por [Kaufman e Rousseeuw 1989], o Índice Silhueta Tradicional (ST) combina coesão e separação. Este índice é calculado para cada um dos $\boldsymbol{n}$ objetos de $\boldsymbol{X}$ e permite determinar a qualidade de uma partição, baseando-se na dissimilaridade de cada objeto em relação aos demais objetos do mesmo grupo (distância *intracluster*) e em relação aos objetos dos demais grupos formados (distância *interclusters*). A partir da silhueta individual de cada objeto, obtém-se a silhueta da solução (FC), uma medida que permite avaliar uma solução de agrupamento de maneira global. Nas Equações de 1 a 4 é apresentado o cálculo do índice para um dado objeto($x_i$), enquanto a silhueta da solução é calculada por meio da Equação 5.

$$a(x_i) = \frac{1}{|G_w|}\sum d(x_i, x_j)\ \forall x_i \neq x_j, \quad x_i \in G_w,\ x_j \in G_w \tag{1}$$

$$d_{ext} = \frac{1}{|G_t|}\sum d(x_i, x_j)\ \ x_i \notin G_t,\ \ \forall x_j \in G_t \tag{2}$$

$$b(x_i) = \min d_{ext}(x_i, G_t),\ \ G_t \neq G_w,\ \ x_i \in G_w \tag{3}$$

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}}\ i = 1, \ldots, n \tag{4}$$

$$FC = \frac{1}{n}\sum_{i=1}^{n} s(x_i) \tag{5}$$

Na equação (1), $\boldsymbol{a(x_i)}$ corresponde à distância média de cada objeto $\boldsymbol{x_i}$ em relação aos demais objetos do mesmo grupo. O termo $\boldsymbol{b(x_i)}$ na equação (3) corresponde à menor distância média de $\boldsymbol{x_i}$ aos demais grupos, obtida a partir da equação (2). Utilizando-se os valores de $\boldsymbol{a(x_i)}$ e $\boldsymbol{b(x_i)}$, calcula-se o valor da silhueta de cada objeto pela equação (4) e a silhueta da solução pela equação (5). Segundo [Kaufman e Rousseeuw 1989], a silhueta é uma medida adimensional, útil para avaliar o quanto a solução produzida pelo algoritmo utilizada na construção dos grupos representa, ou seja, a qualidade da estrutura de agrupamento encontrada. Na Tabela 1, nota-se que quanto mais próximo de 1 é o valor da silhueta, maiores são os indícios de que os dados possuem uma forte estrutura.

**Tabela 1. Intervalos de classificação de *FC* segundo [Kaufman e Rousseeuw 1989].**

| *FC* | Descrição |
|---|---|
| 0,71 – 1,00 | Estrutura forte encontrada nos dados. |
| 0,51 – 0,70 | Estrutura razoável encontrada nos dados. |
| 0,26 – 0,50 | Estrutura fraca, possivelmente artificial; avaliar a aplicação de outros algoritmos nos dados. |
| ≤0,25 | Não foi encontrada estrutura substancial nos dados. |

Também conhecida na Literatura, a Silhueta Simplificada (SS) apresenta resultados de qualidade comparável aos da ST, com a vantagem de demandar menor custo computacional. Para essa variante, $\boldsymbol{a(x_i)}$ é a distância entre $\boldsymbol{x_i}$ e o centroide do mesmo grupo $\boldsymbol{c_r}$, $\boldsymbol{(x_i \in G_r)}$, enquanto $b(x_i)$ é a menor distância entre $\boldsymbol{x_i}$ e o centroide $\boldsymbol{c_s}$ de um outro grupo $\boldsymbol{G_s}$ $\boldsymbol{(r \neq s)}$[Hruschka et al. 2004] (vide equações (6) e (7)).

$$a(x_i) = \ d(x_i, c_r)\ x_i \in G_r \tag{6}$$

$$b(x_i) = \min\ d(x_i, c_s),\ \ x_i \notin G_s \tag{7}$$

Propostas por [Amorim 2013], as Silhueta Alternativas 1 e 2 (SA1 e SA2) proporcionaram soluções equivalentes ou superiores às observadas na literatura, no que diz respeito aos experimentos conduzidos com o DBSCAN em um subconjunto de instâncias de DS2 – descritas na seção quatro. Ambas consideram $\boldsymbol{a(x_i)}$ como a menor distância entre $\boldsymbol{x_i}$ e o objeto mais próximo do mesmo grupo, enquanto $\boldsymbol{b(x_i)}$ corresponde à menor distância entre $\boldsymbol{x_i}$ e um objeto pertencente a outro grupo, como pode ser observado nas equações (8) e (9).

