Clusterização de Artigos da Revista Avian Research desde 2022

Autor: Lucas Perondi Kist (RA: 236202)

1 Introdução

O estudo de pássaros e seus comportamentos é relevante no âmbito da compreensão de modos de conservação da biodiversidade das espécies, seus papéis ecológicos, dentre outros. Nesse sentido, a revista científica open-source *Avian Research* (ISSN 2055-6187, disponível em `https://www.sciopen.com/journal/2055-6187`, acesso em 18/05/2024) surge como um veículo de publicação de *papers* relacionados a esses temas.

Assim, ao analisar os artigos publicados nela, podem surgir questões relacionadas a quantos agrupamentos diferentes de publicações existem e quais suas características. Dessa forma, o objetivo deste trabalho foi identificar os diferentes *clusters* e caracterizá-los, estabelecendo os critérios que seriam utilizados para isso. Todos os resultados apresentados podem ser reproduzidos a partir dos arquivos disponíveis em `https://github.com/lpkist/Trabalho2_ME921/`.

2 Materiais e Métodos

As análises foram realizadas nas linguagens de programação R (R Core Team (2018)) e Python (Van Rossum e Drake (2009)), sendo que a última foi utilizada apenas para *download* dos *papers*. Inicialmente, foi escolhida a revista científica *Avian Research*, e decidiu-se utilizar apenas os artigos publicados desde 2022. A partir disso, foi utilizado o pacote *selenium* para baixar os respectivos arquivos, que totalizaram 131 *papers* distintos.

Na sequência, definiu-se que os atributos extraídos de cada arquivo seriam: número de caracteres (*caracteres*), de palavras totais (*palavras*) e no dicionário (*palavras_dicionario*), de pontos (*pontos*), vírgulas (*virgulas*) e números (*numeros*). Vale ressaltar que o dicionário de palavras utilizado foi o disponível no pacote *words* e foi considerado apenas o texto do artigo localizado entre as palavras-chave e as referências. Ademais, trabalhou-se com a raiz quadrada desses valores, tendo em vista que podem ser consideradas contagens com distribuição Poisson, conforme sugerido por Agresti (2015) para reduzir a assimetria positiva intrínseca aos dados e aproximá-los da normalidade, que é razoável ao considerar que o menor valor observado é 146.

Com os dados já estruturados, foi aplicado o algoritmo de *clusterização* baseada em modelos de mistura de gaussianas com ruído implementada no pacote *mclust* (Scrucca et al. (2023)). Adicionalmente, foram ajustados mais dois modelos: mistura de gaussianas com ruído sem as informações de *caracteres* e Partição em Medoides (PAM) via pacote *cluster* (Maechler et al. (2022)) com todos os atributos padronizados. Por fim, foram realizadas comparações dos agrupamentos, bem como uma análise dos perfis de artigos identificados.

2.1 *Clustering* baseado em modelo de mistura finita de gaussianas

Conforme apresentado em Bouveyron et al. (2019), suponha que há n observações p-dimensionais, $\boldsymbol{y}_1, ...\boldsymbol{y}_n$, com $\boldsymbol{y}_i = (y_{i1}, ...y_{ip})), i = 1, ..., n$. Então, este modelo representa a função de densidade de $\boldsymbol{y}_i$ como uma média ponderada de K densidades de distribuições normais, isto é

$$f(\boldsymbol{y}_i) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} p_k (2\pi)^{-p/2} |\boldsymbol{\Sigma}_k|^{-1/2} exp\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)\}$$

com $p_k \geq 0, k = 1, ..., K$ e $\sum_{k=1}^{K} p_k = 1$. Dado K, a estimação dos parâmetros é realizada a partir do algoritmo EM. Foi utilizada a implementação disponível no pacote *mclust*, que permite a escolha da estrutura de $\boldsymbol{\Sigma}_k$ em relação a volume, formato e orientação.

