Otimização de hiperparâmetros de clusterização hierárquica para identificação de deputados evangélicos de corporações pentecostais eleitos em 2018

DOCUMENTO: SAR-2021-017-JG-v01

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

2021-11-18

SUMÁRIO

Otimização de hiperparâmetros de clusterização hierárquica para identificação de deputados evangélicos de corporações pentecostais eleitos em 2018

Histórico do documento

Versão	Alterações
01	Versão inicial

1 LISTA DE ABREVIATURAS

- AGP: receita que veio do partido ao invés de apoiadores privados (empresariais ou não)

2 CONTEXTO

Dados de deputados federais evangélicos, eleitos em 2018, com labels identificando quais pertencem a corporações pentecostais. As corporações pentecostais foram definidas como as entidades evangélicas com interesse predominantemente financeiro e/ou político, diferenciando-se das instituições religiosas tradicionais.

As corporações pentecostais foram previamente identificadas na base de dados recebida.

2.1 Objetivos

Identificar a seleção de hiperparâmetros que maximiza a silhueta média (global) do agrupamento hierárquico de deputados federais evangélicos.

2.2 Recepção e tratamento dos dados

Base de dados recebida contendo características dos deputados federais eleitos em 2018.

As receitas recebidas pelos deputados evangélicos foram divididas entre duas fontes: AGP e outras. As outras receitas foram a soma total das receitas recebidas subtraída da receita AGP.

Todas as variáveis numéricas foram escalonadas para a clusterização de modo a manter os valores observados em um intervalo de amplitude de aproximada 2 unidades (**SAR-2021-011-JG-v01**). Todas as receitas foram escalonadas por milhão de reais. O número de votos será escalonado por milhão de votos. O posicionamento político varia de -1 a 1 e portanto já está limitado a um intervalo de amplitude 2. A capilaridade e os decis foram mantidos na escala original.

3 METODOLOGIA

3.1 Variáveis

3.1.1 Desfechos primário e secundário

O desfecho primário está definido como a combinação de hiperparâmetros k, métrica de distância e método de ligação que maximiza a silhueta média do agrupamento hierárquico.

3.1.2 Covariáveis

As seguintes características dos deputados federais foram incluídas na análise: Receitas (divididas em AGP e outras fontes), número de votos recebidos, posicionamento político e capilaridade. As seguintes características dos partidos foram consideradas para inclusão na análise: decil do número de deputados eleitos e decil do número de filiados.

3.2 Análises Estatísticas

Modelos de clusters hierárquico foram ajustados aos dados numéricos. Foi criado um algoritmo para percorrer o espaço de hiperparâmetros e calcular a silhueta de cada combinação.

Os índices de silhueta média foram visualizados em um gráfico de dispersão (silhueta por k). Cada ponto foi identificado pela métrica de distância e método de ligação (mapeados em cores e formas) para identificação visual da qualidade dos agrupamentos avaliados. A combinação ótima foi destacada textualmente, e outras combinações com valores de silhueta mais altos podem ser identificados.

Hiperparâmetros a ser avaliados:

- Número de clusters k: variando de 2 a 10
- Métricas de distância
    - Norma 2 (Euclidiana)
    - Norma 1 (Manhattan)
    - Norma infinito (máxima)
    - Norma p (Minkowski) com p = 0.5 e 1.5

  - Canberra
- Métodos de ligação
  - Completa (máximo)
  - Single (mínimo)
  - Média (UPGMA)
  - Ward sem critério de Ward (1963)
  - Ward com critério de Ward (1963)
  - Mediana *
  - Centroide *

* Os métodos de ligação por mediana e centroide podem gerar inversões na clusterização hierárquica e foram excluídos das recomendações. Sua inclusão no algoritmo de otimização terá apenas fins informativos.

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

4 RESULTADOS

4.1 Espaço de hiperparâmetros

Ao todo foram incluídas 378 combinações dos três hiperparâmetros, com suas respectivas clusterizações hierárquicas geradas e silhuetas médias calculadas.

