Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Índice de Densidade da Clusterização: Uma Nova Métrica para Validação Interna de Agrupamentos

Dirceu Scaldelai[1], Solange R. dos Santos[2]
Unespar, Campo Mourão, PR
Luiz C. Matioli[3]
DAMAT/UFPR, Curitiba, PR

Resumo. Neste trabalho propomos uma nova métrica de validação interna de clusterização, o índice de Densidade da Clusterização (índice CD), baseado na máxima razão entre a dispersão interna dos clusters e a separação entre centroides. Visando facilitar a compreensão da nova métrica de validação, a qual foi implementada no Software R, descrevemos detalhadamente sua metodologia e procedimentos, exemplificando cada um dos seus passos por meio de um problema simples, bidimensional, com um número reduzido de observações e uma estrutura bem definida. Na sequência, realizamos experimentos numéricos comparando o índice CD com outras duas métricas de validação já consagradas na literatura, o índice DB e o coeficiente de silhueta. Resultados preliminares revelaram que o índice CD é eficiente para avaliar clusterização de dados multidimensionais, uma vez que apresentou uma concordância substancial com o índice DB, a um custo de execução similar, e uma concordância significativa com o coeficiente de silhueta, a um custo execução consideravelmente menor. Sendo assim, os resultados evidenciam a boa qualidade do índice CD como métrica de validação interna para clusterização de dados multidimensionais.

Palavras-chave. Métrica de validação interna, Clusterização, Comparação.

1 Introdução

O processo de explorar grandes quantidades de dados à procura de padrões, regras ou sequências de informações, para detectar correlações entre as variáveis, faz parte do campo da ciência denominada Ciências de dados, do inglês *Data science*. De acordo com [7, 10], a Ciência de dados se concentra na implementação de decisões fundamentadas em dados e no gerenciamento de suas consequências sendo uma prática interdisciplinar com forte conexão com Estatística, Aprendizado de máquina e *Big data*.

Para [5, 7] a clusterização, do inglês *clustering*, é um dos campos de estudos da Ciências de dados que busca identificar padrões ou grupos[4] de objetos semelhantes em um conjunto de dados de interesse, sendo utilizada nos mais variados campos da pesquisa científica, tendo seus resultados frequentemente utilizados para prever, analisar e replicar fenômenos, além de fomentar novas pesquisas.

De forma simples, a clusterização de um conjunto de dados consiste na identificação de grupos em que os elementos pertencentes a um mesmo grupo devem possuir características semelhantes

[1]dirceu.scaldelai@ies.unespar.edu.br
[2]solaregina@gmail.com
[3]lcmatioli@gmail.com
[4]No decorrer do texto também é adotada a expressão cluster para fazer referência a cada grupo proveniente do agrupamento de dados, da mesma forma que clusterização e agrupamento são considerados sinônimos.

enquanto elementos de grupos distintos devem exibir características diferentes. Logo, a clusterização consiste num processo de partição do conjunto de dados em subconjuntos disjuntos de forma a minimizar a dissimilaridade intra-grupos e maximizar a intergrupos.

Segundo [8], determinar a qualidade da clusterização envolve avaliar a qualidade dos grupos por meio de critérios externos, internos ou relativos. Os critérios internos refletem a compactação de elementos e a separação entre diferentes grupos. Assim, considerando que o objetivo da clusterização é agrupar elementos semelhantes no mesmo grupo e elementos diferentes em grupos distintos, as métricas de validação interna geralmente são baseadas em dois critérios: a coesão e a separação. Nesse contexto, propomos uma nova métrica de validação interna para clusterização multidimensional, baseada na máxima razão entre a dissimilaridade intra-grupos e a intergrupos, denominada índice de Densidade da Clusterização ou, simplesmente, "índice CD".