$$a(x_i) = \min d(x_i, c_r) \quad \forall x_i \neq x_j, \; x_i \in G_w, \; x_j \in G_w \tag{8}$$

$$b(x_i) = \min \; d(x_i, x_j) \quad \forall x_i \neq x_j, \; x_i \in G_w, \; x_j \in G_t \tag{9}$$

A SA2 difere da SA1 por utilizar, em seu cálculo, uma função indicadora, $\boldsymbol{c(x_i)}$, que permite avaliar se $\boldsymbol{x_i}$ está mais próximo de um objeto de seu próprio grupo do que de um objeto de outro grupo, ou seja, se está alocado ao "grupo correto". Quando $\boldsymbol{a(x_i) < b(x_i)}$, $\boldsymbol{c(x_i) = 1}$ e, caso contrário, $\boldsymbol{c(x_i) = 0}$. Em grupos com um único objeto, $\boldsymbol{c(x_i) = 0}$. Portanto, em SA1 a silhueta de um objeto é dada pela equação (1), e para a SA2 corresponde à função $\boldsymbol{c(x_i)}$ (Equação 10). Dessa forma, SA2 assume valores inferiores a 1 quanto ao menos um objeto esteja alocado "incorretamente".

$$c(x_i) = \begin{cases} 1, & se \; a(x_i) < b(x_i) \\ 0, & c.c \end{cases} \tag{10}$$

Por fim, a variante denominada Silhueta Mediana (SM), proposta neste trabalho, é baseada na ST, mas difere por considerar $\boldsymbol{a(x_i)}$ como a distância mediana entre o objeto $\boldsymbol{x_i}$ e os demais objetos do mesmo grupo, enquanto $\boldsymbol{b(x_i)}$ corresponde à menor distância mediana em relação a todos os objetos de um mesmo grupo para cada um dos demais grupos. A utilização da mediana em vez da média tem por objetivo tornar o índice menos suscetível a valores extremos.

## 4. Bases de Dados

Para a realização dos experimentos computacionais relatados na seção cinco, foram utilizadas 51 instâncias somente com atributos quantitativos, todas disponíveis e relatadas em trabalhos da literatura: conjunto DS2[1] , UCI[2] , Atlas Brasil [3], SIDRA[4] e DATASUS[5], descritas nas tabelas 2, 3, 4, 5 e 6, respectivamente. Nessas tabelas $\boldsymbol{p}$, $\boldsymbol{n}$ e $\boldsymbol{k}$ correspondem a quantidade de atributos, quantidade de objetos e número ideal de grupos (conhecido a priori), respectivamente.

[1] Conjunto de Dados DS2 utilizado em diversos trabalhos e reportados em [Semaan 2013].
[2] *University of California*, Irvine - *Machine Learning Repository* (https://archive.ics.uci.edu/ml).
[3] Atlas do Desenvolvimento Humano no Brasil (https://dados.gov.br/dataset/atlasbrasil).
[4] Sistema IBGE de Recuperação Automática (https://sidra.ibge.gov.br/acervo).
[5] Ministério da Saúde - OpenDataSUS (https://opendatasus.saude.gov.br).

Para a realização dos experimentos 1, 2 e 3 foram consideradas as 31 instâncias da Tabela 2, todas com 2 atributos, obtidas do conjunto de bases DS2 [Semaan 2013]. Essas bases são classificadas em "comportadas" (C) ou "não comportadas" (NC). Mais especificamente, as bases ditas "comportadas" possuem grupos bem delimitados, coesos e separados, enquanto as "não comportadas" apresentam padrões mais difusos quanto à separação e distribuição de seus objetos, com grupos menos definidos [Semaan 2013]. É importante salientar que, para essas instâncias, o número ideal de grupos é conhecido a priori, e suas nomenclaturas traz informações importantes. Por exemplo, DS2-100p2c1 indica que a base possui 100 objetos, 2 grupos e que é "não comportada", enquanto DS2-400p3c possui 400 objetos, 3 grupos e é "comportada".