A fim de facilitar o uso dos resultados desta análise, foi escolhido descrever os conjuntos de parâmetros de acordo com os argumentos das funções da linguagem R. Desta forma, a escolha de hiperparâmetros baseada nas análises de silhueta média podem ser diretamente aplicados na implementação. Os detalhes dos argumentos a seguir podem ser vistos na documentação das funções `dist` e `hclust`, do R base (pacote `stats`).

As métricas de distância descritas nesta seção são:

- **canberra**: Métrica de Canberra;
- **euclidian**: Métrica Euclidiana (norma 2);
- **manhattan**: Métrica de Manhattan (norma 1);
- **maximum**: Métrica do supremo (norma infinito);
- **minkowski_0.5**: Métrica de Minkowski (norma p), p = 0.5;
- **minkowski_1.5**: Métrica de Minkowski (norma p), p = 1.5.

Os métodos de ligação descritos nesta seção são:

- **average**: Método de ligação média (distância média);
- **centroid**: Método de ligação centroide;
- **complete**: Método de ligação completa (distância máxima);

- **median**: Método de ligação da mediana;
- **single**: Método de ligação single (distância mínima);
- **ward.D**: Método de ligação de Ward (sem critério de Ward 1963);
- **ward.D2**: Método de ligação de Ward (com critério de Ward 1963).

Como a visualização de um espaço com quatro dimensões é desafiadora, vamos inicialmente considerar as projeções em espaços bidimensionais e avaliar a variabilidade de performance mensurado pelas silhuetas médias em cada hiperparâmetro.

Trocando as métricas de distância (Figura 1) a melhor performance foi obtida pela métrica maximum, com silhueta média 0.461, seguida da métrica de Minkowski com p = 0.5 (silhueta média 0.424). As métricas de Minkowski com p = 1.5 e Canberra tiveram a pior performance geral com as maiores silhuetas médias atingindo 0.383 e 0.335, respectivamente. As menores silhuetas médias foram observadas na métrica de Minkowski com p = 0.5 e Manhattan (silhueta média -0.0857 e 0.00716, respectivamente).

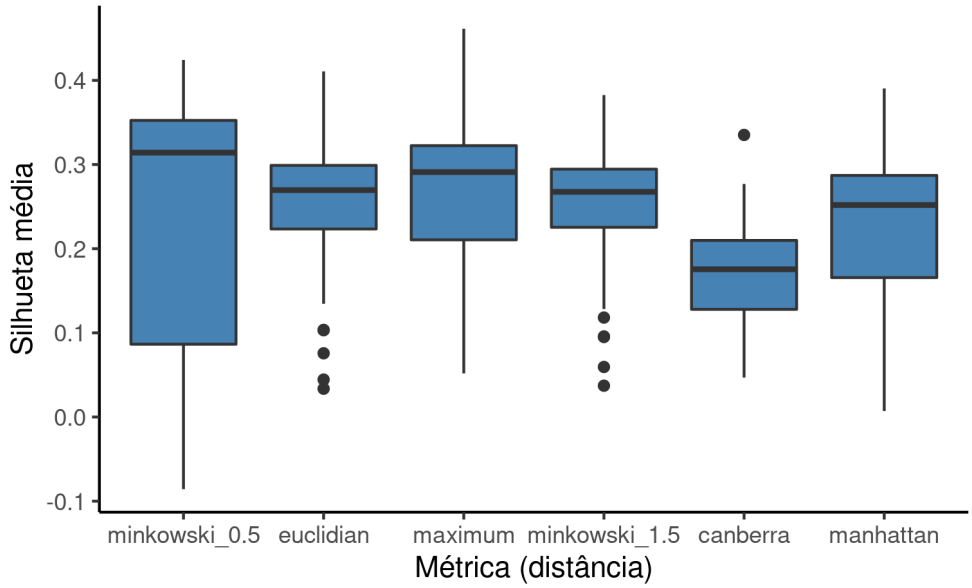

Figura 1 *Silhuetas médias de clusters hierárquicos para diferentes métricas.*

Trocando os métodos de ligação (Figura 2) o maior valor de silhueta foi observado simultaneamente em três métodos: centroide, mediana e single com silhueta média 0.461. O segundo maior valor observado foi outro empate, desta vez entre os métodos ward.D e ward.D2 (0.424). Os dois menores valores de silhueta foram observados com o método single (-0.0857) e mediana (-0.0628).