2 Índice de Densidade da Clusterização

No aspecto geral, o índice CD é definido como a máxima razão entre a dispersão interna de cada grupo (coesão) e a separação do respectivo centroide do grupo ao centroide mais próximo (separação). Supondo que uma clusterização tenha k grupos, o índice CD calcula k razões, e dentre essas, a maior é definida como sendo a métrica de avaliação da clusterização, ou seja, valor do índice CD. De forma mais detalhada, supondo que um conjunto $X \in \mathbb{R}^{m\times n}$ tenha um agrupamento com k grupos provenientes de algum algoritmo de clusterização, a coesão do $j-$ésimo grupo é calculada pela média das distâncias de todas as observações pertencentes a esse grupo à seu respectivo centroide, ou seja,

$$S_j = \left\{ \frac{1}{T_j} \sum_{i=1}^{T_j} \| |\overline{X}_i - \mu_j \| \right\}, \quad j = 1, \ldots, k \tag{1}$$

em que T_j é a quantidade de elementos pertencentes ao grupo j, $\mu_j \in \mathbb{R}^n$ é o centroide do grupo j e $\overline{X} \subset X$ são todos os elementos que pertencem ao grupo j. Por outro lado, a separação entre os grupos é medida da seguinte forma

$$D_j = \{min\{\|\mu_j - \mu_i\|; i = 1, \ldots, k; i \neq j\}\}, \quad j = 1, \ldots, k \tag{2}$$

em que μ_j é o centroide do grupo analisado e μ_i é qualquer outro centroide diferente de μ_j. Na separação entre grupos, cada grupo é analisado de forma isolada, ou seja, fixado o grupo j, calculamos a distância do centroide μ_j a todos os outros $k-1$ centroides. A mínima distância encontrada é considerada como medida de separação do grupo j.

Com base nos valores S_j e D_j definimos a razão sobre cada um dos grupos da seguinte maneira

$$CD_j = \frac{S_j}{D_j}, \quad j = 1, \ldots, k.$$

O índice CD corresponde ao valor máximo obtido, ou seja,

$$\text{índice CD} = max\{CD_j; j = 1, \ldots, k\}.$$

Esse resultado exprime quantitativamente a qualidade do agrupamento de um conjunto de dados X em k grupos, onde quanto mais próximo de zero for o valor do índice, melhor é a qualidade do agrupamento.

A justificativa para que o índice CD ideal seja um valor próximo de zero é simples e intuitiva. Como o índice CD consiste em uma razão, logo devemos avaliar o comportamento do seu numerador

e do seu denominador. O numerador mede a coesão do grupo, dessa forma, quanto mais compacto o grupo j, menor o valor de S_j. Por outro lado, o denominador avalia a separação entre grupos e, é evidente que, para um agrupamento ser considerando bom, almejamos que os centroides dos grupos estejam o mais distante possível um do outro, logo D_j deverá ser o maior possível. Nestes termos, um agrupamento é considerado adequado quando S_j for o menor possível e D_j o maior possível, o que conduz $\frac{S_j}{D_j}$ a um número pequeno. Note que, quando o grupo consistir de apenas um elemento, este será o próprio centroide e S_j será nulo. Por outro lado, em virtude da estrutura dos problemas de clusterização, D_j poderá assumir um grande valor, no entanto nunca assumirá realmente um valor infinito.

Em contrapartida, se o agrupamento for formado por grupos dispersos e os centroides forem próximos, a ordem de grandeza do numerador e do denominador da razão se invertem, fazendo que $\frac{S_j}{D_j}$ seja um número grande, caracterizando um agrupamento insatisfatório. No entanto, para que um agrupamento seja considerado inadequado, não há a necessidade de que o índice CD seja um número muito grande, basta que seja superior a 1.

Supondo que o valor da métrica seja superior a 1 e considerando a natureza da razão $\frac{S_j}{D_j}$, então $S_j > D_j$, isto é, a distância média dos elementos do grupo ao seu centroide é superior a distância deste ao centroide vizinho mais próximo, ou seja, o centroide mais próximo apresenta uma distância menor que alguns elementos do próprio grupo, o que caracteriza o agrupamento como insatisfatório. Dessa forma, o valor do índice CD superior a 1 caracteriza um agrupamento inconsistente, com centroides deslocados ou com sobreposição de centroides num mesmo conjunto de observações tornando o agrupamento indesejável.

De acordo com a estrutura do índice CD, k razões são calculadas e a máxima é definida como métrica final de validação do agrupamento. A escolha por associar o valor do índice CD com a máxima razão se deve ao fato que, se o grupo que propiciou a pior razão é um grupo considerado bom, compacto e distante de outro centroide, então todos os demais $k-1$ grupos terão características semelhantes ou melhores, justificando a escolha pela máxima razão. Porém, a recíproca dessa afirmação não é verdadeira, visto que um valor do índice CD alto não implica que todos os grupos sejam considerados ruins e sim que existe ao menos um mensurado como insatisfatório.

