FACOM32701 - Ciência de Dados II

Projeto 6: Análise de Agrupamentos de Desmatamento e Uso do Solo

Grupo: Gustavo Antunes de Souza - 12311BSI276

Millena Lins dos Santos - 12311BSI292

RELATÓRIO TÉCNICO

1. Introdução

O desmatamento é um dos principais desafios ambientais do Brasil, impactando biodiversidade, clima e uso do solo.

Este projeto aplica técnicas de aprendizado não supervisionado para identificar padrões estruturais de desmatamento e uso do solo nos municípios brasileiros, buscando responder:

- Existem perfis distintos de municípios quanto à dinâmica de desmatamento?
- Como o desmatamento se relaciona com a estrutura atual de uso da terra?
- É possível identificar padrões de maior pressão ou estabilização?

2. Descrição dos Dados

Foram utilizadas três bases públicas nacionais:

2.1. PRODES - INPE

- Série histórica de desmatamento anual por município, representando a dinâmica temporal deste.
- Foram extraídas métricas agregadas:
  - Desmatamento total acumulado;
  - Média anual;
  - Desvio padrão;
  - Tendência temporal (coeficiente angular regressão linear);
  - Desmatamento nos últimos 3 anos;
  - Número de anos sem desmatamento.

2.2. MapBiomas - Uso do Solo (Coleção 10)

- Área por classe de uso do solo, representando a estrutura territorial atual.
- Foram agregadas classes em:
  - Fração Natural;
  - Fração Antrópica;
  - (*Farming* removida por colinearidade).

2.3. IBGE - Bioma Predominante

- Padronização espacial via código IBGE;
- Interpretação ambiental dos clusters.

3. Metodologia

3.1. Pré-processamento

- Remoção de valores infinitos;
- Imputação por mediana;
- Padronização para evitar dominância por escala (StandardScaler);
- Remoção de colinearidade da variável *Farming* (frações somam 1).

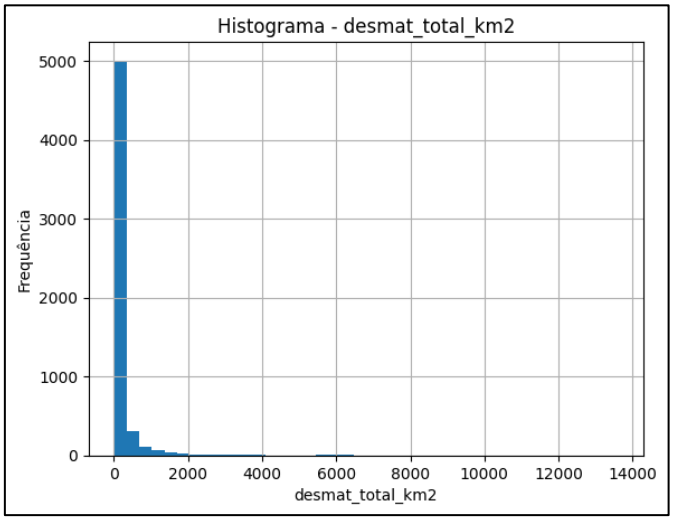
A série histórica do PRODES foi considerada até o ano de 2025, conforme disponibilidade na base TerraBrasilis/INPE. As variáveis de uso e cobertura do solo derivadas do MapBiomas foram consideradas até 2024, último ano consolidado disponível. Essa pequena defasagem temporal não compromete a análise estrutural, uma vez que o objetivo é caracterizar regimes territoriais de médio e longo prazo.

3.2. Análise Exploratória (EDA)

- Estatísticas descritivas;
- Histogramas;
- Boxplots;
- Matriz de correlação (Spearman).

A análise exploratória revelou uma alta assimetria nas variáveis de desmatamento total, além de forte correlação entre total acumulado, média anual e intensidade recente, e a relação inversa entre fração Natural e fração Antrópica.