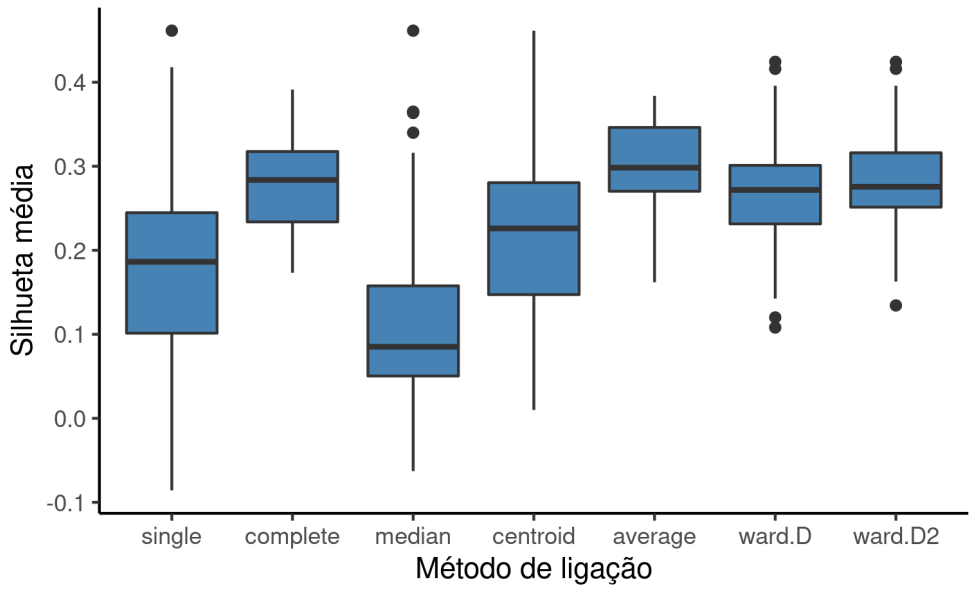

Figura 2 *Silhuetas médias de clusters hierárquicos para diferentes métodos de ligação.*

Trocando o número k de clusters (Figura 3) o maior valor de silhueta média foi observado com k = 2 (0.461), seguido de k = 10 (0.424) e k = 9 (0.416). Os valores mais baixos de silhueta observados foram -0.0857 para k=8 e -0.0628 para k=6.

