Clusterização hierárquica para determinação do número ótimo de clusters de deputados federais evangélicos eleitos em 2018

DOCUMENTO: SAR-2021-011-JG-v01

De: Felipe Figueiredo **Para:** Josir Gomes

Data: 2021-10-27

SUMÁRIO

1 LISTA DE ABREVIATURAS 2

2 INTRODUÇÃO 2

2.1 Objetivos 3

2.2 Recepção e tratamento dos dados 3

3 METODOLOGIA 3

3.1 Variáveis 3

3.1.1 Desfechos primário e secundário 3

3.1.2 Covariáveis 4

3.2 Análises Realizadas 4

4 RESULTADOS 5

4.1 Análise descritiva 5

4.2 Determinação de clusters para k-means: cluster hierárquico 6

4.2.1 Número ótimo de clusters 6

4.2.2 Descrição dos clusters 8

4.3 Avaliação dos clusters do método k-means 9

4.4 Caso particular com k=2 clusters 10

4.5 Sensibilidade da análise à exclusão de variáveis 11

5 OBSERVAÇÕES E LIMITAÇÕES 13

6 CONCLUSÕES 15

7 REFERÊNCIAS 15

8 APÊNDICE 15

8.1 Disponibilidade 15

8.2 Dados utilizados 15

8.3 Análise exploratória de dados 16

Histórico do documento

Versão	Alterações
01	Versão inicial

1 LISTA DE ABREVIATURAS

- BSS: soma dos quadrados inter grupos (*between sum of squares*)
- DP: desvio padrão
- TSS: soma dos quadrados total (*total sum of squares*)
- WSS: soma dos quadrados intra grupos (*within sum of squares*)

2 INTRODUÇÃO

Avaliação da qualidade do agrupamento de acordo com dois critérios: altura da árvore e número putativo de clusters. Criação do elbow plot para auxiliar a tomada de decisão no uso do k-means.

Esta análise testou a hipótese de que deputados federais evangélicos que foram eleitos em 2018 podem ser identificados com base nas doações recebidas durante a campanha eleitoral, número de votos recebidos e outras características.

O posicionamento político dos deputados foi identificado pelo índice de Power e Silveira-Rodrigues que varia de -1 a 1, onde -1 é mais à esquerda e 1 mais à direita.

A capilaridade é um índice de 0 a 1 que indica quão ampla (sob o ponto de vista geográfico) foi a votação do candidato. Quanto maior o número de zonas eleitorais onde o candidato recebeu votos, maior o índice. Este indicador é a média dos percentis de cada zona eleitoral em que o candidato concorreu.

A classificação dos deputados como pertencendo ou à classe evangélica foi estabelecida por autodenominação, isto é, foram considerados evangélicos os deputados que se autodenominaram como tal.

2.1 Objetivos

1. Avaliar número ideal de clusters de acordo em um dendrograma de clusterização hierárquica, para uso do k-means.
2. Avaliar em particular a acurácia do agrupamento aplicado em 2 clusters com a proposta de identificar os deputados da bancada evangélica vs outros, de acordo com a classificação pré estabelecida.

2.2 Recepção e tratamento dos dados

Base de dados recebida contendo características dos deputados federais eleitos em 2018. Todas as variáveis da tabela de dados analíticos foram identificadas de acordo com as descrições das variáveis, e os valores foram identificados de acordo com o dicionário de dados providenciado. Estas identificações possibilitarão a criação de tabelas de resultados com qualidade de produção final.

Foram feitos ajustes de escala em variáveis para viabilizar a performance dos algoritmos, mantendo a interpretabilidade dos dados. O número de votos, o número de filiados do partido foram padronizados por milhão e todas as variáveis relativas às receitas foram padronizados em milhões de reais – com estas novas escalas todas as variáveis da base tiveram suas amplitudes reduzidas entre 0 e 2.5. O posicionamento foi mantido em sua escala original por já ter amplitude de tamanho 2 (entre -1 e 1).

Uma observação não tinha o número de votos registrada, e foi removida para a análise de clusters.

3 METODOLOGIA

3.1 Variáveis

3.1.1 Desfechos primário e secundário

O desfecho primário está definido como o número ótimo de clusters para uso no algoritmo k-means. O número de clusters foi definido de acordo com a largura média da silhueta no agrupamento hierárquico.

O desfecho secundário é a avaliação do k-means com dois clusters (k=2) para a classificação entre deputados da bancada evangélica e outros deputados explicada pela receita total recebida.

3.1.2 Covariáveis

As seguintes características dos deputados federais foram consideradas para inclusão na análise: Número de votos recebidos, posicionamento político, capilaridade, a unidade da federação, o partido (sigla), o sexo e se é o primeiro mandato. As seguintes

características dos partidos foram consideradas para inclusão na análise: decil do número de deputados eleitos e decil do número de filiados.