Além disso, é possível incluir chutes iniciais para observações que correspondem a ruído (obtidas a partir de uma estimação robusta da covariância). Para seleção de modelos, pode-se utilizar tanto BIC quanto a incerteza do *cluster* de uma observação, dada pela probabilidade de não pertencer ao agrupamento ao qual foi atribuída.

2.2 Partição em Medoides (PAM)

De acordo com o disposto em Brian S. Everitt (2011), o PAM é um algoritmo de *clustering* baseado em distâncias que busca encontrar as K observações (medoides) que minimizam a dissimilaridade em relação às demais. Uma forma de medir a qualidade do agrupamento é a partir da média das silhuetas, que quantifica a distância média de uma observação às demais do próprio *cluster* em relação às do segundo agrupamento mais razoável para ela.

3 Resultados

Inicialmente, foi realizada uma estimação robusta das covariâncias a fim de identificar possíveis valores discrepantes. O gráfico de atributos aos pares, colorido de acordo com um valor ser ou não discrepante, está apresentado na Figura 1. Segundo essa análise, foram apontadas 14 observações como chutes iniciais de ruídos.

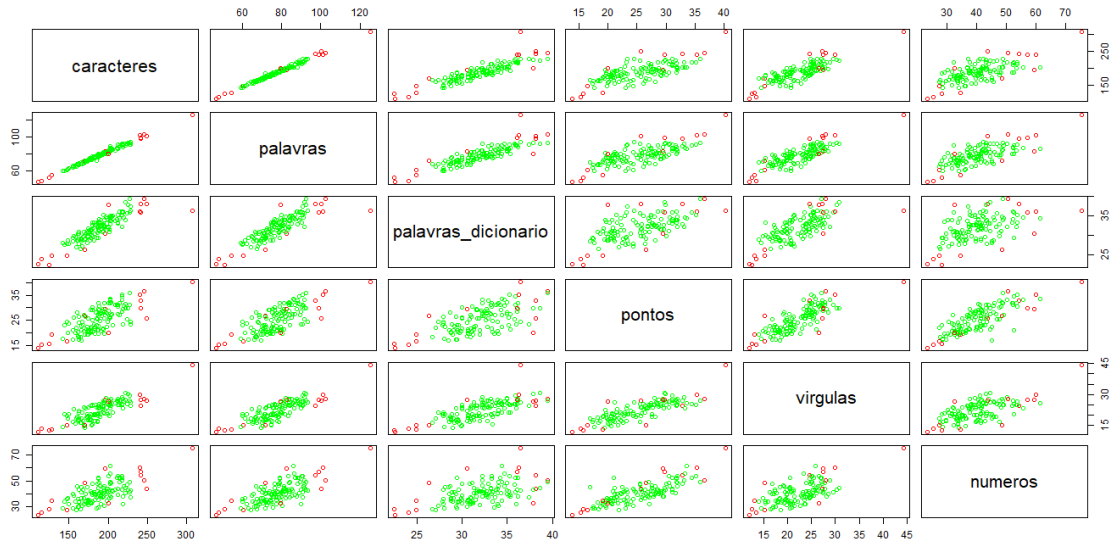

Figura 1: Dispersão de atributos, identificando potenciais ruídos (vermelho) ou não (verde)

Foram ajustados modelos de mistura de gaussianas usando todas as variáveis disponíveis e os BICs estão apresentados na Figura 2. Os três melhores modelos segundo esse critério, todos com dois clusters, e respectivos BICs são: EEE (-4188), VEE (-4193) e EVE (-4195). Os gráficos de incerteza de cada um deles são apresentados na Figura 3.