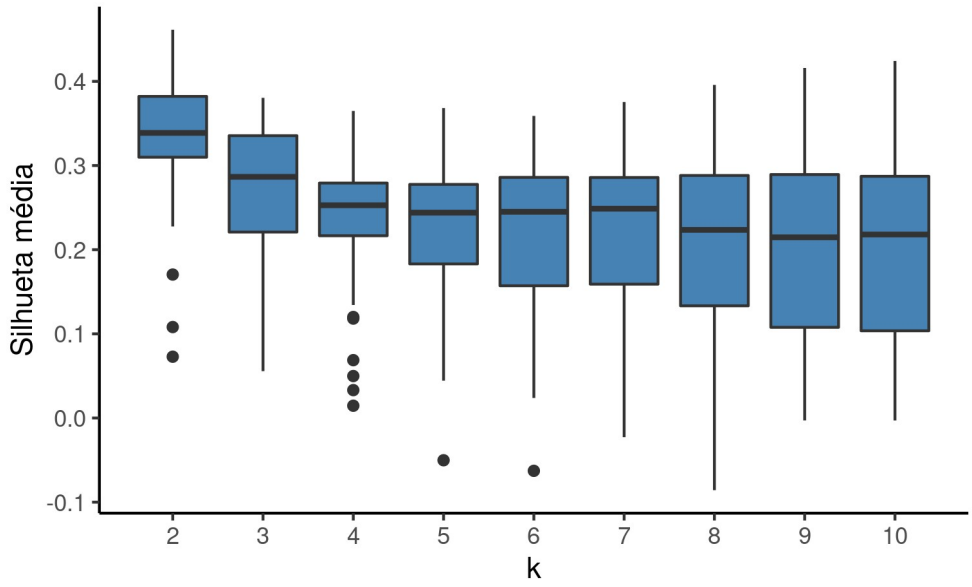

Figura 3 *Silhuetas médias de clusters hierárquicos para diferentes valores de k.*

No geral, as silhuetas médias observadas variaram entre -0.0857 e 0.461. A Tabela 1 mostra os dez maiores valores de silhueta, e as combinações de hiperparâmetros que os geraram. A Figura 4 mostra todas as 378 combinações de hiperparâmetros avaliadas. Foi necessário acrescentar um pequeno jitter horizontal para permitir a visualização de pontos sobrepostos, e a paleta de cores foi organizada para desenfatizar os métodos de ligação que não serão recomendados na aplicação dos resultados desta análise (centroide e mediana). A listagem completa de valores de silhueta média podem ser consultados neste link.

Tabela 1 *Combinações de hiperparâmetros com os dez maiores valores de silhueta média.*

	k	Métrica (distância)	Método de ligação	Silhueta média
1	2	maximum	single	0.4614
2	2	maximum	median	0.4614
3	2	maximum	centroid	0.4614
4	10	minkowski_0.5	ward.D	0.4243
5	10	minkowski_0.5	ward.D2	0.4243
6	2	minkowski_0.5	single	0.4178
7	9	minkowski_0.5	ward.D	0.4159
8	9	minkowski_0.5	ward.D2	0.4159
9	2	euclidian	centroid	0.4107
10	8	minkowski_0.5	ward.D	0.3958
11	8	minkowski_0.5	ward.D2	0.3958