**Tabela 2. Bases do conjunto DS2.**

| Base | n | k | Padrão | Base | n | k | Padrão |
|---|---|---|---|---|---|---|---|
| DS2-1000p5c1.csv | 1000 | 5 | NC | DS2-200p8c1.csv | 200 | 8 | NC |
| DS2-1000p6c.csv | 1000 | 6 | C | DS2-300p2c1.csv | 300 | 2 | NC |
| DS2-100p2c1.csv | 100 | 2 | NC | DS2-300p3c.csv | 300 | 3 | C |
| DS2-100p3c.csv | 100 | 3 | C | DS2-300p3c1.csv | 300 | 4 | NC |
| DS2-100p3c1.csv | 100 | 3 | NC | DS2-300p4c1.csv | 300 | 4 | NC |
| DS2-100p5c1.csv | 100 | 5 | NC | DS2-300p6c1.csv | 300 | 6 | NC |
| DS2-100p7c1.csv | 100 | 7 | NC | DS2-400p3c.csv | 400 | 3 | C |
| DS2-100p8c1.csv | 100 | 8 | NC | DS2-400p4c1.csv | 400 | 4 | NC |
| DS2-1100p6c1.csv | 1100 | 6 | NC | DS2-500p3c.csv | 500 | 3 | C |
| DS2-1500p6c1.csv | 1500 | 6 | NC | DS2-500p4c1.csv | 500 | 4 | NC |
| DS2-2000p9c1.csv | 2000 | 9 | NC | DS2-500p6c1.csv | 500 | 6 | NC |
| DS2-200p2c1.csv | 200 | 2 | NC | DS2-600p3c1.csv | 600 | 3 | NC |
| DS2-200p3c1.csv | 200 | 3 | NC | DS2-700p4c.csv | 700 | 4 | C |
| DS2-200p4c.csv | 200 | 4 | C | DS2-800p4c1.csv | 800 | 4 | NC |
| DS2-200p4c1.csv | 200 | 4 | NC | DS2-900p5c.csv | 900 | 5 | C |
| DS2-200p7c1.csv | 200 | 7 | NC | | | | |

**Tabela 3. Bases UCI.**

| Base | p | n | k |
|---|---|---|---|
| avila_ts | 11 | 10437 | 12 |
| breast_cancer | 10 | 116 | 2 |
| breast_tissue | 11 | 106 | 6 |
| cardiotocography | 40 | 2126 | 10 |
| ecoli | 9 | 336 | 8 |
| machine_failure | 14 | 10000 | 2 |
| seeds | 8 | 199 | 3 |
| vertebral column | 7 | 310 | 3 |

No Experimento 4 foram utilizadas oito instâncias, obtidas a partir do *Machine Learning Repository da University of California, Irvine (UCI)*. Elas são comumente utilizadas em problemas de classificação, e o número de grupo considerado como o ideal corresponde à sua quantidade de classes. Suas características estão na Tabela 3. Na Tabela 4 estão as bases retiradas do Atlas Brasil, que têm atributos relativos a indicadores de educação, renda, longevidade, entre outros, para municípios de cinco estados brasileiros. A Tabela 5 contém as bases retiradas do DATASUS, que têm atributos relativos à mortalidade e doenças, também para municípios de cinco estados brasileiros. Por fim, na Tabela 6 estão as bases do SIDRA, relativas ao Censo Agropecuário 2017 e PIB 2018 para municípios das regiões Norte e Nordeste do Brasil. Destaca-se que para as bases relatadas nas Tabelas 4, 5 e 6 a quantidade de grupos não é conhecida. Além disso, foram realizados experimentos com Estatística de Hopkins (descrita na Seção 5) para confirmar se, nessas bases, existe tendência à formação de agrupamentos [Semaan et al. 2019].

**Tabela 4. Bases Atlas Brasil (ano 2010).**

| Base | p | n | Descrição |
|---|---|---|---|
| AM | 2 | 61 | Esperança de vida ao nascer e mortalidade infantil. |
| GO | 2 | 245 | Taxa de atividade (maiores de 10 anos) e de desocupação (maiores de 10 anos). |
| MA | 2 | 216 | Índice de Gini e Renda per capita. |
| RS | 3 | 495 | Taxa de fecundidade total, Razão de dependência e Taxa de envelhecimento. |
| SP | 3 | 644 | IDHM (Índice de Desenv. Humano Municipal) Renda, Longevidade e Educação. |

