

## **Modelos de agrupamento aplicados aos resultados setoriais do índice de efetividade da gestão municipal**

Marcos Dionísio Ferreira<sup>1</sup>; Regina Ávila Santos<sup>2</sup>

<sup>1</sup> Governo do estado de São Paulo. Secretaria de comunicação. Av. Morumbi, 4.500 – Morumbi; 05650-905  
São Paulo, SP, Brasil

<sup>2</sup> Universidade de São Paulo. Doutorado em Economia Aplicada. Av. Pádua Dias, 235 – Agronomia; 13418-900

\*autor correspondente: marcosdf\_84@hotmail.com

## **Modelos de agrupamento aplicados aos resultados setoriais do índice de efetividade da gestão municipal**

### **Resumo**

O estudo abordou a importância da análise dos resultados do Índice de Efetividade da Gestão Municipal [IEG-M] no estado de São Paulo. O IEG-M foi introduzido para avaliar a eficiência das prefeituras em sete áreas administrativas: saúde, planejamento, educação, gestão fiscal, proteção aos cidadãos, meio ambiente e governança em tecnologia da informação. No entanto, apesar de ser um índice reconhecido e premiado, seu cálculo possui uma ponderação enviesada pelo orçamento público, o que limita a análise dos dados e subvaloriza setores essenciais. Este trabalho propôs uma abordagem alternativa, utilizando técnicas não supervisionadas de aprendizado de máquina para agrupar as prefeituras de maneira representativa e homogênea em relação aos resultados dos índices setoriais, sem considerar o peso orçamentário. Isso proporcionou uma facilidade maior na análise dos perfis das prefeituras e ampliou as possibilidades de análise, permitindo a formulação de ações mais amplas, direcionadas e assertivas por parte dos entes públicos. Foram testados e comparados os resultados de modelos hierárquicos aglomerativos e não hierárquicos para agrupamento de dados. Neste estudo, os modelos foram avaliados a partir da representatividade e qualidade dos grupos formados, e o modelo não hierárquico K-Means apresentou os melhores resultados, tornando-se assim a melhor opção para agrupar as prefeituras.

**Palavras-chave:** clustering; modelos hierárquicos aglomerativo; modelos não-hierárquicos; K-Means.

### **Clustering models applied to sectoral results of the Municipal Management Effectiveness Index**

### **Abstract**

The study addressed the importance of analyzing the results of the Municipal Management Effectiveness Index [IEG-M] in the state of São Paulo. The IEG-M was introduced to assess the efficiency of municipalities in seven administrative areas: healthcare, planning, education, fiscal management, citizen protection, environment, and information technology governance. However, despite being a recognized and awarded index, its calculation is biased by the public budget, limiting data analysis and undervaluing essential sectors. This work proposed an alternative approach, using unsupervised machine learning techniques to group municipalities in a representative and homogeneous manner in relation to sectoral index results, without considering budgetary weight. This provided greater ease in analyzing municipal profiles and expanded analysis possibilities, allowing for the formulation of broader, targeted, and more assertive actions by public entities. The results of agglomerative hierarchical and non-hierarchical models for data clustering were tested and compared in this study. The models were evaluated based on the representativeness and quality of the formed groups, and the non-hierarchical K-Means model yielded the best results, thus becoming the best option for grouping municipalities.

**Keywords:** clustering; agglomerative hierarchical models; non-hierarchical models; K-Means.

## **Introdução**

No cenário brasileiro, São Paulo emerge como uma das unidades da federação mais densamente povoadas, abrigando uma população que se aproxima dos 45 milhões de habitantes, distribuídos em uma estrutura geopolítica composta por 645 municípios. Além disso, o estado desempenha um papel crucial na economia nacional, contribuindo com mais de 31% do Produto Interno Bruto [PIB] brasileiro (IBGE, 2023). Essa importância não apenas em termos de população e riqueza financeira, mas também em sua influência geopolítica, gera desafios substanciais relacionados à organização, administração e fiscalização da máquina estatal. Este cenário complexo engendra uma rede intrincada de relações entre os poderes de estado, constituindo um formidável desafio para qualquer entidade política que busque enfrentá-lo.