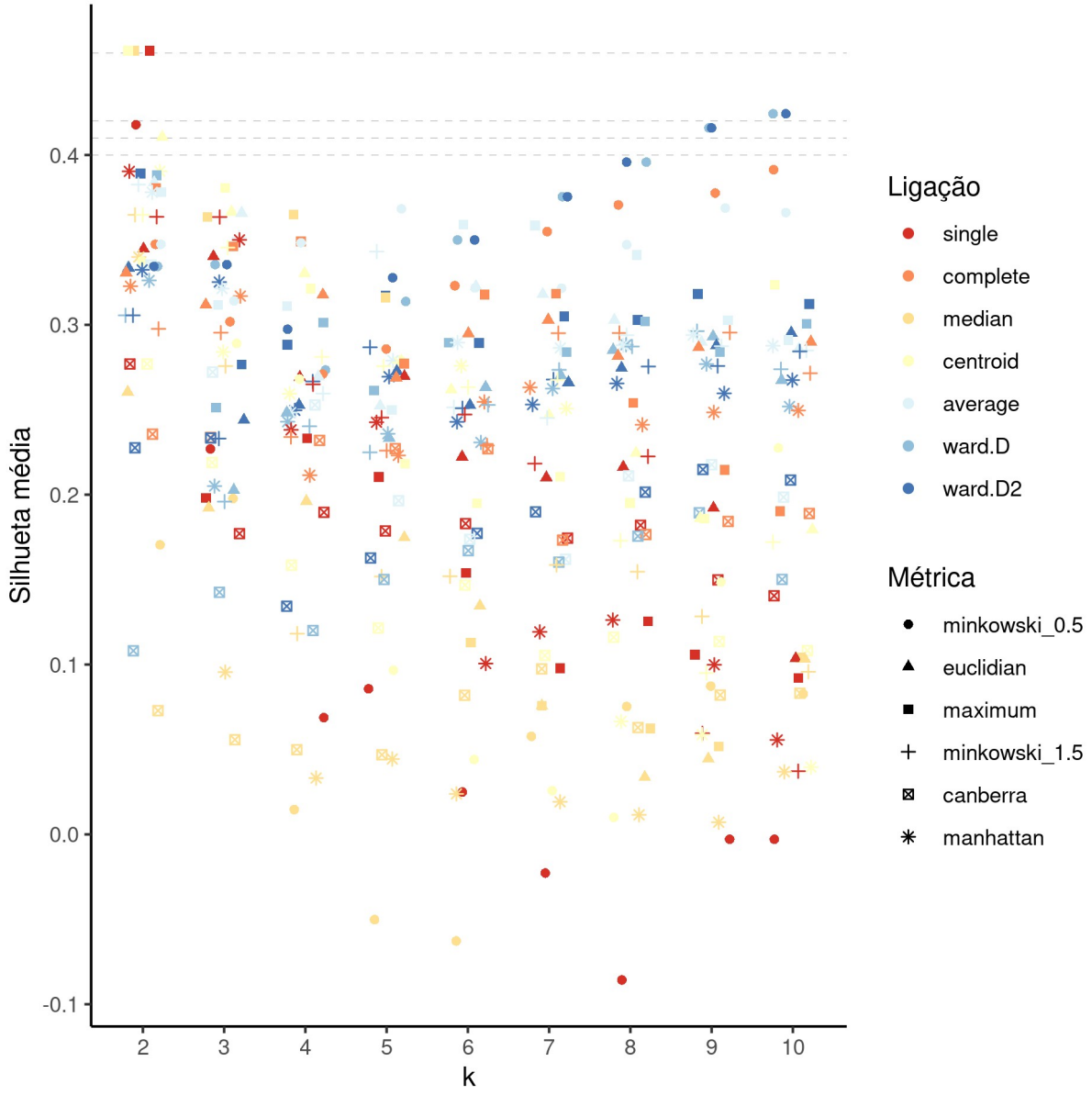

Figura 4 *Silhuetas médias das combinações de hiperparâmetros avaliadas. Linhas tracejadas indicam os dez maiores valores de silhueta média observados.*

4.2 Silhuetas das soluções ótimas

De acordo com os critérios de exclusão especificados no plano analítico (**SAP-2021-017-JG-v01**) os métodos de ligação da mediana e do centroide não são recomendados pois podem gerar inversões nas clusterizações hierárquicas.

O maior valor de silhueta foi observado com três conjuntos de hiperparâmetros. Duas destas combinações incluem métodos excluídos da avaliação (mediana e centroide). O terceiro conjunto é composto pela métrica “maximum”, o método de ligação “single” e k=2. Esta combinação de hiperparâmetros gera uma clusterização hierárquica com silhueta média 0.4614 (Figura 5). As silhuetas médias de cada cluster são, respectivamente 0.47 e 0. Este conjunto de hiperparâmetros gerou um cluster com um deputado, e como este resultado não atende ao objetivo estabelecido, esta combinação de hiperparâmetros foi rejeitada.