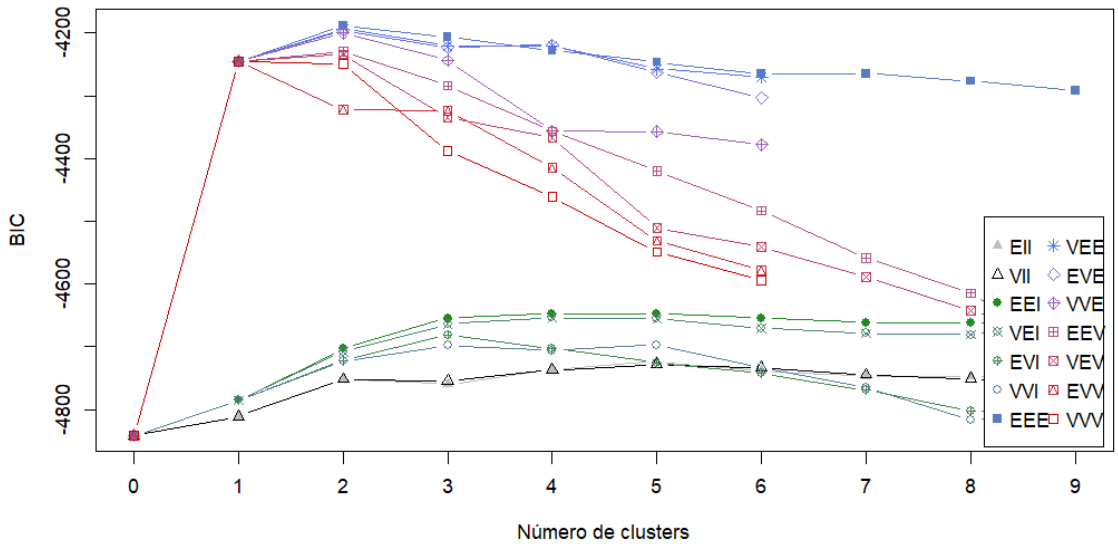

Figura 2: BICs dos modelos de misturas de gaussianas

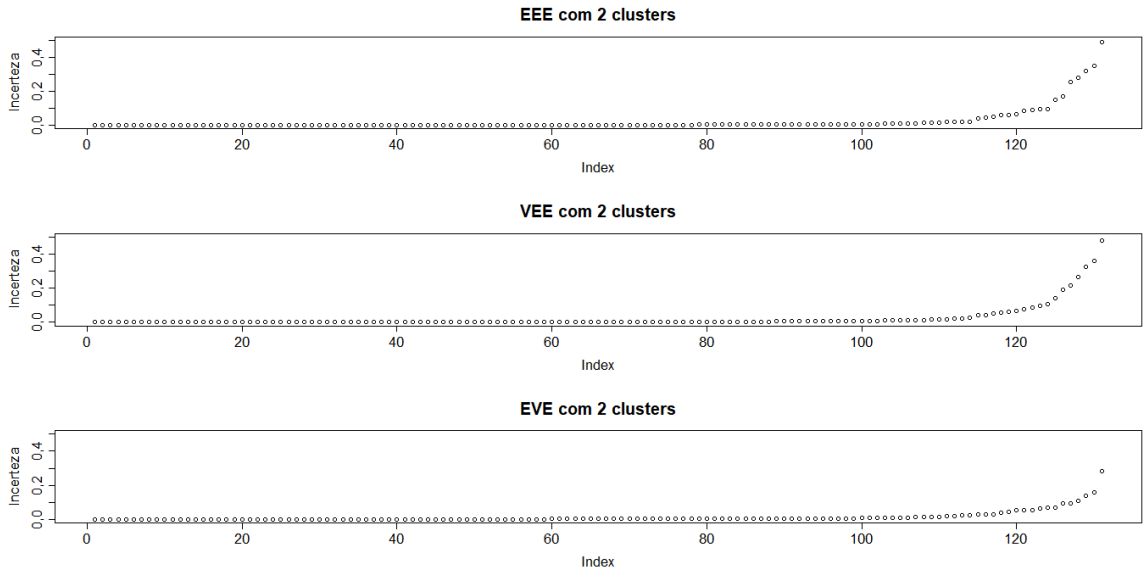

Figura 3: Incertezas dos três melhores modelos segundo critério de BIC

O *cluster* a que cada ponto foi atribuído segundo o modelo EVE com 2 *clusters* (que será referido como EVE,2) está apresentado na Figura 4. Os agrupamentos em preto (ruído), azul e vermelho possuem, respectivamente, 12, 58 e 61 pontos. A análise dos perfis de cada um deles pode ser feita com base na Figura 5.