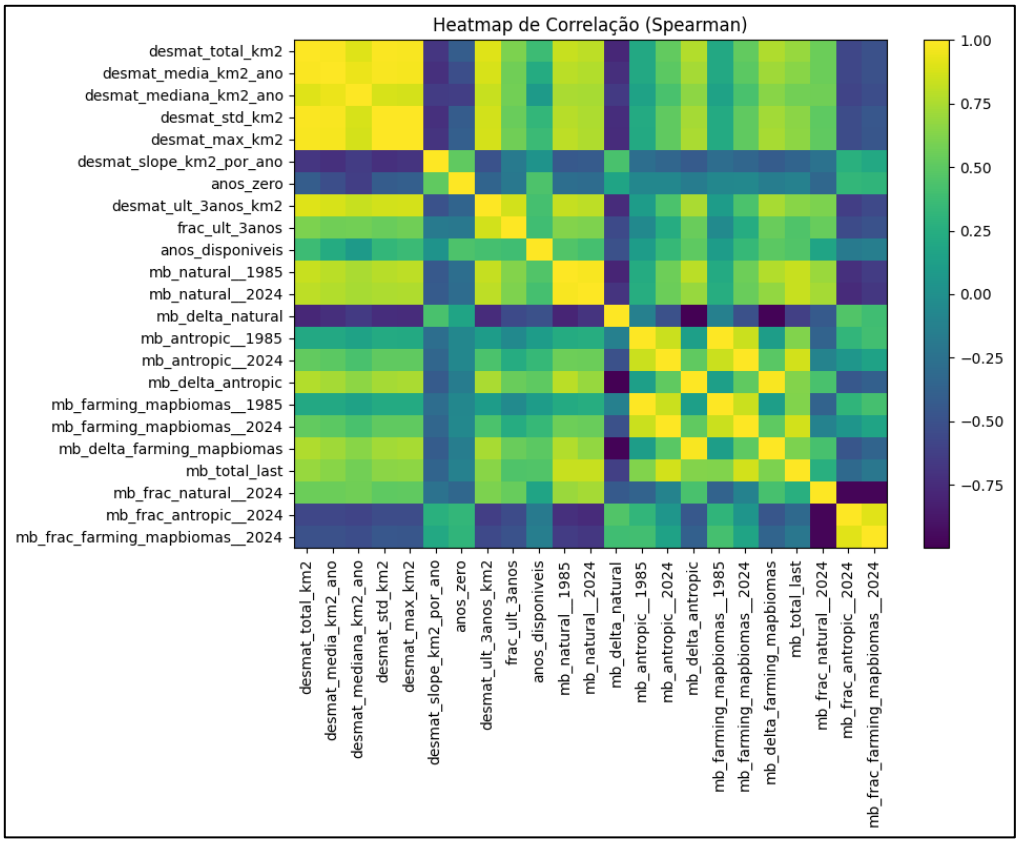
Figura 1. Distribuição das principais variáveis de desmatamento municipal.



Observa-se que a maioria dos municípios apresenta baixos valores de desmatamento, enquanto poucos municípios concentram valores extremamente elevados, justificando a padronização e o uso de técnicas robustas de clusterização.

O heatmap de correlação indicou estrutura não aleatória, sugerindo que técnicas de redução de dimensionalidade poderiam capturar bem a variância estrutural.

Figura 2. Matriz de correlação (Spearman) entre variáveis estruturais de desmatamento e uso do solo.



Observa-se forte correlação entre desmatamento total, média anual e intensidade recente, indicando redundância parcial entre variáveis. Também se confirma a relação inversa entre fração Natural e fração Antrópica, justificando a remoção da variável *Farming* por colinearidade perfeita.