O Tribunal de Contas do Estado de São Paulo [TCE-SP], órgão encarregado pela supervisão das finanças, contabilidade, orçamento, operações e patrimônio do estado de São Paulo e de seus Municípios, com exceção da Capital, tem direcionado seus esforços, por meio do seu plano estratégico, para estabelecer e implementar mecanismos que colaborem na fiscalização e orientação das prefeituras. Com uma abordagem preventiva e corretiva, avaliando atos e resultados, o TCE-SP busca assegurar que os recursos públicos sejam empregados de modo adequado e transparente, em prol da sociedade. Essas iniciativas visam garantir a eficiência e a transparência na administração dos recursos, contribuindo para o bem-estar e desenvolvimento de toda a comunidade (TCESP, 2022).

Uma iniciativa implementada em 2015 foi a introdução do Índice de Efetividade da Gestão Municipal [IEG-M], cujo propósito foi avaliar a eficiência das 644 prefeituras do estado de São Paulo, excluindo a capital. Esse índice concentra-se na análise da infraestrutura e dos processos, bem como na avaliação da eficácia das políticas públicas em sete áreas administrativas: saúde, planejamento, educação, gestão fiscal, proteção aos cidadãos, meio ambiente e governança em tecnologia da informação. Através desse mecanismo, o TCE-SP busca promover melhorias contínuas na administração pública municipal, identificando pontos fortes e áreas que necessitam de maior fiscalização ou aprimoramento de serviços prestados à população. Além disso, ele disponibiliza informações que têm sido utilizadas por prefeitos e vereadores para ajustar suas estratégias, reavaliar prioridades e fortalecer o planejamento dos municípios, dado que oferece uma avaliação de setores de extrema importância para o município (TCESP, 2022). Uma prova da relevância do IEG-M foi o recebimento de uma menção honrosa no Prêmio Inovare de 2018, na categoria Tribunais, onde foi um dos dois finalistas escolhidos entre 89 práticas inscritas por tribunais de vários estados.

O IEG-M é construído a partir de uma média ponderada dos índices setoriais, os quais, por sua vez, são obtidos por meio de um questionário respondido pelas administrações das prefeituras, abrangendo todos os sete setores de avaliação. A ponderação do índice é realizada com base em critérios de importância orçamentária, ou seja, os setores com maior peso no orçamento público recebem uma participação maior na composição do índice. Esse tipo de composição, embora eficaz para fins orçamentários, introduz um viés na interpretação dos dados, uma vez que subestima setores que, apesar de possuírem menor importância orçamentária, são de extrema relevância para a população em geral, como é o caso dos setores relacionados ao meio ambiente e à proteção dos cidadãos, ambos com participação reduzida no IEG-M.

E é neste ponto, considerando a relevância e a abrangência do IEG-M no contexto do estado de São Paulo, bem como a necessidade de iniciativas que simplifiquem a organização e a administração, dada a complexidade inerente à gestão de um estado como São Paulo, este estudo propõe uma análise alternativa dos resultados dos índices setoriais do IEG-M. Para isso, sugere o agrupamento das prefeituras de uma forma representativa e homogênea em relação aos resultados dos índices, partindo de uma classificação sem o viés orçamentário. O objetivo é agrupar prefeituras com perfis semelhantes, ampliando assim as possibilidades de análise e permitindo a formulação de ações mais abrangentes, assertivas e direcionadas, promovendo uma visão mais completa e equilibrada dos índices setoriais do IEG-M.

Para esta análise, propõe-se a utilização de técnicas não supervisionadas de aprendizado de máquina para realização dos agrupamentos, e a comparação dos resultados obtidos em cada técnica, incluindo a comparação com os resultados do próprio IEG-M.

Conforme Fávero e Belfiore (2017) afirmam, a análise de agrupamentos consiste em um conjunto de técnicas exploratórias utilizadas com o propósito de verificar a existência de comportamentos semelhantes entre observações, que podem ser indivíduos, empresas, municípios, países ou outros exemplos, em relação a determinadas variáveis. O objetivo é criar grupos nos quais prevaleça a homogeneidade interna, alinhando-se assim com o propósito deste trabalho.