As receitas discriminadas em suas diversas origens não foram consideradas na análise principal, pois a densidade das distribuições indica que a maior parte dos deputados não recebeu valores em cada uma destas fontes (figura A1). O número de filiados do partido também não foi incluído em favor do decil de filiados do partido. O impacto da exclusão destas variáveis na clusterização foi avaliado em uma análise de sensibilidade.

3.2 Análises Realizadas

As características dos deputados foram descritas como média (DP) ou frequência e proporção (%), conforme apropriado. As distribuições foram sumarizadas em tabelas e visualizadas em gráficos exploratórios.

O número ótimo de clusters foi determinado por dois métodos: agrupamento hierárquico e o elbow plot do k-means.

O agrupamento hierárquico foi realizado usando a distância euclidiana e o critério de ligação completa (*complete linkage*). O agrupamento hierárquico foi avaliado pelo método da silhueta, onde as larguras das silhuetas de cada observação são calculadas individualmente resultando em uma medida interpretável do pertencimento ao cluster avaliado. Valores positivos de silhueta indicam um bom ajuste da observação ao cluster. A média das larguras das silhuetas foi calculada como critério de avaliação global do agrupamento hierárquico.

O agrupamento diretamente pelo método k-means também foi avaliado pela redução do WSS total (elbow plot). Após a determinação do número ótimo de clusters, dois agrupamentos k-means foram realizados para diagnóstico usando a distância euclidiana e 10 centroides iniciais aleatórios – o primeiro usando o número ótimo de clusters (desfecho primário) e o segundo usando dois clusters (desfecho secundário). O agrupamento do k-means foi avaliado globalmente com o WSS total e foi apresentada a proporção da variância entre grupos que é explicada pelo agrupamento. Esta proporção foi calculada como $\text{variância explicada} = \frac{BSS}{TSS} \times 100$.

Esta análise foi realizada utilizando-se o software R versão 4.1.1.

4 RESULTADOS

4.1 Análise descritiva

Em 2018 foram eleitos 116 (23% dos deputados avaliados) deputados federais que se autodenominaram evangélicos. Destes, 29 (25%) estão filiados à igreja AD, 21 (18%) à

igreja IURD e 15 (13%) à igreja Batista. Observa-se nas duas classes de deputados federais uma predominância do sexo masculino, com 91 (78%) homens entre os deputados evangélicos e 345 (87%) homens dentre as demais classes (Tabela 1). O deputado federal evangélico parece ter posicionamento político mais alinhado à direita, com índice de Power e Silveira-Rodrigues médio 0.7.

A maior parte dos deputados federais eleitos em 2018 foram eleitos em primeiro mandato. Dentro os deputados evangélicos 94 (81%) se elegeram pela primeira vez e 22 (19%) foram reeleitos. Os candidatos evangélicos obtiveram, na média, 127 mil votos, quando os outros deputados obtiveram 97 mil votos. Apesar da discrepância nas médias de performance de votos, a variabilidade deste indicador é substancialmente maior na classe de deputados evangélicos, com desvio padrão superior à média. A variabilidade entre os dois grupos pode ser comparada pelo CV – o CV dos deputados evangélicos foi 157% enquanto nos outros foi 60%. Ambos os grupos tiveram capilaridade semelhante, em torno de 0.8 na média.

Os partidos tiveram desempenho comparáveis, onde tanto o decil do número de deputados eleitos como o decil do número de filiados ficaram na faixa entre 0,7 e 0,8 (Tabela 1). Os partidos que mais abrigaram os deputados evangélicos eleitos foram o PRB com 23 (20%) deputados, PSL com 16 (14%), PR com 10 (8.6%) e os partidos com menor representatividade desta classe foram PATRIOTA, PMN, PRP e PTC todos com 1 (0.9%) deputado.

Tabela 1 *Características dos deputados federais eleitos em 2018.*

Características	Outros, N = 397[1]	Evangélico, N = 116[1]
Receita total (milhão R$)	1.12 (0.68)	1.08 (0.69)
Capilaridade	0.77 (0.15)	0.80 (0.16)
Releição vs primeiro mandato		
Primeiro mandato	345 (87%)	94 (81%)
Reeleito	52 (13%)	22 (19%)
Sexo		
Masculino	345 (87%)	91 (78%)
Feminino	52 (13%)	25 (22%)
Votos (milhão)	0.10 (0.06)	0.13 (0.20)
Índice de Power e Silveira-Rodrigues	0.17 (0.50)	0.42 (0.30)
Decil do núm. de filiados	0.79 (0.22)	0.70 (0.20)
Decil do núm. de deputados	0.78 (0.22)	0.74 (0.22)
[1]Média (Desvio Padrão); n (%)		

