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

DOCUMENTO: SAP-2021-017-JG-v01

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

2021-11-16

**Sumário**

[1 Lista de abreviaturas 2](#__RefHeading___Toc3157_1439423365)

[2 Contexto 2](#__RefHeading___Toc3159_1439423365)

[2.1 Objetivos 2](#__RefHeading___Toc3161_1439423365)

[2.2 Hipóteses 2](#__RefHeading___Toc3163_1439423365)

[3 Dados 2](#__RefHeading___Toc3165_1439423365)

[3.1 Dados brutos 2](#__RefHeading___Toc3167_1439423365)

[3.2 Tabela de dados analíticos 3](#__RefHeading___Toc3169_1439423365)

[4 Variáveis do estudo 3](#__RefHeading___Toc3171_1439423365)

[4.1 Desfechos primário e secundários 3](#__RefHeading___Toc3173_1439423365)

[4.2 Covariáveis 4](#__RefHeading___Toc3175_1439423365)

[5 Métodos estatísticos 4](#__RefHeading___Toc3177_1439423365)

[5.1 Análises estatísticas 4](#__RefHeading___Toc3179_1439423365)

[5.1.1 Análise descritiva 4](#__RefHeading___Toc3181_1439423365)

[5.1.2 Análise inferencial 4](#__RefHeading___Toc3183_1439423365)

[5.1.3 Modelagem estatística 4](#__RefHeading___Toc3185_1439423365)

[5.2 Significância e Intervalos de Confiança 5](#__RefHeading___Toc3187_1439423365)

[5.3 Tamanho da amostra e Poder 5](#__RefHeading___Toc3189_1439423365)

[5.4 Softwares utilizados 5](#__RefHeading___Toc3191_1439423365)

[6 Observações e limitações 5](#__RefHeading___Toc3193_1439423365)

[7 Referências 5](#__RefHeading___Toc3195_1439423365)

[8 Apêndice 5](#__RefHeading___Toc3197_1439423365)

[8.1 Disponibilidade 5](#__RefHeading___Toc3199_1439423365)

Plano Analítico para 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 |

# Lista de abreviaturas

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

# Contexto

Dados de deputados federais evangélicos, eleitos em 2018, com labels identificando quais pertencem a coporações 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.

## Objetivos

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

## Hipóteses

O agrupamento hierárquico pode ser usado para identificar os deputados evangélicos que pertencem a uma corporação pentecostal.

# Dados

## Dados brutos

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

## Tabela de dados analíticos

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

Todas as variáveis numéricas serão 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 serão escalonadas por milhão de reais. O número de votos séra 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 serão mantidos na escala original.

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.

Depois dos procedimentos de limpeza e seleção 9 variáveis foram incluídas na análise com 116 observações. A Tabela 1 mostra a estrutura dos dados analíticos.

**Tabela 1** Estrutura da tabela de dados analíticos após seleção e limpeza das variáveis.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **corp\_pentecostal** | **receita\_agp** | **receita\_outras** | **num\_votos** | **capilaridade** | **posicao** | **decil\_filiados** | **decil\_deputados** |
| 1 |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| 116 |  |  |  |  |  |  |  |  |

A tabela de dados analíticos serão disponibilizados na versão privada do relatório, e serão omitidas da versão pública do relatório.

# Variáveis do estudo

## Desfechos primário e secundários

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.

## Covariáveis

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

# Métodos estatísticos

## Análises estatísticas

### Análise descritiva

N/A.

### Análise inferencial

N/A.

### Modelagem estatística

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

Os índices de silhueta média serão visualizados em um gráfico de dispersão (silhueta por k). Cada ponto será 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 será 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 \*
  + Centróide \*

\* Os métodos de ligação por mediana e centróide podem gerar inversões na clusterização, e não serão selecionados na otimização. Sua inclusão no algoritmo de otimização terá apenas fins informativos.

## Significância e Intervalos de Confiança

N/A.

## Tamanho da amostra e Poder

N/A.

## Softwares utilizados

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

# Observações e limitações

N/A.

# Referências

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

# Apêndice

## Disponibilidade

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

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