Para ilustrar o funcionamento do índice CD, escolhemos um exemplo simples com 45 observações de 2 atributos e com uma clusterização constituída por 3 grupos. A clusterização juntamente com os centroides são representados no quadro (a) da Figura 1.

Primeiramente calculamos uma razão para cada um dos grupos e selecionamos como métrica de validação o seu valor máximo, ou seja, $CD = max\{CD_j\}$, com $CD_j = \frac{S_j}{D_j}$, $j = 1, \ldots, 3$.

O quadro (b) da Figura 1 ilustra a determinação de CD_1. Para tanto, inicialmente determinamos as distâncias Euclidianas, representadas por segmentos pontilhados na cor vermelha, de cada elemento do grupo 1 ao seu respectivo centroide. Em seguida, calculamos o valor médio dessas distâncias, $S_1 = 0,142$, que representa a coesão do grupo 1. Já a medida de separação do grupo 1 em relação aos demais grupos é dada pelo valor mínimo entre a distância Euclidiana do centroide 1 aos centroides 2 e 3, representadas no quadro (b) da Figura 1 pelos segmentos na cor preta. Conforme é possível observar, a distância entre o centroide 1 aos centroides 2 e 3 é de 1,427 e 1,091, respectivamente. Logo $D_1 = min\{1,427; 1,091\} = 1,091$ é a medida de separação do grupo 1 aos demais grupos do problema. Dessa forma, temos que $CD_1 = \frac{0,142}{1,091} = 0,13$.

De forma análoga, calculamos os valores $CD_2 = 0,103$ e $CD_3 = 0,129$, conforme apresentado nos quadros (c) e (d) da Figura 1. Com base nas três razões, temos $CD = max\{0,13; 0,103; 0,129\} = 0,13$, o qual quantifica a qualidade da clusterização do conjunto de dados do exemplo em 3 grupos.

Na seção seguinte, conduzimos experimentos numéricos que demonstram o desempenho do índice CD, comparado a outras métricas de validação internas amplamente utilizadas na literatura.

