Plano Analítico para Clusterização hierárquica para determinação do número ótimo de clusters para classificação de deputados federais como bancada evangélica

DOCUMENTO: SAP-2021-011-JG-v01

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

Data: 2021-10-12

SUMÁRIO

1 LISTA DE ABREVIATURAS 2

2 INTRODUÇÃO 2

2.1 Contexto 2

2.2 Objetivos 2

2.3 Hipóteses 2

3 DADOS 3

3.1 Dados brutos 3

3.2 Tabela de dados analíticos 3

4 VARIÁVEIS DO ESTUDO 3

4.1 Desfechos primário e secundários 3

4.2 Covariáveis 4

5 MÉTODOS ESTATÍSTICOS 4

5.1 Análises estatísticas 4

5.1.1 Análise descritiva 4

5.1.2 Análise inferencial 4

5.1.3 Modelagem estatística 4

5.2 Significância e Intervalos de Confiança 4

5.3 Tamanho da amostra e Poder 4

5.4 Softwares utilizados 4

6 EXCEÇÕES E OBSERVAÇÕES 5

7 REFERÊNCIAS 5

8 APÊNDICE 5

8.1 Disponibilidade 5

8.2 Análise exploratória de dados 6

Histórico do documento

Versão	Alterações
01	Versão inicial

1 LISTA DE ABREVIATURAS

- DP: Desvio padrão

2 INTRODUÇÃO

2.1 Contexto

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 kmeans.

2.2 Objetivos

1. Avaliar número ideal de clusters de acordo em um dendograma de clusterização hierárquica, para uso do kmeans.
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.

2.3 Hipóteses

Deputados federais da bancada evangélica que foram eleitos em 2018 podem ser identificados com base nas doações recebidas durante a campanha eleitoral, número de votos recebidos e outras características.

3 DADOS

3.1 Dados brutos

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

3.2 Tabela de dados analíticos

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 12 variáveis foram incluídas na análise com 514 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	partido	uf	capilaridade	primeira	sexo	evangelico	num_votos	posicao	decil_filiados	decil_deputados	total_receita
1											
2											
3											
…											
514											

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

4 VARIÁVEIS DO ESTUDO

4.1 Desfechos primário e secundários

O desfecho primário está definido como a classificação entre deputados da bancada evangélica e outros deputados explicada pela receita total recebida.

4.2 Covariáveis

As seguintes características dos deputados federais serão 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 serão 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 serão consideradas na análise, devido à baixa representatividade de valores em suas distribuições (figura A1).

5 MÉTODOS ESTATÍSTICOS

5.1 Análises estatísticas

5.1.1 Análise descritiva

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

5.1.2 Análise inferencial

Não serão realizadas análises inferenciais.

5.1.3 Modelagem estatística

Será ajustado um modelo de clusters hierárquico aos dados numéricos. O dendograma associado ao modelo de agrupamento será cortado em diferentes alturas e números de clusters para obter o número ótimo de clusters que melhor explique a classificação do desfecho primário.

5.2 Significância e Intervalos de Confiança

Todas as análises serão realizadas ao nível de significância de 5%. Todos os testes de hipóteses e intervalos de confiança calculados serão bicaudais.

5.3 Tamanho da amostra e Poder

N/A

5.4 Softwares utilizados

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

6 EXCEÇÕES E OBSERVAÇÕES

7 REFERÊNCIAS

- **SAR-2021-011-JG-v01** – Clusterização hierárquica para determinação do número ótimo de clusters para classificação de deputados federais como bancada evangélica