**Tabela 5. Bases DATASUS (ano 2015).**

| Base | p | n | Descrição |
|---|---|---|---|
| BA_AT | 1 | 417 | Taxa de mortalidade por causas externas (acidentes de trânsito) (100 mil hab.). |
| MS_H | 1 | 77 | Taxa de mortalidade por causas externas (homicídio) (100 mil hab.). |
| MT_DNT | 1 | 140 | Taxa de mortalidade por doenças crônicas não transmissíveis (100 mil hab.). |
| PA_AG | 1 | 142 | Taxa de mortalidade por agressão (100 mil habitantes). |
| RJ_AIDS | 1 | 92 | Taxa de incidência de AIDS (100 mil habitantes). |

**Tabela 6. Bases SIDRA.**

| Base | p | n | Descrição |
|---|---|---|---|
| CENSOAGRO2017_Nordeste | 1 | 1338 | Estabelecimentos com produtos da extração veg. (2017). |
| PIB_Norte | 1 | 450 | PIB a preços correntes (Mil reais) (2018). |

## 5. Experimentos Computacionais

Para a realização dos cinco experimentos foram utilizadas 51 instâncias, com características diversificadas em relação: (i) às dimensões - em objetos ($\boldsymbol{n}$), atributos ($\boldsymbol{p}$) e grupos ($\boldsymbol{k}$); (ii) estrutura dos dados - grupos bem definidos, coesos e bem separados ou

com padrões difusos; (iii) origem dos dados - reais (naturais) ou gerados artificialmente; (iv) problema: para o PA (***k*** é conhecido), Problema de Classificação (classes conhecidas) ou instâncias sem grupos e classes conhecidos. As instâncias utilizadas possuem entre 61 e 10437 registros (objetos), entre 1 e 40 atributos (todos quantitativos), e de 2 a 12 grupos (quando reportado), conforme apresenta a Seção 4. Adicionalmente, todas as bases com mais de um atributo foram padronizadas antes de sua utilização, de forma que as variáveis ficassem com a mesma ordem de grandeza. Os algoritmos usados, PAM, DBSCAN e BK, estão disponíveis no CRAN[6], nos pacotes cluster (função *pam*), dbscan e stats (função *kmeans*). Foi utilizado um computador dotado de processador AMD Ryzen 5 de 2.1GHz, com 12 GB de RAM e Windows 10.

A Tabela 7 sumariza os resultados de quatro experimentos, contemplando: os algoritmos, as distâncias, as instâncias e o percentual de acertos de cada variante do índice silhueta. Para o cálculo do percentual de acertos, foi observado em qual ***k*** (número de grupos) o maior valor de silhueta ocorreu. Caso este ***k*** esteja a, no máximo, uma unidade do ***k*** considerado como ideal para a respectiva instância, para mais ou para menos, contabiliza-se um acerto.

**Tabela 7. Percentual de acertos das silhuetas por tipo de algoritmo, silhueta e distância.**

| Exp. | Algoritmo | Dist. | Instâncias | ST | SA1 | SA2 | SS | SM |
|---|---|---|---|---|---|---|---|---|
| 1 | PAM | DE | DS2 | 93,5% | 45,2% | 45,2% | 96,8% | **100,0%** |
| 1 | PAM | DM | DS2 | 93,5% | 45,2% | 45,2% | 96,8% | **100,0%** |
| 2 | DBSCAN | DE | DS2 | 80,6% | 58,1% | 61,3% | 77,4% | **83,9%** |
| 2 | DBSCAN | DM | DS2 | **61,3%** | 48,4% | 45,2% | **61,3%** | **61,3%** |
| 3 | BK | DE | DS2 | 80,6% | 35,5% | 35,5% | 80,6% | **83,9%** |
| 3 | BK | DM | DS2 | 77,4% | 35,5% | 35,5% | 77,4% | **80,6%** |
| 4 | PAM | DE | UCI | 37,5% | 50,0% | **62,5%** | 37,5% | 37,5% |
| 4 | PAM | DM | UCI | 50,0% | 50,0% | **62,5%** | 50,0% | 50,0% |