A receita total obtida pelos deputados evangélicos foi comparável aos demais deputados, com ambas as classes atingindo receita média superior a 1 milhão de reais. Os deputados evangélicos obtiveram, na média, 1.08 milhões de reais (desvio padrão 0.7 milhões, CV 64%) e os demais deputados 1.12 milhões de reais (desvio padrão 0.7 milhões, CV 60%). A amplitude das receitas observadas variou entre R$ 21648 e R$ 2507377 entre os deputados evangélicos e R$ 12075 e R$ 2500500 entre os demais (Tabela 1 e Figura 1).

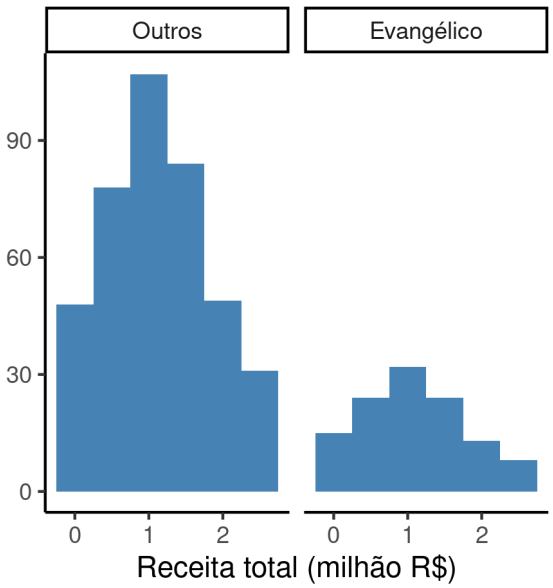

Figura 1 *Distribuição da receita total dos deputados federais eleitos em 2018.*

4.2 Determinação de clusters para k-means: cluster hierárquico

4.2.1 Número ótimo de clusters

O agrupamento hierárquico foi avaliado para números de clusters (k) entre 2 e 10, e as silhuetas médias de cada agrupamento particular foram calculadas (Figura 2). O número de clusters com maior silhueta média é k = 4.

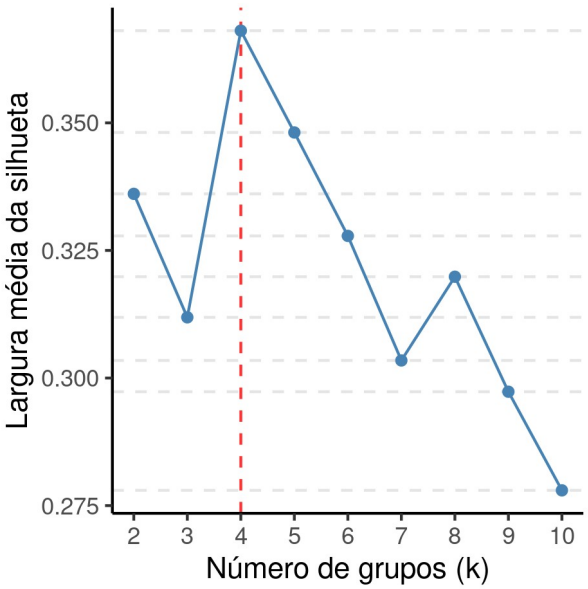

Figura 2 *Variação da silhueta média do cluster hierárquico para diferentes valores de k.*

Estipulando o número ótimo de clusters em k = 4, as silhuetas das observações variam entre -0.28 e 0.66 com média 0.37 e mediana 0.40. Este agrupamento em particular pode ser identificado cortando-se o dendrograma do agrupamento hierárquico na altura h = 2.2 (Figura 3). Isto significa que a maior distância entre dois clusters não supera h = 2.2.

As silhuetas médias dos clusters indicam que o agrupamento foi bem-sucedido em discriminar características dos deputados com os dados avaliados. O cluster 1 incluiu n = 226 deputados e possui silhueta média 0.28, o cluster 2 incluiu n = 153 deputados e possui silhueta média 0.50, o cluster 3 incluiu n = 132 deputados e possui silhueta média 0.36 e o cluster 4 incluiu n = 2 deputados e possui silhueta média 0.50.