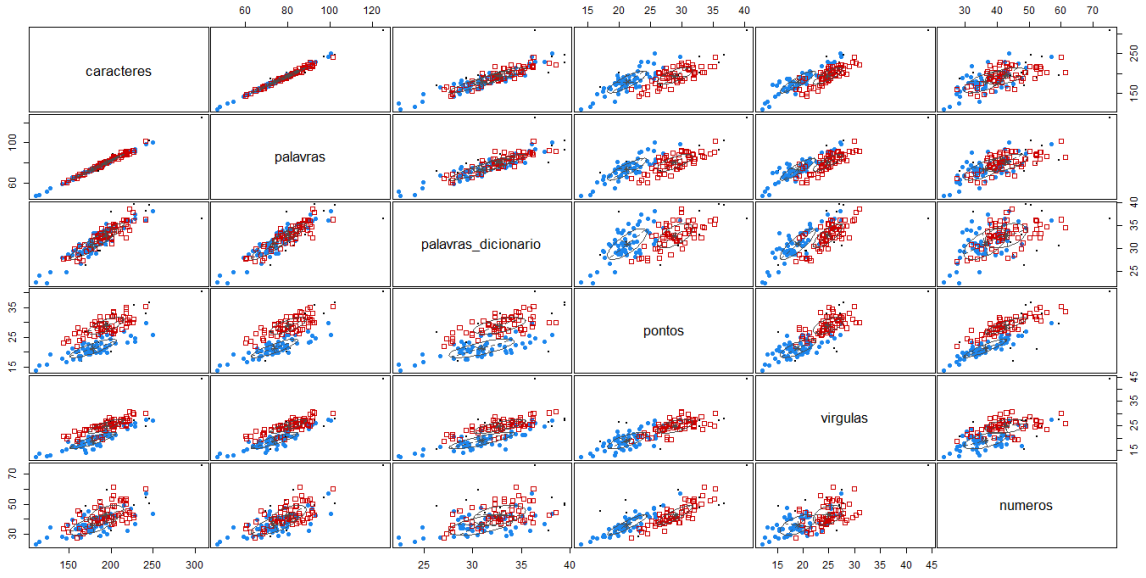

Figura 4: Dispersão de atributos, identificando etiquetas atribuídas aos pontos via EVE,2