No experimento 1, o algoritmo PAM foi aplicado nas bases do conjunto DS2, sendo utilizadas as distâncias Euclidiana e de Manhattan. Na primeira etapa do experimento, definiu-se a variação do parâmetro ***k*** da função pam entre 2 e 9, sendo tais valores correspondentes, respectivamente, ao número mínimo e máximo de grupos observados no conjunto de bases de dados. Em uma etapa posterior, considerando as 248 soluções (bases $x$ valores de $k$) produzidas pelo algoritmo PAM, foram aplicados o índice de silhueta tradicional e suas 4 variantes. Em seguida, identificou-se, em cada base, para qual valor de ***k*** foi observado o maior valor de silhueta, considerando cada tipo de silhueta e se este correspondia ao número de grupos definido previamente na Tabela 2. Na Tabela 7 pode ser observado que a distância utilizada não foi determinante para a qualidade das soluções, visto que os mesmos percentuais de acertos foram obtidos para ambas as distâncias. A ST foi superada em acertos pela SM e SS. Das silhuetas analisadas, SM foi a que apresentou o melhor desempenho, permitindo identificar corretamente o número de grupos ideal em todas as bases de dados. As silhuetas com pior desempenho foram SA1 e SA2, ambas com percentual de acertos igual a 45,2%.

[6] The Comprehensive R Archive Network (https://cran.r-project.org).

No experimento 2 aplicou-se o algoritmo DBSCAN em conjunto às distâncias euclidiana e de Manhattan nas bases do conjunto DS2. Preliminarmente à aplicação do algoritmo, foi necessário realizar uma calibração de parâmetros de entrada e, para esse fim, escolheu-se a abordagem *DistK* [Semaan 2013]. Destaca-se que todos os objetos devem fazer parte das soluções finais, ou seja, objetos inicialmente classificados como ruídos são alocados ao grupo mais próximo. Considerando a Tabela 7, observa-se que a distância euclidiana proporcionou maior percentual de acertos que a distância de Manhattan. As variantes com melhor desempenho para a distância euclidiana são SM e ST. Já para a distância de Manhattan, houve um empate em relação ao percentual de acertos das silhuetas Mediana, Simplificada e Tradicional. SA1 e SA2 apresentaram desempenho inferior às demais silhuetas, independentemente do tipo de distância. Entretanto, observa-se que seus percentuais de acertos foram mais elevados com a utilização da distância euclidiana. Ainda para SA1 e SA2, nota-se que seus percentuais de acertos são similares e inferiores a 50% para a distância de Manhattan, assim como o observado no experimento 1.

No Experimento 3, foi utilizado o algoritmo BK nas instâncias do conjunto DS2. Optou-se por utilizar somente a DE como métrica, e o (BK) utilizou como critério de seleção do grupo a ser particionado o menor valor de silhueta para cada uma das versões da silhueta abordadas nesse trabalho. Definiu-se que o algoritmo iria produzir soluções para $\boldsymbol{k}$ variando de 2 a 5, dado que das 31 bases de DS2, 21 possuem $\boldsymbol{k}$ nesse intervalo. Para cada $\boldsymbol{k}$ a partir de $\boldsymbol{k} = 2$, a escolha do grupo a ser particionado se dá pela aplicação de cada um dos 5 tipos de silhueta em cada um dos grupos obtidos até o momento (no processo de divisão). O grupo que apresenta o menor valor de silhueta é dividido em dois novos grupos e, assim, o algoritmo segue até que $\boldsymbol{k} = 5$ seja obtido. Ao final de cada iteração (divisão), as silhuetas são calculadas com as soluções parciais obtidas, ou seja, com as soluções em que o total de grupos definido pelo usuário ainda não foi atingido. Tendo por base a Tabela 7, SM, SS e ST obtiveram desempenho superior às demais silhuetas, quando utilizadas como critério de divisão dos grupos. Em especial, nota-se que, para esses critérios, a utilização da SM na avaliação das soluções parciais resultou em acerto no número de grupos em mais de 80% das bases. A SM foi a variante com maior percentual de acertos em todos os critérios de divisão utilizados.