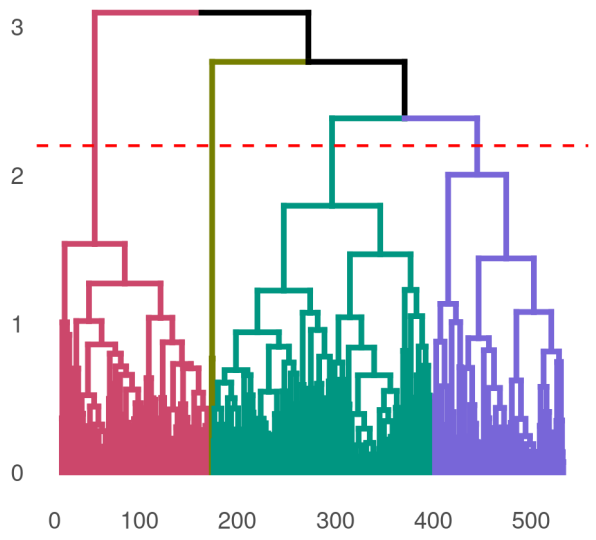

Figura 3 *Dendrograma do cluster hierárquico, corte na altura h = 2.2.*

4.2.2 Descrição dos clusters

Os quatro clusters obtidos não foram suficientes para discriminar os deputados evangélicos do restante, mas a maioria destes ficou distribuída em dois dos grupos (clusters 1 e 2 na Tabela 2). O cluster 1 tem 226 deputados (verde na figura 3) e em sua composição 30% dos deputados incluídos são evangélicos, enquanto o cluster 2 tem 153 deputados dos quais 24% são evangélicos (vermelho na figura 3). A principal característica que difere entre esses clusters é a receita total onde o cluster 1 obteve em torno de R$ 700 milhões e o cluster dois obteve quase R$ 2 milhões.

Tabela 2 *Características dos clusters de deputados federais eleitos em 2018, determinadas por agrupamento hierárquico com k = 4 clusters.*

Características	1, N = 226[1]	2, N = 153[1]	3, N = 132[1]	4, N = 2[1]
Evangélico				
Outros	159 (70%)	116 (76%)	122 (92%)	0 (0%)
Evangélico	67 (30%)	37 (24%)	10 (7.6%)	2 (100%)
Receita total (milhão R$)	0.71 (0.43)	1.94 (0.34)	0.87 (0.44)	0.24 (0.03)
Capilaridade	0.77 (0.16)	0.78 (0.16)	0.78 (0.15)	0.84 (0.00)
Votos (milhão)	0.10 (0.07)	0.10 (0.06)	0.10 (0.05)	1.46 (0.54)
Índice de Power e Silveira-Rodrigues	0.50 (0.23)	0.44 (0.22)	-0.48 (0.23)	0.76 (0.00)

Decil do núm. de filiados	0.67 (0.21)	0.85 (0.18)	0.84 (0.21)	0.60 (0.00)
Decil do núm. de deputados	0.73 (0.25)	0.81 (0.16)	0.80 (0.20)	1.00 (0.00)
[1]n (%); Média (Desvio Padrão)				

O cluster 3 (azul na Figura 3) parece identificar a maioria dos deputados com posicionamento mais à esquerda (Índice de Power e Silveira-Rodrigues negativo), e possui baixa representatividade de deputados evangélicos (n=10, 8% do grupo).

É possível identificar um cluster pequeno (n=2) que agrupou dois deputados com mais de 1 milhão de votos (cluster 4 na Tabela 2, verde-escuro na Figura 3). Ambos são evangélicos.

4.3 Avaliação dos clusters do método k-means

O método k-means foi utilizado de forma exploratória para determinação do número ótimo de clusters conforme mensurado pela queda do WSS. Foram avaliados os possíveis números de clusters (k) de 1 a 10. O número ótimo de clusters obtido foi k = 3, onde houve a queda de WSS mais expressiva (Figura 4). Usando k = 3 clusters o WSS total foi 178.47 que explica 79.7% da variância entre os grupos.

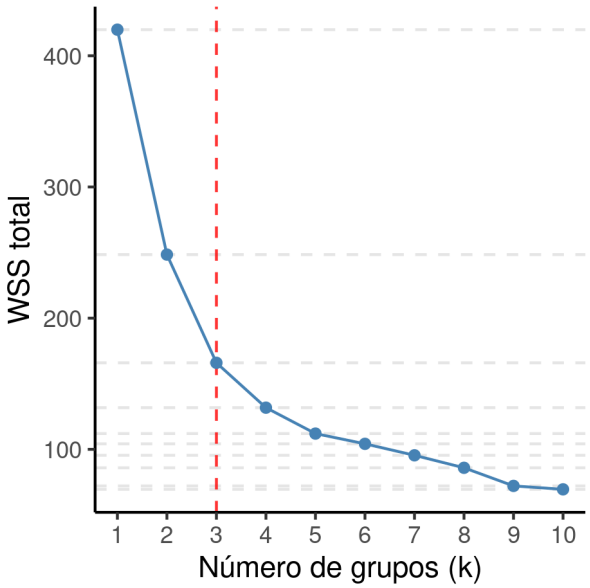

Figura 4 *Elbow plot do método k-means.*

4.4 Caso particular com k=2 clusters

Os métodos descritos nas seções anteriores foram aplicados ao caso particular de se buscar exatamente 2 agrupamentos que discriminassem os deputados evangélicos dos demais.