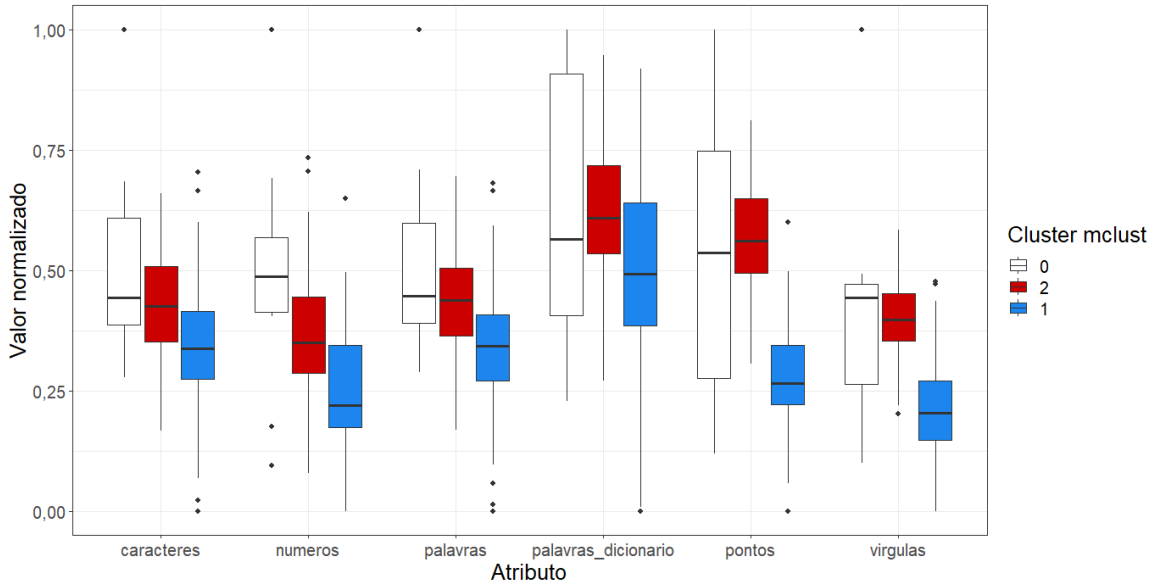

Figura 5: Boxplots dos valores normalizados dos atributos segundo clusters do EVE,2

Como a coluna com o número de caracteres é fortemente correlacionada com o número de palavras, a análise foi repetida desconsiderando-a. O comportamento da evolução dos BICs é bastante semelhante ao retratado na Figura 2. Os três melhores modelos, todos com 2 *clusters*, e respectivos BICs foram: EEE (-3524), VEE (-3528) e EVE (-3538). Os gráficos de incertezas estão apresentados na Figura 6.

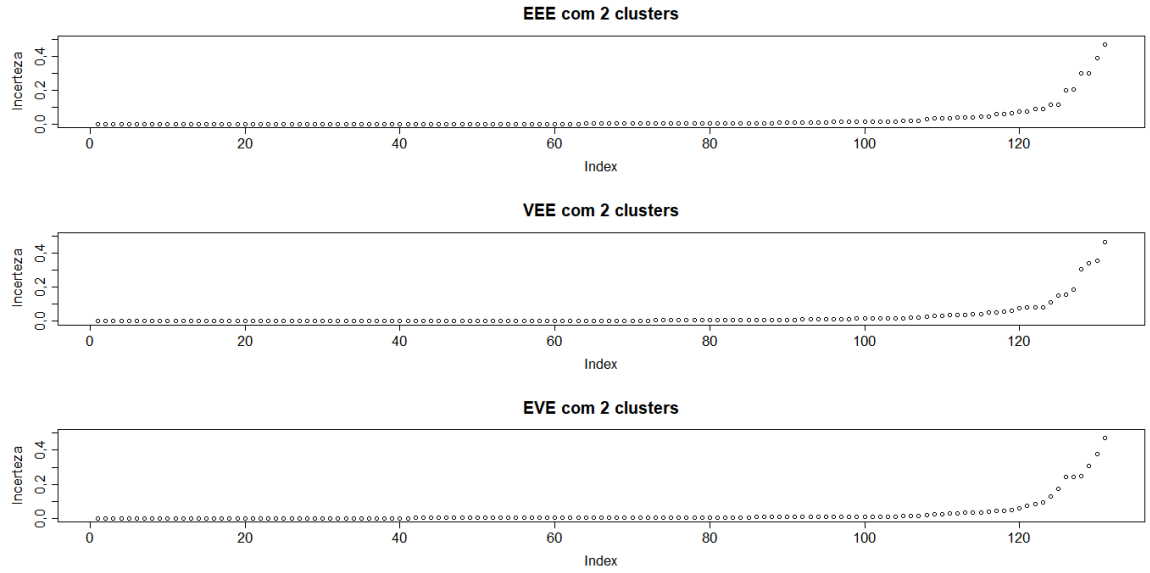

Figura 6: Incertezas dos três melhores modelos segundo critério de BIC (sem *caracteres*)

Na sequência, foi realizado o agrupamento dos dados utilizando todas as características padronizadas a partir do algoritmo PAM. A silhueta média para seis valores de k está apresentada na Tabela 1. Já a dispersão colorida de acordo com as etiquetas atribuídas pelo PAM com $k = 2$ (que será referido como PAM,2) pode ser observada na Figura 7. Adicionalmente, os atributos do medoides estão na Tabela 2 e a concordância das etiquetas atribuídas pelo EVE,2 e pelo PAM,2 podem ser observadas na Tabela 3.