8 APÊNDICE

8.1 Disponibilidade

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

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

8.2 Análise exploratória de dados

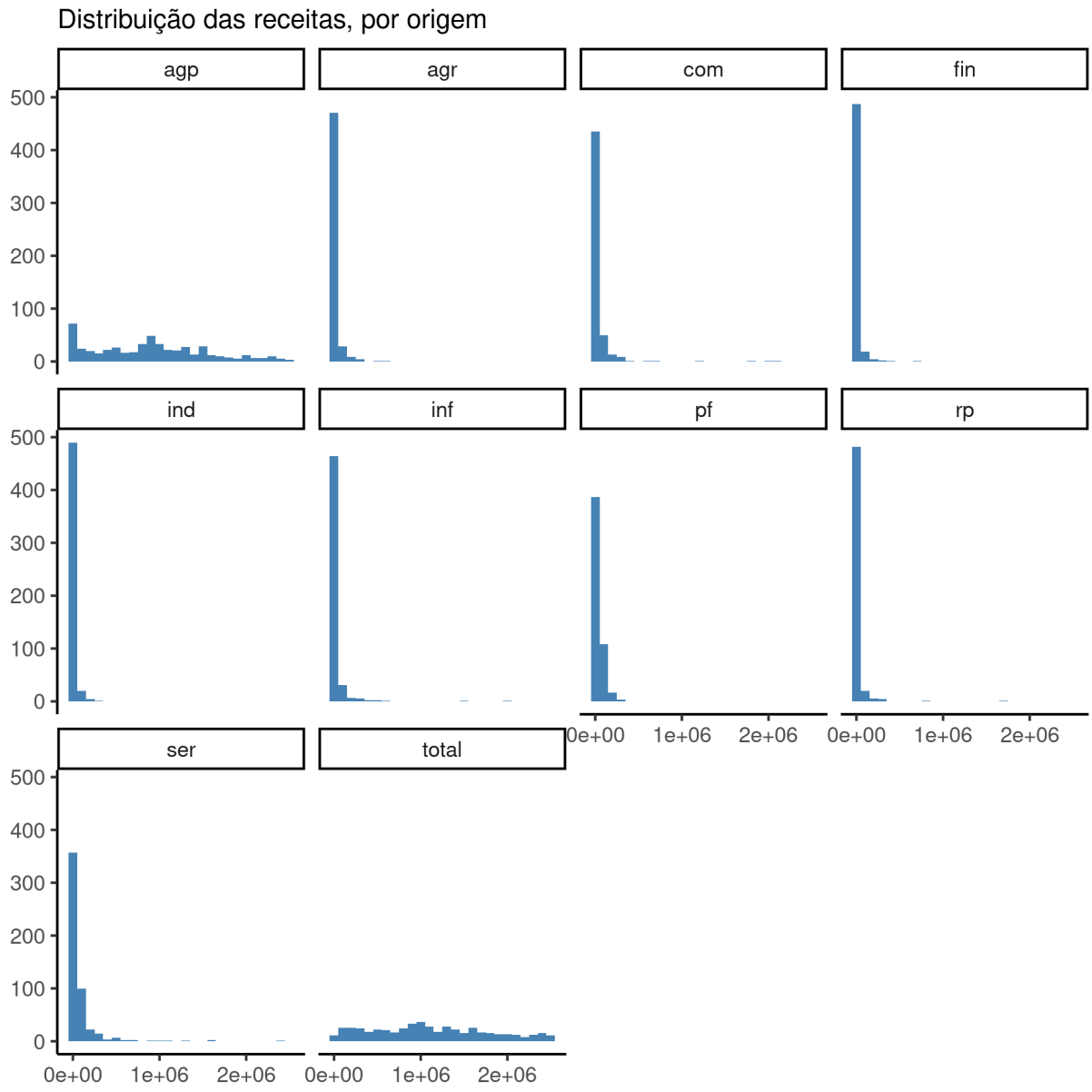

Figura A1 *Distribuição das receitas de deputados federais, por origem (agp = receita que veio do Partido ao invés de apoiadores privados (empresariais ou não); agr = setor agrícola; com = setor do comércio; fin = setor específico dos bancos e outras instituições financeiras e imobiliárias; ind = setores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