Clusterização hierárquica

Estipulando o número de clusters em k = 2 no agrupamento hierárquico, as silhuetas das observações variam entre -0.38 e 0.67 com média 0.34 e mediana 0.38. O cluster 1 incluiu n = 360 deputados e possui silhueta média 0.25 e o cluster 2 incluiu n = 153 deputados e possui silhueta média 0.54.

A tentativa de agrupar os deputados em k = 2 clusters não identificou um grupo com maioria de evangélicos, com ambos os clusters incluindo cerca de 20% de deputados evangélicos (Tabela 3) com os dados incluídos na análise. A principal característica que parece discriminar estes grupos é a receita total recebida, onde o cluster 1 obteve em média R$ 760 mil, enquanto o cluster 2 obteve próximo de R$ 2 milhões.

Tabela 3 *Características dos clusters de deputados federais eleitos em 2018, determinadas por agrupamento hierárquico com k = 2 clusters.*

Características	1, N = 360[1]	2, N = 153[1]
Evangélico		
Outros	281 (78%)	116 (76%)
Evangélico	79 (22%)	37 (24%)
Receita total (milhão R$)	0.76 (0.44)	1.94 (0.34)
Capilaridade	0.78 (0.15)	0.78 (0.16)
Votos (milhão)	0.11 (0.12)	0.10 (0.06)
Índice de Power e Silveira-Rodrigues	0.14 (0.53)	0.44 (0.22)
Decil do núm. de filiados	0.73 (0.23)	0.85 (0.18)
Decil do núm. de deputados	0.76 (0.24)	0.81 (0.16)

[1]n (%); Média (Desvio Padrão)

Método k-means

A qualidade do agrupamento por k-means também foi avaliada para este caso particular. Usando k = 2 clusters o WSS total foi 465.37 que explica 47.0% da variância entre os grupos.

4.5 Sensibilidade da análise à exclusão de variáveis

Método k-means

O método elbow plot para determinação do número de clusters do k-means não demonstrou sensibilidade à exclusão das variáveis de origens individuais de receita e número de filiados, tendo resultado no mesmo número k=3 de clusters da análise principal (Figura 5). Com o número ótimo de clusters (k = 3) o WSS total foi 514.73 que explica 56.4% da variância entre os grupos. A exclusão das variáveis causou uma redução do WSS total para 336.26, associada a um aumento de 23.3% na variância explicada ente os grupos.

Usando k = 2 clusters o WSS total foi 681.26 que explica 42.3% da variância entre os grupos. A exclusão das variáveis causou uma redução do WSS total para 215.89, associada a um aumento de 4.7% na variância explicada entre os grupos.

Clusterização hierárquica

O método de agrupamento hierárquico se mostrou sensível à exclusão destas variáveis. Com o dataset completo o número ótimo de clusters é k=8 onde a maior silhueta média foi 0.29 (figura 6). Esta silhueta média é substancialmente menor que o resultado obtido na análise principal, indicando uma menor coesão dos grupos gerados e portanto uma menor qualidade dos agrupamentos. Ao comparar o mesmo número de clusters da análise principal (k=4) observa-se silhueta média foi 0.28. A exclusão das variáveis causou um aumento da silhueta média do agrupamento em 0.14.

Os dois agrupamentos acima podem ser observados na figura 7, com seus respectivos cortes nas alturas do dendrograma. Observa-se em ambos os casos que a maior parte dos deputados encontram-se nos primeiros clusters. No caso k=4 (Figura 7A) os deputados são majoritariamente agrupados nos 3 primeiros clusters, e o último cluster possui 6 deputados. No caso k=8 (Figura 7B) apenas 10 deputados foram alocados nos últimos 4.