3.3. Redução de Dimensionalidade (PCA)

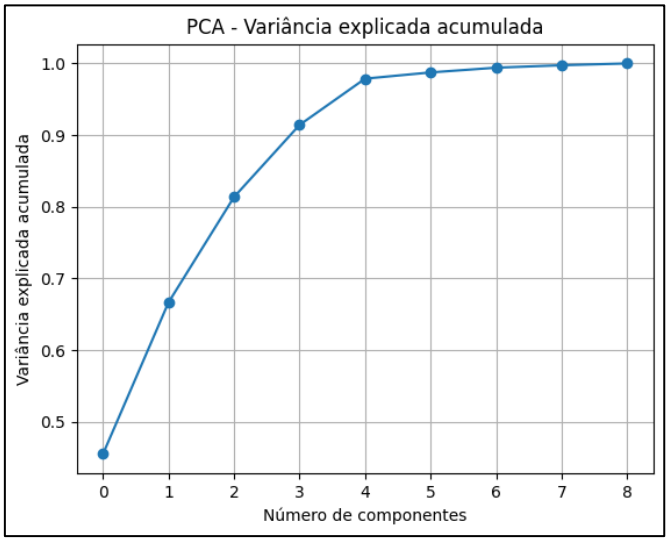
A Análise de Componentes Principais (PCA) foi aplicada às variáveis padronizadas com o objetivo de reduzir a dimensionalidade do conjunto de dados e identificar os principais eixos estruturais associados ao desmatamento e ao uso do solo.

A variância explicada por cada componente foi:

- PC1: ≈ 45,5%
- PC2: ≈ 21,1%
- PC3: ≈ 14,7%
- PC4: ≈ 10,0%

Os dois primeiros componentes explicam aproximadamente 66% da variância total, enquanto os quatro primeiros acumulam cerca de 91%.

Figura 3. Variância explicada acumulada pelos componentes principais.

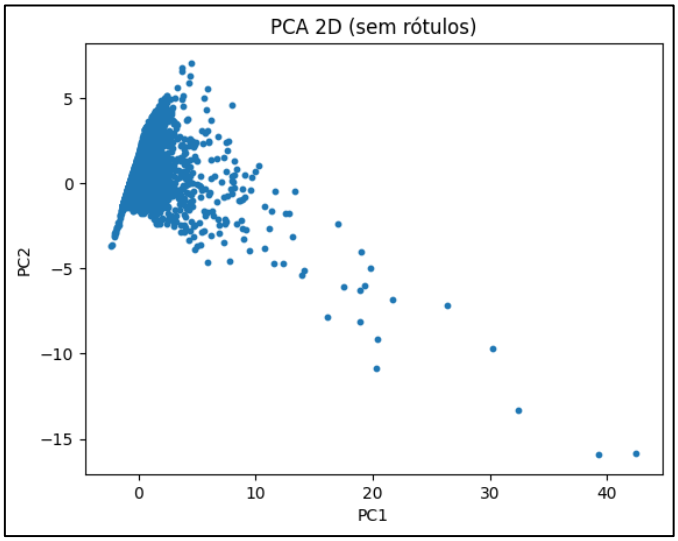


A Figura 3 mostra que há inflexão clara após o terceiro ou quarto componente, indicando que componentes adicionais contribuem marginalmente para a explicação da variabilidade. Isso evidencia que a estrutura do fenômeno analisado pode ser representada adequadamente por poucos fatores latentes.

A análise das cargas indica que PC1 está fortemente associado à magnitude do desmatamento, incluindo: desmatamento total acumulado, média anual, intensidade recente e variabilidade histórica. Esse componente representa um eixo de pressão estrutural de desmatamento.

PC2 está relacionado à dinâmica temporal, capturando: tendência (slope), intensidade recente relativa e número de anos sem desmatamento. Esse eixo distingue municípios com crescimento recente daqueles com estabilização ou redução.

Figura 4. Projeção dos municípios no espaço bidimensional definido por PC1 e PC2.



A Figura 4 apresenta a distribuição dos municípios no plano PC1 × PC2:

- Alta concentração de municípios com baixos valores de PC1 (menor magnitude de desmatamento).
- Presença de observações dispersas ao longo do eixo PC1, indicando municípios com comportamento extremo.
- Estrutura não aleatória, sugerindo viabilidade da aplicação de algoritmos de clusterização.