No Experimento 4, o algoritmo PAM foi aplicado nas 8 instâncias de classificação do repositório do UCI, vide Tabela 3, ($\boldsymbol{n}$ entre 106 e 10437, $\boldsymbol{p}$ entre 7 e 40 e $\boldsymbol{k}$ (classes) entre 2 e 12), com o parâmetro $\boldsymbol{k} \in \{2, 3, \dots, 12\}$. Foi utilizada a Estatística de Hopkins (EH) para verificar a existência de tendência de agrupamento nessas instâncias, embora as classes sejam conhecidas. Em linhas gerais, a EH consiste em um teste de hipóteses, em que a hipótese nula é de que não existe estrutura de grupos no conjunto de dados, enquanto a hipótese alternativa é de que o conjunto de dados possui estrutura de grupos. Para essa estatística, quanto mais próximo de 1 for o seu valor, maior a indicação de que os dados possuem uma tendência de agrupamento. Como mostra a Figura 1, para essas oito instâncias, todos os resultados de EH foram superiores a 0,7, o que indica a existência de uma boa estrutura dos dados (tendência a agrupamento) [Semaan 2013]. Ao avaliar o percentual de acertos por tipo de distância, nota-se um percentual maior para a distância de Manhattan. De acordo com a Tabela 7, a silhueta com melhor desempenho foi a SA2 que, em ambas as distâncias, possibilitou identificar o número ideal de grupos em 62,5% dos casos. Diferentemente dos experimentos anteriores, SA2 e SA1 produziram percentual de acertos comparável ou superior às demais variantes de silhueta.

Em geral, a partir da Tabela 7, observa-se que nos Experimentos 1, 2 e 3, houve melhor performance (acertos) da SM frente às demais alternativas e no Experimento 4, ocorre uma inversão, principalmente considerando a SA2 e a DE; fato que pode ser decorrente do grau de assimetria dos dados, o que tende a produzir valores $\boldsymbol{b(x_i)}$ bem superiores aos de $\boldsymbol{a(x_i)}$, implicando alta prevalência de $\boldsymbol{c(x_i) = 1}$.

Por fim, o Experimento 5 contemplou 12 instâncias ($\boldsymbol{n}$ entre 61 e 1338, $\boldsymbol{p}$ entre 1 e 3) retiradas dos repositórios Atlas Brasil, DATASUS e SIDRA, descritas nas tabelas 4, 5 e 6, cujo número ideal de grupos não é conhecido. O algoritmo PAM foi aplicado nessas instâncias, com o parâmetro $k \in \{2, \ldots, \frac{\sqrt{n}}{2}\}$, onde $\boldsymbol{n}$ é o número de objetos da respectiva instância. Novamente a EH foi utilizada, sendo confirmada a tendência à formação de agrupamentos para todas as instâncias, como mostra a Figura 2.