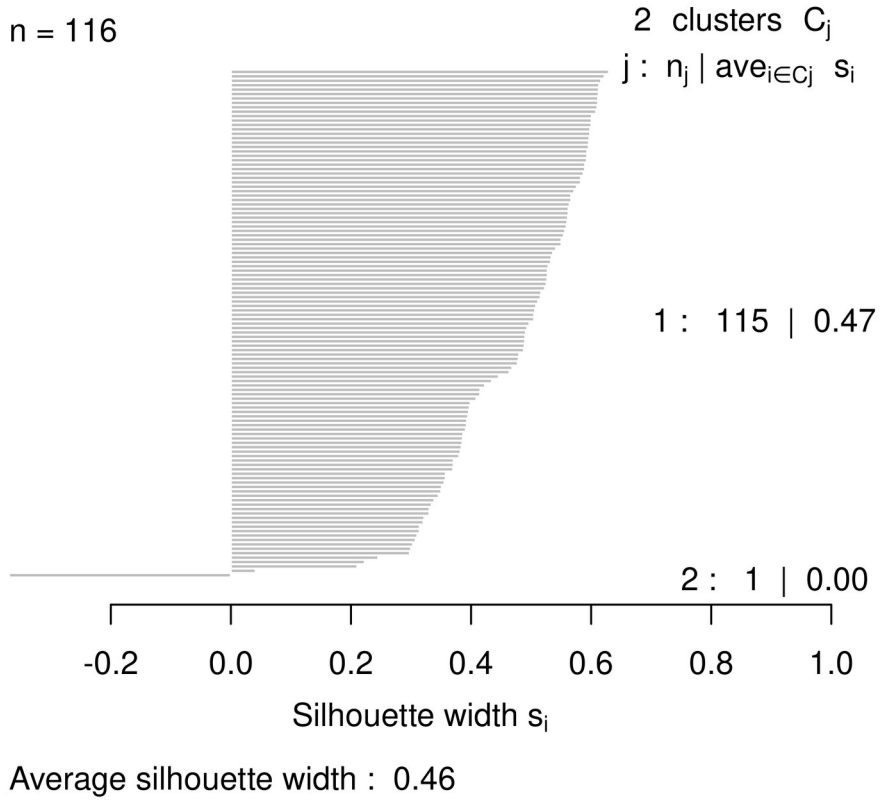

Figura 5 *Silhuetas do cluster hierárquico gerado com métrica “maximum”, o método de ligação “single” e k=2.*

O segundo maior valor de silhueta média (0.4243) foi observado com dois conjuntos de hiperparâmetros. Ambas as combinações usaram a métrica de Minkowski com p=0.5, e k=10, diferindo apenas pelo método de ligação: um com ward.D e a outra com ward.D2. Optando assim pelo método de ligação ward.D2 (ver seção Observações) vemos as silhuetas da clusterização hierárquica gerada com este conjunto de hiperparâmetros na

figura 6. Os tamanhos dos clusters gerados variaram entre 3 e 33 observações. As silhuetas médias dos clusters variaram entre 0.1526 e 0.7376. As silhuetas das observações individuais variaram entre -0.2507 e 0.8139.