A aplicação do PCA demonstrou que:

- O fenômeno do desmatamento municipal possui estrutura latente consistente.
- A maior parte da variabilidade pode ser representada por poucos componentes.
- Há organização suficiente nos dados para justificar a aplicação de técnicas de agrupamento.

3.4. Algoritmos de Clusterização

Foram aplicados três métodos de agrupamento:

- K-Means
- Agrupamento Hierárquico (Ward)
- DBSCAN (exploratório)

A seleção do número ótimo de clusters foi baseada em métricas internas:

- Método do cotovelo (SSE)
- Silhouette
- Davies-Bouldin
- Calinski-Harabasz

A solução final adotada foi k = 2, por apresentar melhor desempenho conjunto nas métricas avaliadas (maior silhouette).

3.5. Estabilidade

A estabilidade do K-Means foi avaliada utilizando diferentes seeds [0, 1, 2, 42, 100].

Resultados médios:

- ARI ≈ 0,99
- NMI ≈ 0,98

Esses valores indicam altíssima consistência entre execuções, evidenciando que os agrupamentos não dependem da inicialização do algoritmo.

3.6. Agrupamento Hierárquico (Ward)

O método de Ward foi aplicado como abordagem complementar. Devido ao grande número de observações, utilizou-se dendrograma truncado para visualização.

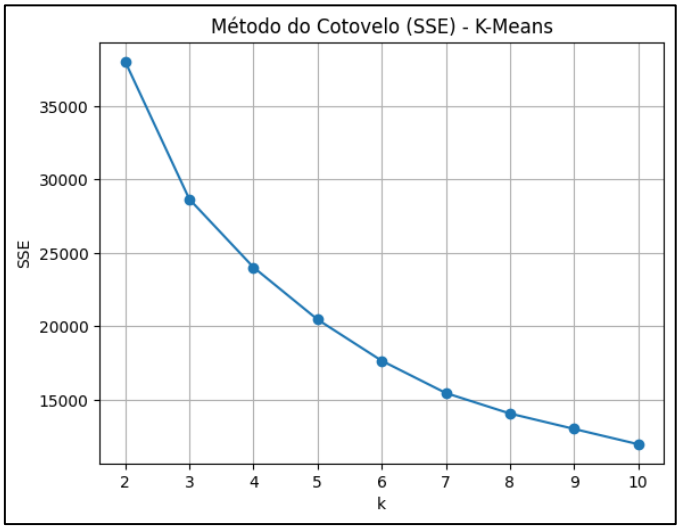
O Silhouette obtido foi:

- Ward ≈ 0,80
- K-Means ≈ 0,79

4. Resultados

4.1. Determinação do Número de Clusters

Figura 5. Método do Cotovelo.



Observa-se que a redução do SSE é acentuada até k = 2. A partir desse ponto, a curva passa a apresentar declínio mais suave, indicando retornos marginais decrescentes na redução da variabilidade interna dos clusters.

Esse comportamento caracteriza o “cotovelo” do gráfico, sugerindo que a estrutura latente dos dados é adequadamente representada por dois grupos principais.

4.2. Validação Interna dos Agrupamentos

Silhouette obtido:

- K-Means: 0,79
- Ward: 0,80

Valores acima de 0,7 indicam separação muito consistente entre grupos.