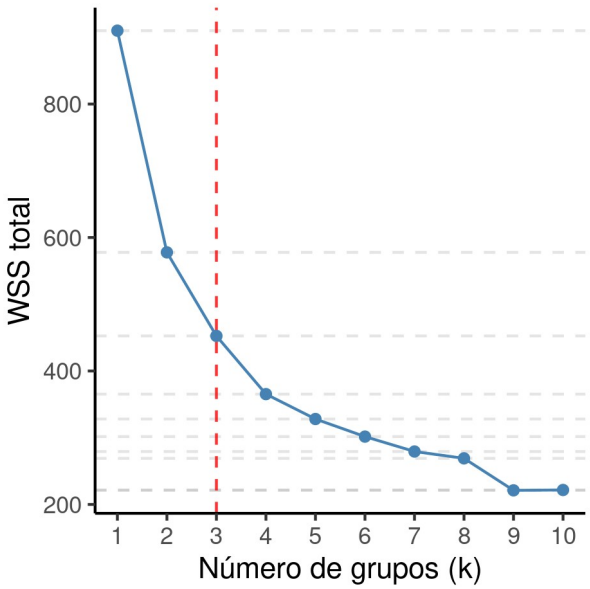

Figura 5 *Elbow plot do método k-means com o dataset completo.*

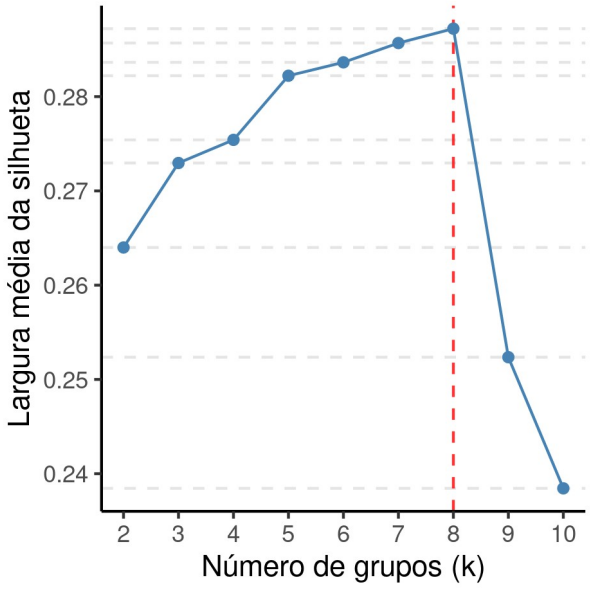

Figura 6 *Variação da silhueta média do cluster hierárquico para diferentes valores de k, com o dataset completo.*