DOI: 10.5540/03.2022.009.01.0328

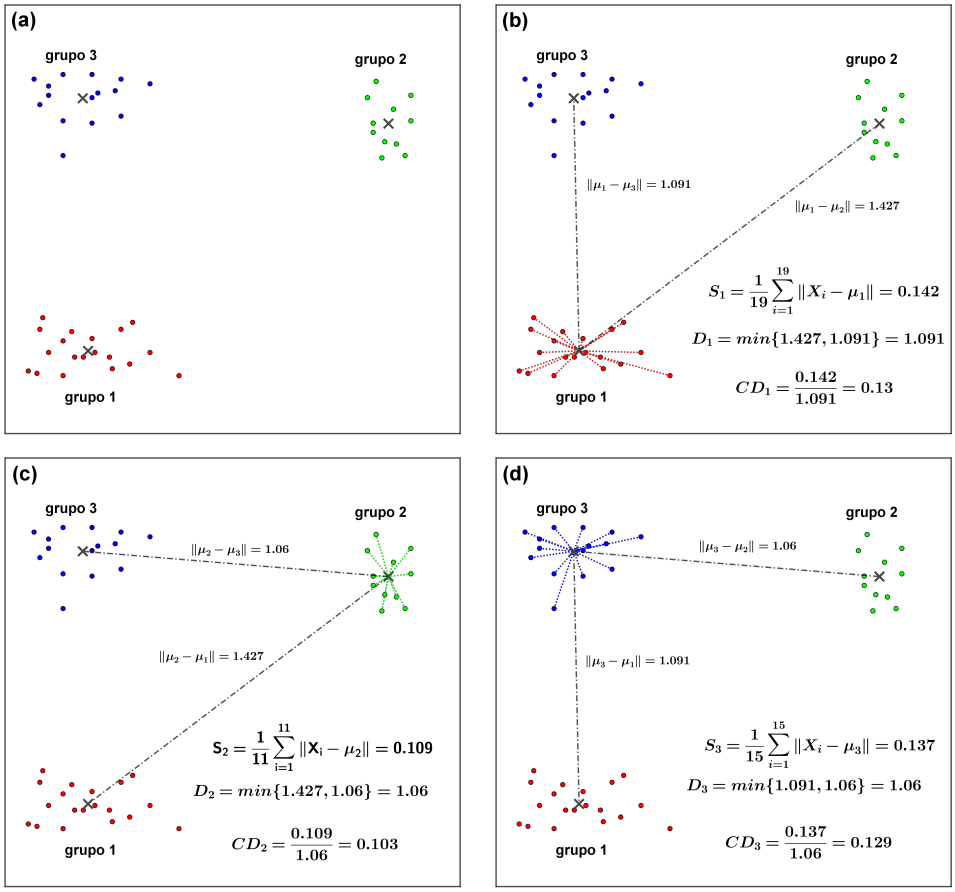

Figura 1: Ilustração do índice CD.

3 Experimentos numéricos

Para realizar a verificação do desempenho do índice CD como métrica de validação interna, optamos por utilizar como base para os testes o Algoritmo CLARA, uma vez que esse necessita apenas da informação do número de grupos como parâmetro de entrada e possui um tempo de execução consideravelmente baixo. Cabe observar que outros algoritmos, tais como: K-means, K-medoids e GMM, também poderiam ser utilizados para essa verificação.

Para analisar o desempenho do índice CD, executamos 9 configurações de agrupamentos, ou seja, consideramos $k = 2, \ldots, 10$, sendo k o número de grupos, em 14 problemas diferentes, totalizando 126 execuções. A partir dos agrupamentos quantificamos os resultados por meio de 3 métricas de validação interna: o índice DB, proposto por [2], o coeficiente de silhueta, proposto por [9] e o índice CD. Os resultados são apresentados na Figura 2.

Na Figura 2 em cada um dos gráficos apresentados, o eixo horizontal expressa o número de grupos, enquanto o eixo vertical expressa os valores das métricas de validação. Adotamos nesse trabalho como critérios de comparação a indicação de cada uma das métricas do número de grupos considerado ideal para cada um dos problemas (representada nos gráficos pelo símbolo "+") e o tempo de execução.

Analisando primeiramente os agrupamentos indicados como ótimos pelas 3 métricas, encontra-