Isso significa que:

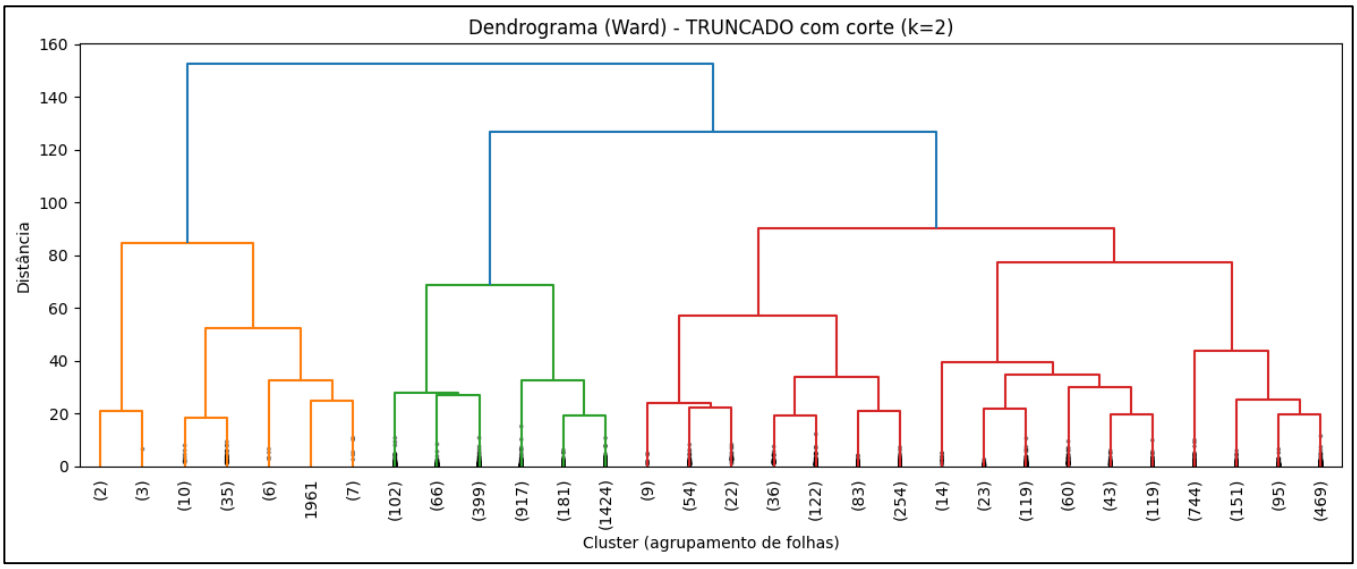
- Municípios dentro do mesmo cluster são similares entre si.

- Municípios de clusters diferentes são estruturalmente distintos.

A convergência entre K-Means e Ward indica que a estrutura identificada é robusta e não dependente do algoritmo utilizado.

4.3. Estrutura Hierárquica

Figura 6. Dendrograma truncado do método hierárquico (Ward), evidenciando a divisão em dois clusters principais.



O dendrograma evidencia uma divisão hierárquica clara entre dois grandes blocos de municípios. Observa-se que a distância de ligação entre esses dois grupos principais é significativamente superior às distâncias internas dentro de cada bloco, indicando uma ruptura estrutural bem definida nos dados.

Internamente, cada cluster apresenta subestruturas menores com distâncias relativamente reduzidas, sugerindo alta similaridade entre municípios pertencentes ao mesmo grupo. Já a grande altura do último nível de fusão confirma que a separação entre os dois clusters principais não é arbitrária, mas decorre de diferenças consistentes nas variáveis analisadas.

Esse comportamento reforça o resultado obtido pelo K-Means (k = 2) e pelo valor elevado do índice Silhouette (≈ 0,80), indicando que a estrutura binária identificada representa um padrão real e robusto de organização territorial dos municípios brasileiros segundo indicadores de desmatamento e uso do solo.

4.4. Avaliação do DBSCAN

O algoritmo DBSCAN foi aplicado de forma exploratória com diferentes valores de ε (0,5, 0,8, 1,0, 1,2) e parâmetro `min_samples = 10`.

Os resultados indicaram elevada sensibilidade ao parâmetro ε, produzindo cenários com fragmentação excessiva (muitos clusters pequenos) ou identificação predominante de ruído.