Tabela 1: Silhueta média do PAM com k *clusters* aplicado aos dados padronizados

k	2	3	4	5	6	7
Silhueta média	0,42	0,30	0,29	0,26	0,22	0,23

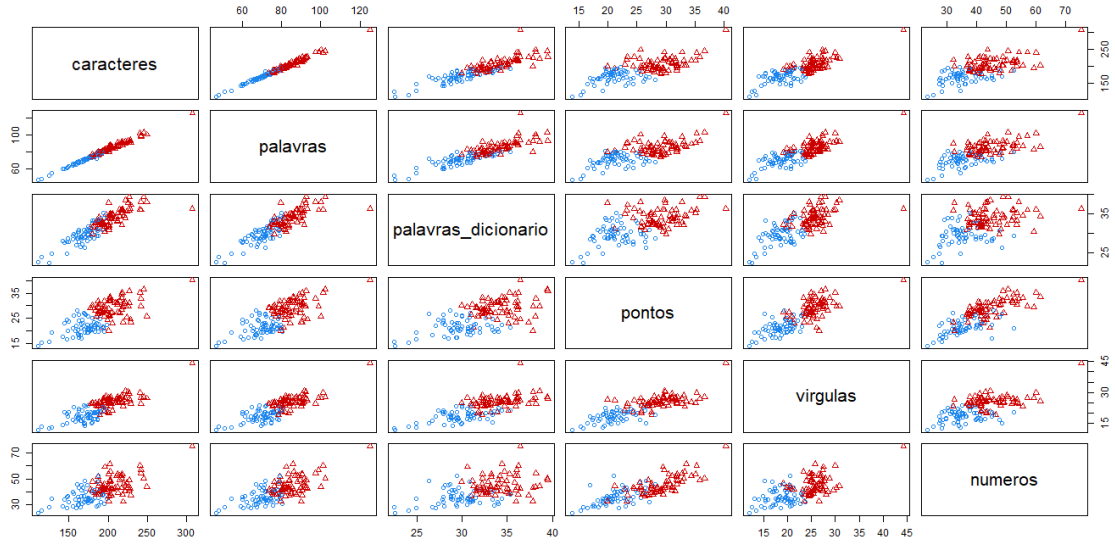

Figura 7: Dispersão de atributos, identificando etiquetas atribuídas aos pontos via PAM,2

Tabela 2: Atributos padronizados dos dois medoides de PAM,2

Medoide	caracteres	palavras	palavras_dicionario	pontos	virgulas	numeros
1	-0,58	-0,64	-0,40	-0,81	-0,71	-0,72
2	0,72	0,51	0,52	1,10	0,72	0,30

Tabela 3: Comparação das etiquetas atribuídas pelos métodos EVE,2 e PAM,2

	Etiqueta EVE,2			
Etiqueta PAM,2	0	1	2	Total
1	4	47	13	64
2	8	11	48	67
Total	12	58	61	131

4 Discussão

Conforme pode ser observado na Figura 1, há alguns pontos destoantes dos demais, que estão relacionados, em geral, a textos excessivamente longos ou curtos. Ademais, é evidente que há forte correlação entre as variáveis. Dessa forma, foi ajustado o modelo de mistura de normais utilizando chutes iniciais de ruídos. A evolução do BIC para os 14 tipos de estrutura de covariância para 0 a 9 *clusters* pode ser visualizada na Figura 2, de onde se conclui que o número adequado de agrupamentos sob essa ótica é dois, além do grupo de ruído.

Considerando apenas os três modelos com menores BICs (os quais diferem do melhor por menos de 0,2%), foram comparadas as incertezas sobre a etiqueta de cada observação na Figura 3. A partir dela, é evidente que o modelo EVE com dois *clusters* (EVE,2) é o melhor, por possuir não só a menor dentre as maiores incertezas, mas também a menor incerteza média. Esse modelo possui $1 + (p + 2 * C) * (p - 1)/2 = 26$ parâmetros para serem estimados, mais do que o EEE (21) e o VEE (22), sendo caracterizado por elipses de volumes variados, mas de mesmos formatos e orientações.