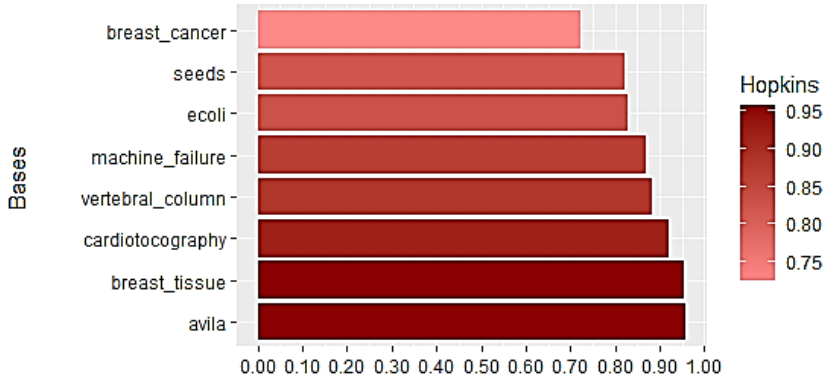


**Figura 1. Estatística de Hopkins para as bases do experimento 4.**

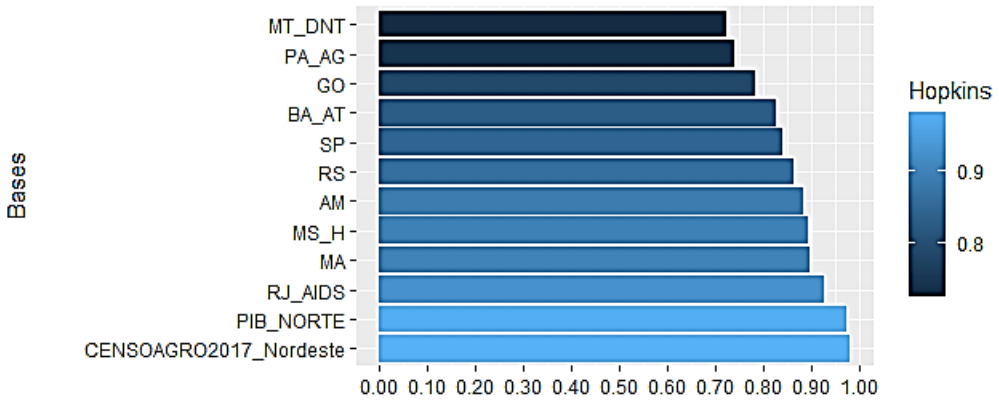


**Figura 2. Estatística de Hopkins para as bases do experimento 5.**

Nesse último experimento, após obter as soluções via algoritmo PAM, a coincidência dos valores ideais de $\boldsymbol{k}$ entre os pares de silhuetas foi calculada, bem como os $\boldsymbol{k}$ associados aos maiores valores de cada versão, conforme apresenta a Tabela 8. Destaca-se o par SA1 e SA2, que obteve 100% de coincidência para a distância de Manhattan, bem como o par SS e SM, com coincidência superior a 80% em ambas as distâncias.

**Tabela 8. Coincidência do valor de *k* por tipo de distância.**

| Distância | ST e SA1 | ST e SM | SA1 e SA2 | SA1 e SM | SS e SM |
|---|---|---|---|---|---|
| Euclidiana | **83,3%** | **83,3%** | 75,0% | **83,3%** | **83,3%** |
| Manhattan | 75,0% | 75,0% | **100,0%** | 66,7% | **91,7%** |

## 6. Conclusões e Trabalhos Futuros

Neste trabalho foram estudadas e avaliadas quatro variantes (três da literatura) para o índice de validação silhueta, medida utilizada para verificar a qualidade de um agrupamento e que combina coesão e separação. A SM, baseada na distância mediana, foi aplicada com o intuito de tornar o índice menos suscetível a valores extremos. Para avaliar o desempenho dessas variantes foram realizados cinco experimentos que contemplaram os algoritmos PAM, DBSCAN e BK, 51 bases de dados da literatura e dois tipos de distâncias. Nos experimentos 1 e 2, a combinação da SM propiciou um percentual de acertos superior a 80%. No experimento 3 o destaque ocorreu com o uso da SM tanto como critério de seleção quanto na avaliação das soluções. Já nos experimentos 4 e 5, a DM ocasionou maiores percentuais de coincidência entre os pares de silhuetas que a DE, em geral. Em especial, no experimento 5 a correspondência do valor de *k* ideal foi de 100% entre as SA1 e SA2.

Com base nos resultados obtidos, a SM pode constituir-se como uma boa opção, quando comparada à ST, nos casos em que as bases de dados possuem grupos bem estruturados. Como trabalhos futuros pretende-se analisar com mais profundidade as vantagens e desvantagens de cada uma das versões do índice. Além disso, deve-se utilizar outros algoritmos, como o *Clustering Large Applications* (CLARA) e algoritmos hierárquicos como *Single Linkage* e *Complete Linkage.*

## Agradecimentos

Os autores agradecem ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela bolsa concedida a autora principal, à FAPERJ (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro) e à PROPPI/UFF (Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação da Universidade Federal Fluminense, edital FOPESQ 2021) pelo financiamento parcial da pesquisa realizada.

## Referências

Amorim, E. R. (2013). Novos Índices Relativos para a Identificação da Quantidade Ideal de Grupos. Trabalho de conclusão de curso, Universidade Anhanguera, Niterói - RJ.

Bussab, W. O., Miazaki, E. S., Andrade, D. F. (1990). Introdução à Análise de Agrupamentos. IME - USP, São Paulo.

Han, J., Kamber, M., Pei, J. (2012). Data Mining: Concepts and Techniques: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Hruschka, E. R., Campello, R. J. G. B., Castro, L. N. (2004). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In IEEE International Conference on Data Mining, pages 403–406.

Kaufman, L., Rousseeuw, P. J. (1989). Finding Groups in Data - An Introduction to

Clusters Analysis. Wiley-Interscience Publication.

Semaan, G. S. (2013). Algoritmos para o Problema de Agrupamento Automático. Tese de doutorado, Universidade Federal Fluminense, Niterói - RJ.

Semaan, G.S., Fadel, A.C., Brito, J.A.M., Ochi, L.S. (2019). *A Hybrid Efficient Heuristic with Hopkins Statistic for the Automatic Clustering Problem*. IEEE Latin America Transactions, v.19, n.1.