Além disso, os valores de silhouette (considerando apenas pontos não classificados como ruído) foram inferiores aos obtidos por K-Means e Ward.

Dessa forma, optou-se por não utilizar o DBSCAN como modelo principal, uma vez que os métodos baseados em partição e hieráricos apresentaram estrutura mais estável, interpretável e coerente com a dinâmica ambiental.

4.5. Caracterização dos Clusters

- **Cluster 0 - Alta pressão de desmatamento**
  - Tendência temporal positiva mais intensa;
  - Maior intensidade recente;
  - Menor número de anos com desmatamento zero;
  - Fração Natural ainda relevante.

Representa municípios em processo ativo de conversão territorial, característicos de fronteiras agrícolas e regiões com dinâmica recente de expansão ou contínua de desmatamento.

- **Cluster 1 - Menor pressão ou estabilização**
  - Baixo desmatamento acumulado;
  - Tendência estável ou negativa;
  - Maior número de anos com desmatamento zero;
  - Estrutura territorial já consolidada.

Representa municípios com menor pressão recente, podendo indicar áreas já consolidadas ou com menor dinamismo de conversão.

4.6. Relação com Biomas

Figura 7. Distribuição percentual dos biomas (IBGE) por cluster identificado pelo K-Means.

bioma_ibge	Amazônia	Caatinga	Cerrado	Mata Atlântica	Pampa	Pantanal
cluster_kmeans						
0	65.06	0.00	31.33	0.00	1.2	2.41
1	8.18	19.96	18.88	49.95	2.9	0.13

A distribuição dos biomas evidencia forte associação entre estrutura ambiental e padrão de desmatamento. O Cluster 0, caracterizado por maior pressão de desmatamento, concentra-se majoritariamente na Amazônia (65,06%) e no Cerrado (31,33%), biomas historicamente associados à expansão da fronteira agropecuária.

Já o Cluster 1 apresenta predominância da Mata Atlântica (49,95%) e maior presença relativa na Caatinga e Cerrado, refletindo municípios com estrutura territorial mais consolidada e menor dinâmica recente de conversão.

Essa associação reforça a coerência ecológica dos agrupamentos e indica que os clusters capturam diferenças regionais estruturais no regime de uso e cobertura do solo.

5. Discussão

Os clusters revelam dois regimes territoriais distintos:

1. Municípios com fronteira ativa de conversão territorial.
2. Municípios com território consolidado ou estabilizado.

A forte associação entre bioma e cluster indica que a dinâmica de desmatamento possui componente estrutural regional.

A remoção da variável *Farming* foi metodologicamente correta devido à colinearidade perfeita (frações somam 1).

A alta variância explicada pelo PCA e os altos valores de Silhouette indicam que a estrutura identificada não é aleatória.

Além disso:

- A consistência entre K-Means e Ward reforça robustez.
- A associação com biomas demonstra validade ambiental.
- A engenharia de atributos temporais foi fundamental para capturar a dinâmica histórica.

6. Limitações

- PRODES não cobre todos os biomas de forma homogênea.
- Agregação municipal mascara heterogeneidade intramunicipal.
- Ausência de variáveis socioeconômicas.
- Uso apenas de métricas internas (ausência de validação externa).

7. Conclusão

A aplicação de técnicas não supervisionadas permitiu identificar padrões claros e estáveis de desmatamento municipal no Brasil. Foram identificados dois regimes territoriais distintos:

a. Municípios com dinâmica ativa de expansão territorial.
b. Municípios com consolidação ou baixa pressão recente.

Principais achados:

- Estrutura de dados altamente organizada (PCA explica ≈80% com três componentes).
- Dois regimes ambientais bem definidos.
- Alta estabilidade dos agrupamentos (ARI > 0,98).
- Boa separação estrutural (Silhouette ≈ 0,80).

Os resultados demonstram o potencial de técnicas de clusterização para caracterização de regimes territoriais em larga escala e indicam aplicabilidade prática no apoio a políticas públicas ambientais.