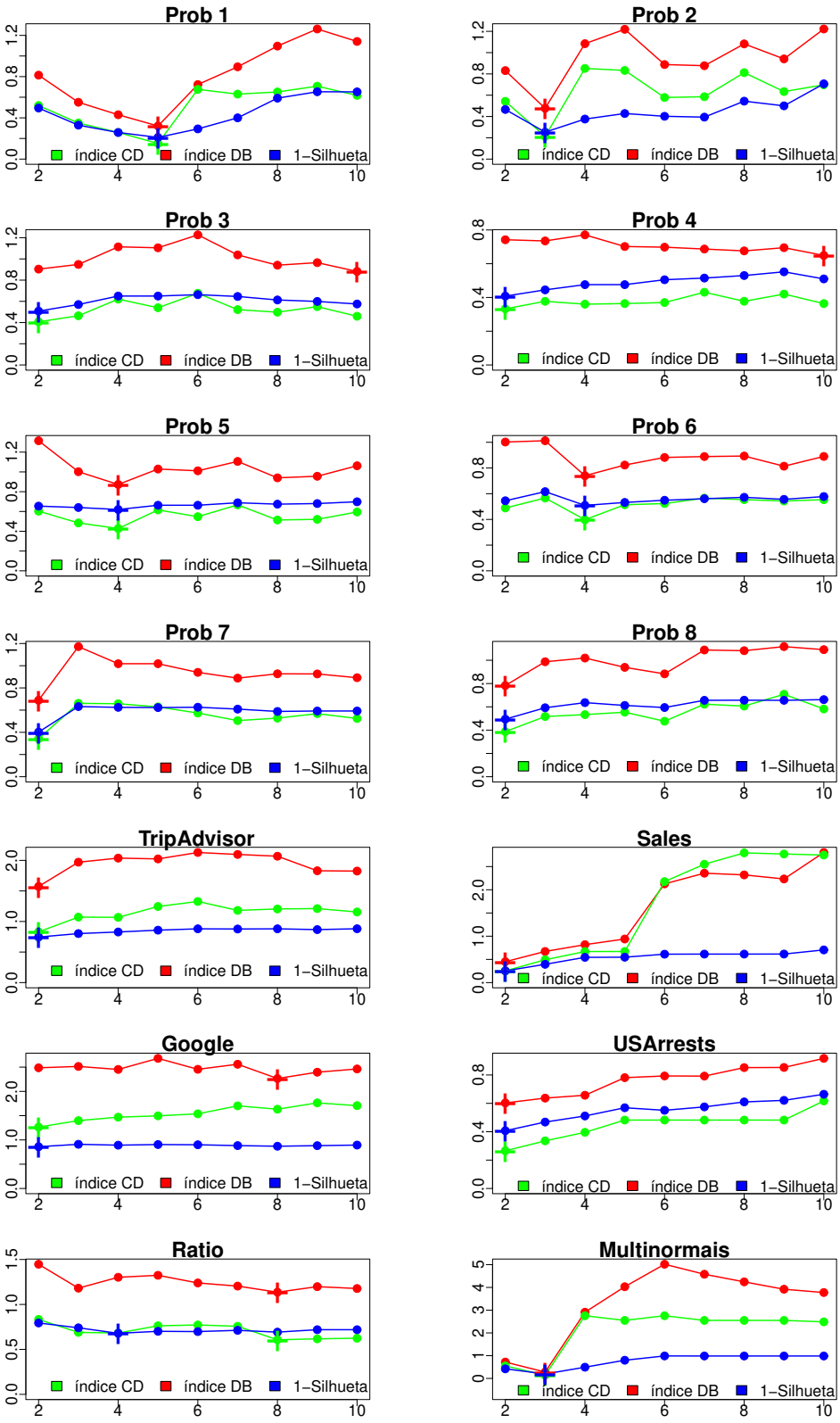

Figura 2: Avaliação das 3 métricas para os diferentes problemas com variação das configurações de clusterização

mos uma correspondência mútua em 10 dos 14 problemas testados, 71,4% dos problemas. Outro fato interessante é que não houve divergência entre as 3 métricas, sempre ao menos duas delas indicaram a mesma quantidade de grupos como sendo ótima. Nesse sentido, os resultados do índice CD coincidiram em 85,7% dos casos com o coeficiente de silhueta e em 78,6% com o índice DB.

A avaliação do desempenho do índice CD por meio, exclusivamente, do percentual de correspondência é simples e de fácil visualização, contudo pode gerar questionamentos da real confiabilidade da qualidade da métrica em avaliar a clusterização multidimensional. Nesse sentido, para garantir uma análise consistente usamos dois testes de concordância. O primeiro, denominado Kappa (κ), proposto por [1], e executado por meio da rotina "Kappa2", disponível no *Software R* pelo pacote "irr"[4]. O segundo, proposto por [3], denominado Fleiss-κ, calculado pela rotina "kappam.fleiss", também disponível no pacote "irr" do *Software R*.

O coeficiente Kappa avalia a concordância entre dois avaliadores, neste caso, entre os índices DB e CD, entre o índice CD e o coeficiente de silhueta e, por último, entre o índice DB e o coeficiente de silhueta. Enquanto o coeficiente de Fleiss-κ, avalia a concordância entre n avaliadores, nesse caso os 3 avaliadores, índices DB, CD e o coeficiente de silhueta. Considerando o número ótimo de grupos para cada uma das 3 métricas sobre os 14 problemas analisados aplicamos o teste de concordância de Kappa e Fleiss-κ e os resultados são apresentados na Tabela 1.

Tabela 1: Resultados dos testes de concordâncias.

	CD e DB	CD e silhueta	DB e silhueta	Todas as métricas
κ	0,71	0,884	0,614	0,724
Concordância	substancial	quase perfeita	substancial	substancial
p-valor	2,3e-08	4,01e-08	1,03e-06	0

De acordo com [6], resultados para κ e Fleiss-κ entre 0,61 e 0,80 indicaram que existe uma concordância substancial entre os avaliadores, enquanto valores superiores a 0,81 indicaram uma concordância quase perfeita. Além disso, o p-valor inferior a 0,001 demonstra que a concordância entre as métricas não é puramente aleatória. Analisando a Tabela 1, percebemos que os valores para κ, foram todos superiores a 0,61, com destaque para a correspondência de 0,884 entre o índice CD e o coeficiente de silhueta, indicando uma correspondência quase perfeita. O coeficiente de Fleiss-κ também apresentou uma concordância substancial entre as 3 métricas analisadas com um valor de 0,724. Por fim, todas os testes de Kappa e Fleiss-κ, apresentaram um p-valor $< 0,001$, com isso rejeitamos a hipótese de que a concordância entre as métricas é puramente aleatória, caracterizando-as como concordantes.