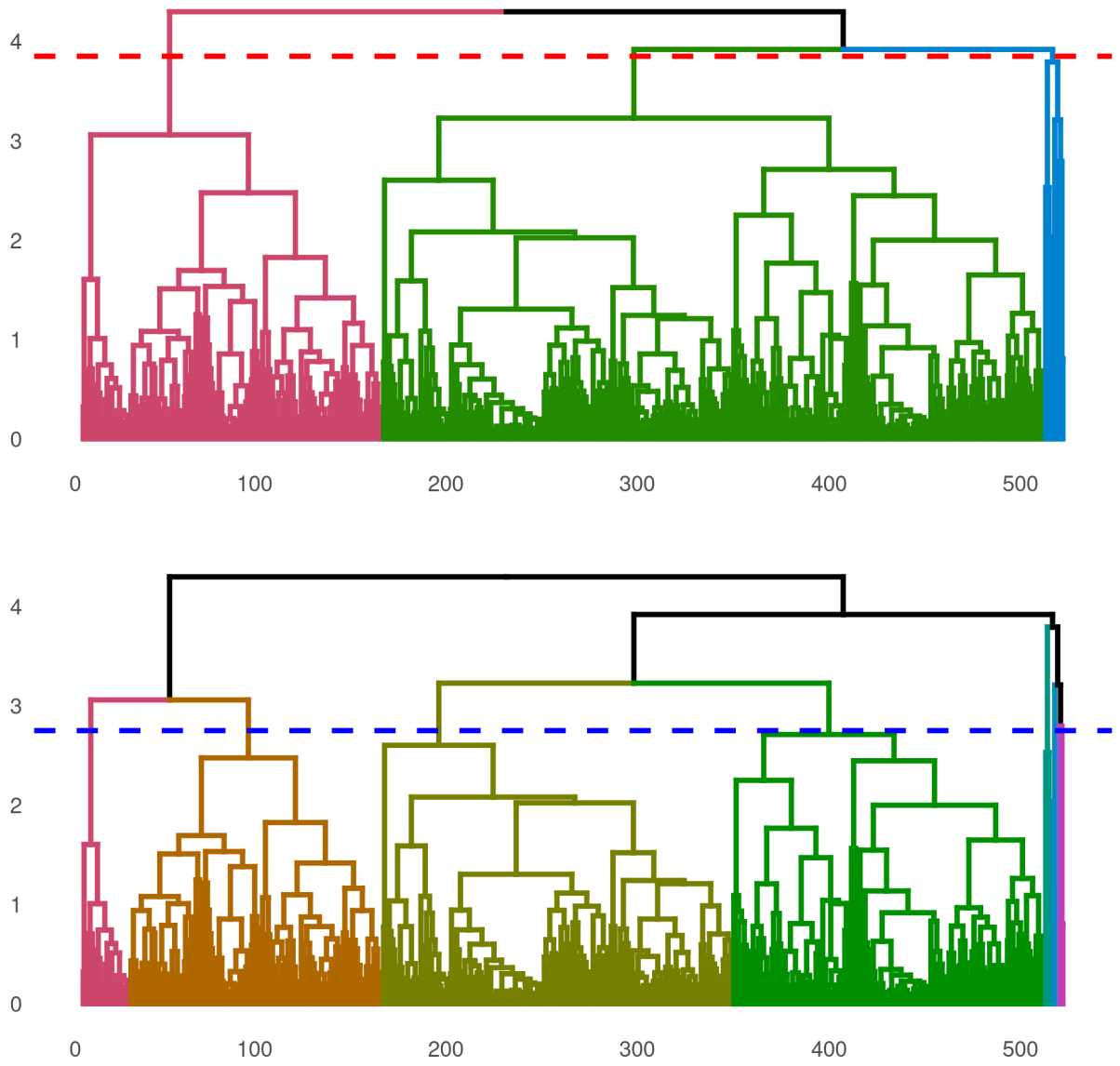

Figura 7 *Dendrograma do cluster hierárquico com o dataset completo, cortes nas alturas h = 3.85 (vermelho) e h = 2.75 (azul).*

5 OBSERVAÇÕES E LIMITAÇÕES

As escalas das variáveis foram ajustadas de acordo com os critérios arbitrados nos métodos ao invés da abordagem usual de padronizar em relação à média e DP dos dados. Esta escolha foi feita de modo a manter a interpretabilidade dos valores, em

detrimento de um possível ganho de estabilidade numérica do algoritmo. Não há consenso na literatura sobre a abordagem de escalonamento das variáveis e por isto esta escolha metodológica deve ser apresentada explicitamente, junto com a justificativa técnica para tal.

O algoritmo de clusterização hierárquico se mostrou sensível à exclusão de variáveis com baixa informação presumida. Como a maior parte dos deputados possuía valor nulo obtidas destas fontes de receita (Figura A1) é possível que estas variáveis contribuam para agrupamentos mais sensíveis à presença ou ausência de receitas em fontes específicas. Embora a qualidade geral do agrupamento seja maior com a exclusão, é possível que os agrupamentos gerados com todas as variáveis tenham maior poder de explicar os dados

Além disso o agrupamento hierárquico realizado com todos os dados gerou alguns agrupamentos com poucas observações. Especula-se que um aumento do número de clusters nesse dataset seja capaz de isolar os *outliers* em grupos pequenos, criando outros grupos mais balanceados. Talvez isso resulte em um agrupamento que identifique os deputados evangélicos em um ou mais clusters e discrimine as características que os diferem dos demais deputados.

As variáveis do dataset que possuem *outliers* foram estatisticamente detectáveis em outra análise associada ao problema aqui avaliado (**SAR-2021-012-JG-v01**).

Por fim, a presença de *outliers* é sempre delicada, e os resultados precisam ser interpretados com cautela. Com base nos resultados desta análise, duas recomendações podem ser observadas para o tratamento de *outliers* neste dataset:

1. **Remover os *outliers***. Isto implica redefinir a população estudada (por exemplo, deputados federais com número de votos, ou receita, próximos do típico) e, portanto, ajustar o escopo do objetivo do estudo.
2. **Avaliar uma clusterização com maior número (k) de clusters, para tentar isolar os outliers em grupos pequenos**.

6 CONCLUSÕES

O número ótimo de clusters identificado no agrupamento hierárquico foi k = 4 clusters.

Este número de clusters possibilita identificar dois grupos de evangélicos com receitas totais diferentes. Um terceiro grupo parece conter a maioria dos deputados com posicionamento à esquerda política e o último grupo é integrado por 2 deputados que receberam um número de votos desproporcionalmente maior que os demais.

O caso particular com dois clusters foi avaliado mas não foi possível discriminar os deputados evangélicos nesta base de dados. Os dois grupos gerados pelo método parecem diferir principalmente pela renda total recebida.