Definido o melhor modelo, foi avaliada a suposição de normalidade conjunta dos atributos, condicionados no *cluster*, a partir da Figura 4. Com base em sua análise, pode-se concluir que tal hipótese é razoável, tendo em vista as formas elípticas presentes nos gráficos, com pontos concentrados dentro da elipse desenhada. Como a variável *caracteres* é fortemente correlacionada com *palavras*, foi realizado um ajuste análogo, mas desconsiderando-a. O comportamento dos BICs foi semelhante ao da Figura 2 e os três melhores modelos foram os mesmos. Para comparar seus desempenhos,

utilizaram-se as incertezas apresentadas na Figura 6, que revela a superioridade do modelo EVE,2 (com todas as características) em relação a eles por possuir menores incertezas máximas e médias.

A análise de perfil dos agrupamentos produzidos pelo modelo EVE,2 com todas as características pode ser realizada com base na Figura 5, na qual são apresentados os atributos normalizados. Com base nela, conclui-se que o *cluster* 1 (azul) corresponde a textos mais curtos, sobretudo com menos vírgulas e pontos, mas apresentando todas as medianas menores do que o primeiro quartil do *cluster* 2 (vermelho), que corresponde a artigos mais longos. Além disso, existe o *cluster* 0 (ruído), que se diferencia, principalmente, por ter mais números e o maior *paper* dentre todos os analisados.

Por fim, foi aplicado o algoritmo PAM aos dados padronizados, pois, dessa forma, não há efeito de escala na distância (o modelo de mistura de gaussianas corrige essas diferenças estimando as variâncias). A Tabela 1 apresenta a silhueta média para alguns valores de k, de onde se depreende que $k = 2$ é a escolha adequada. A dispersão colorida pelos agrupamentos identificados está na Figura 7, que ressalta a existência de um agrupamento de textos menores (azul) e outro com textos maiores (vermelho), fato corroborado pela Tabela 2, que apresenta os atributos dos medoides.

Nesse sentido, percebe-se que ambas as técnicas concordam, identificando dois tipos de *papers*: um com textos menores e outro com textos maiores. Assim, a fim de avaliar o grau de concordância das etiquetas (no sentido de pertencimento ao grupo de textos pequenos ou grandes) foi criada a Tabela 3. A partir dela, conclui-se que, ao todo, 95 etiquetas concordam de um total de 131 (72,5%) ou, desconsiderando o *cluster* de ruídos, 119 (79,8%), indicando alto grau de concordância dos agrupamentos.

5 Conclusão

Com base no exposto, conclui-se que todas as variáveis utilizadas são positivamente correlacionadas, existindo algumas observações que podem ser consideradas ruído. Ademais, apesar da existência de uma correlação quase perfeita, nota-se que há uma piora no ajuste, em termos de incerteza, ao remover *caracteres* da análise, sendo o melhor modelo de covariâncias aquele com estrutura EVE e 2 agrupamentos.

Além disso, após a aplicação do PAM, também foram apontados 2 *clusters*. É interessante ressaltar que ambos os métodos identificaram um grupo de textos menores e outro de artigos maiores, concordando em mais de 72% dos casos. A principal diferença entre os agrupamentos produzidos está no fato de EVE,2 discriminá-los principalmente em termos de pontuação e números. Por fim, apesar de similares, EVE,2 parece ser mais adequado justamente por esse fato.

Referências

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Bouveyron, C., Celeux, G., Murphy, T. B., e Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Brian S. Everitt, Dr Sabine Landau, D. M. L. D. D. S. (2011). *Cluster Analysis, Fifth Edition (Wiley Series in Probability and Statistics)*. Wiley Series in Probability and Statistics. Wiley, 5th edition.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., e Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source).

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scrucca, L., Fraley, C., Murphy, T. B., e Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC.

Van Rossum, G. e Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.