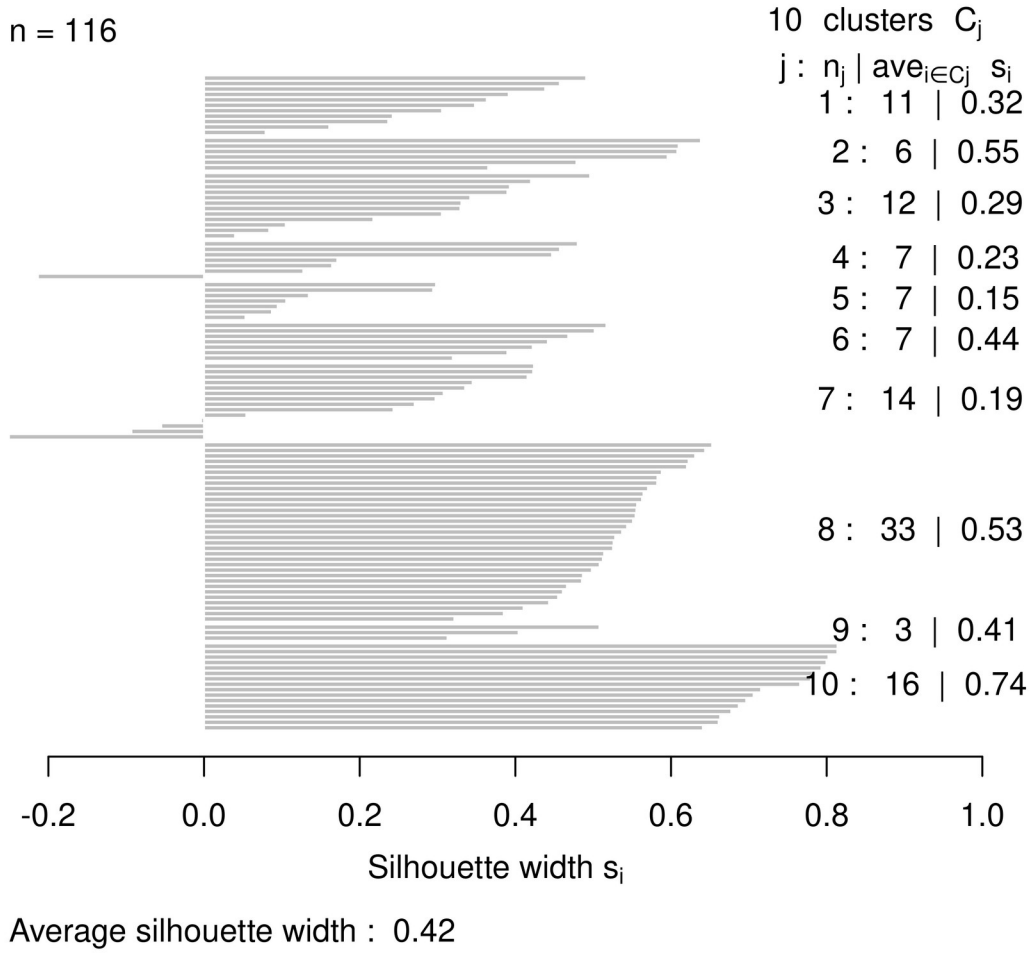

Figura 6 *Silhuetas do cluster hierárquico gerado com métrica de Minkowski com p=0.5, o método de ligação de Ward e k=10.*

5 OBSERVAÇÕES E LIMITAÇÕES

O método de ligação ward.D era usado historicamente em versões anteriores da linguagem R e a documentação indica que o método ward.D2 é sugerido na literatura específica, além de ser a única opção em outra implementação (na função agnes do

pacote `cluster`). Por estes motivos, em caso de empate entre as duas versões do método de Ward, foi recomendado aqui o uso do método de ligação de Ward `ward.D2`.

6 CONCLUSÕES

O maior valor de silhueta média foi observado com o conjunto de hiperparâmetros composto pela métrica "maximum", o método de ligação "single" e k=2. Este conjunto de hiperparâmetros gerou um cluster com um deputado, e como este resultado não atende ao objetivo estabelecido, esta combinação de hiperparâmetros foi rejeitada.

O segundo maior valor de silhueta média válido foi observado com a métrica de Minkowski com p=0.5, o método de ligação de Ward e k=10. A silhueta média deste conjunto de hiperparâmetros foi 0.4243.

7 REFERÊNCIAS

- **SAP-2021-017-JG-v01** – Plano Analítico para Otimização de hiperparâmetros de clusterização hierárquica para identificação de deputados evangélicos de corporações pentecostais eleitos em 2018
- **SAR-2021-011-JG-v01** – Clusterização hierárquica para determinação do número ótimo de clusters de deputados federais evangélicos eleitos em 2018

8 APÊNDICE

8.1 Disponibilidade

Tanto este documento como o plano analítico correspondente (**SAP-2021-017-JG-v01**) podem ser obtidos no seguinte endereço:

https://philsf-biostat.github.io/SAR-2021-017-JG/

8.2 Dados utilizados

Os dados utilizados neste relatório não podem ser publicados online por questões de sigilo.

Tabela A1 *Estrutura da tabela de dados analíticos*

id	corp_pentecostal	receita_agp	receita_outras	num_votos	capilaridade	posicao	decil_filiados	decil_deputados
1								
2								
3								

...								
116								