7 REFERÊNCIAS

- **SAP-2021-011-JG-v01** – Plano Analítico para Clusterização hierárquica para determinação do número ótimo de clusters para classificação de deputados federais como bancada evangélica
- **SAR-2021-012-JG-v01** – Quantificação do efeito da receita recebida na autodenominação como evangélicos em deputados federais de 2018

8 APÊNDICE

8.1 Disponibilidade

Tanto este documento como o plano analítico correspondente (**SAP-2021-011-JG-v01**) podem ser obtidos no seguinte endereço:

https://github.com/philsf-biostat/SAR-2021-011-JG/

8.2 Dados utilizados

Tabela A1 *Estrutura da tabela de dados analíticos*

id	partido	uf	capilaridade	primeira	sexo	evangelico	num_votos	posicao	decil_filiados	decil_deputados	total_receita
1											
2											
3											
…											
513											

8.3 Análise exploratória de dados

Distribuição das receitas, por origem

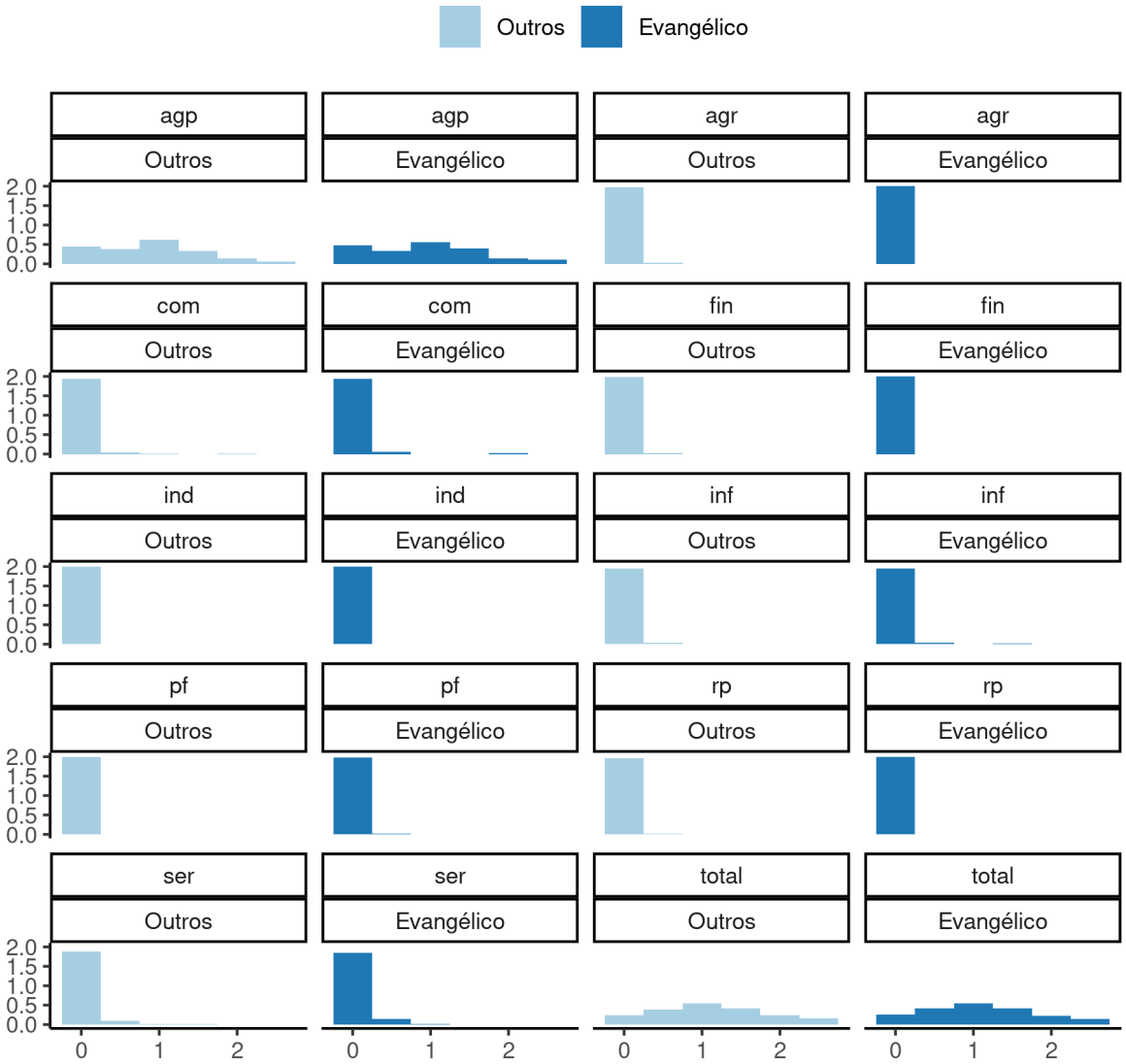

Figura A1 *Distribuição das receitas de deputados federais, por origem (agp = receita que veio do Partido ao invés de apoiadores privados (empresariais ou não); agr = setor agrícola; com = setor do comércio; fin = setor específico dos bancos e outras instituições financeiras e imobiliárias; ind = setores da indústria; inf = setor de infraestrutura; pf = pessoa física; rp = recursos próprios; ser = setor de serviços).*