Quanto ao tempo de execução utilizado pelas 3 métricas para avaliar os 9 diferentes agrupamentos para cada um dos 14 problemas, ou seja, para as 126 avaliações, temos que o índice CD precisou de um total de 1,778 segundos, o índice DB levou 1,470 segundos, enquanto o coeficiente de silhueta precisou de 276,131 segundos para avaliar todas clusterizações. Observamos uma proximidade de valores do tempo gasto pelos índices CD e DB, com uma leve vantagem para o índice DB. Por outro lado, o coeficiente de silhueta apresentou tempo muito superior as outras duas, indiciando a inviabilidade da sua utilização principalmente para problemas de dimensões mais elevadas.

Com base nos resultados expostos é possível garantir que o índice CD, proposto nesse trabalho, é eficiente para medir a qualidade de agrupamentos quando comparado às métricas já consagradas na literatura tais como o índice DB [2] e o coeficiente de Silhueta [9]. O índice CD mostrou uma correspondência substancial com o índice DB com tempo de execução similar, enquanto que com o coeficiente de silhueta sua correspondência foi quase perfeita com um tempo de execução muito menor, evidenciando seu uso como métrica de validação interna para clusterização.

DOI: 10.5540/03.2022.009.01.0328

4 Considerações Finais

Apresentamos neste trabalho uma nova métrica de validação interna de clusterização, o índice de Densidade da Clusterização, o índice CD. Tal índice possui a característica de avaliar cada grupo por meio da máxima razão entre a dispersão interna dos grupos com a separação entre grupos. De acordo com os experimentos, o índice CD apresentou uma concordância significativa com as outras duas métricas de validação já consagradas na literatura, o índice DB e o coeficiente de silhueta. Com relação ao tempo de execução, o índice CD mostrou-se similar ao índice DB e superior ao coeficiente de silhueta, principalmente para problemas com elevado número de observações.

Mediante ao exposto, julgamos que o índice CD é uma contribuição significativa para o campo de estudo da clusterização.

Referências

[1] Jacob Cohen. "A coefficient of agreement for nominal scales". Em: **Educational and psychological measurement** 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[2] David L Davies e Donald W Bouldin. "A cluster separation measure". Em: **IEEE transactions on pattern analysis and machine intelligence** 2 (1979), pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

[3] Joseph L Fleiss. "Measuring nominal scale agreement among many raters." Em: **Psychological bulletin** 76.5 (1971), p. 378. DOI: 10.1037/h0031619.

[4] Matthias Gamer, Jim Lemon e Ian Fellows Puspendra. **irr: Various Coefficients of Interrater Reliability and Agreement**. R package version 0.84.1. 2019. URL: https://CRAN.R-project.org/package=irr.

[5] Alboukadel Kassambara. **Practical guide to cluster analysis in R: unsupervised machine learning**. Vol. 1. STHDA, 2017. ISBN: 978-1542462709.

[6] J Richard Landis e Gary G Koch. "The measurement of observer agreement for categorical data". Em: **biometrics** (1977), pp. 159–174. DOI: 10.2307/2529310.

[7] John Mount e Nina Zumel. **Practical data science with R**. Simon e Schuster, 2019. ISBN: 978-1-617-29587-4.

[8] Quynh H Nguyen e Victor J Rayward-Smith. "Internal quality measures for clustering in metric spaces". Em: **International Journal of Business Intelligence and Data Mining** 3.1 (2008), pp. 4–29. DOI: 10.1504/IJBIDM.2008.017973.

[9] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". Em: **Journal of computational and applied mathematics** 20 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[10] Alan Said e Vicenç Torra. **Data Science in Practice**. Springer, 2019. ISBN: 978-3-319-97555-9.