Essa estratégia tem como objetivo facilitar a análise dos perfis das prefeituras e proporcionar uma abordagem mais específica para a formulação de ações e políticas públicas. Com base nesses resultados, o governo de São Paulo, por exemplo, poderia moldar políticas públicas de acordo com as características de cada grupo, concentrando-se em setores críticos para cada grupo. Da mesma forma, o TCE-SP poderia criar métodos de fiscalização específicos para um grupo que obtenha, por exemplo, pontuações elevadas nos setores ambientais e de saúde, mas notas baixas na gestão fiscal, tornando a ação mais assertiva.

## Material e Métodos

### IEG-M

O IEG-M é um índice que avalia a eficiência das políticas públicas nas 644 prefeituras do estado de São Paulo, com exceção da capital, abrangendo sete setores da administração pública: Educação, Saúde, Planejamento, Gestão Fiscal, Meio Ambiente, Proteção dos Cidadãos/Defesa Civil e Governança em Tecnologia.

A avaliação é conduzida por meio de um questionário desenvolvido pelo TCE-SP e respondido pelas administrações das prefeituras. Posteriormente, uma equipe de fiscalizadores do TCE-SP valida os resultados por amostragem. Este questionário abrange cada um dos setores e é composto por uma série de perguntas correlacionadas aos Objetivos de Desenvolvimento Sustentável [ODS] elaborados pela Organização das Nações Unidas [ONU]. Esses objetivos representam um plano de ação voltado para o futuro do planeta, com o objetivo de promover a paz e a prosperidade, e são sintetizados em 17 metas, conforme ilustrado na Figura 1.



Figura 1. 17 objetivos de desenvolvimento sustentável no Brasil  
Fonte: ONU (2023)

Cada questionário do IEG-M tem como objetivo estabelecer correlações entre os setores de referência e os 17 ODS. Essa análise possibilita avaliar como as ações das administrações das prefeituras contribuem para a realização dos ODS, proporcionando uma visão holística do desempenho e dos esforços voltados para o desenvolvimento sustentável. A seguir, uma síntese dos objetivos de cada questionário.

- O i-Educ avalia os resultados do setor por meio de critérios relacionados à Educação Infantil e ao Ensino Fundamental, com foco na infraestrutura escolar;
- O i-Saúde mensura os resultados da área por meio de critérios relacionados à Atenção Básica, às Equipes de Saúde da Família, aos Conselhos Municipais de Saúde, bem como tratamentos e vacinação;
- O i-Plan avalia a consistência entre o planejado e o efetivamente implementado, além da coerência entre as metas e os recursos empregados;
- O i-Fiscal analisa os resultados da administração fiscal por meio da análise da execução financeira e orçamentária, assim como o cumprimento da Lei de Responsabilidade Fiscal (LRF);
- O i-Amb avalia os resultados das ações relacionadas ao ecossistema que impactam os serviços e a qualidade de vida dos cidadãos. Isso envolve a análise de dados sobre resíduos sólidos, educação ambiental e estrutura dos conselhos ligados ao setor, entre outros fatores;
- O i-Cidade mede o grau de planejamento de ações relacionadas à segurança dos municípios diante de acidentes e desastres naturais;
- O i-Gov TI avalia o grau de utilização de recursos tecnológicos em áreas como capacitação de pessoal, transparência e segurança da informação.

Com base nos resultados do questionário, o TCE-SP calcula um índice para cada setor, que é então classificado em cinco faixas de resultados, conforme ilustrado na Tabela 1.

Tabela 1. Faixa de resultados IEG-M

Nota	Descrição	Faixa
A	Altamente efetiva	IEG-M 90% da nota máxima e mínimo de 5 índices com nota A
B+	Muito efetiva	IEG-M entre 75% e 89,9% da nota máxima
B	Efetiva	IEG-M entre 60% e 74,9% da nota máxima
C+	Em fase de adequação	IEG-M entre 50% e 59,9% da nota máxima
C	Baixo nível de adequação	IEG-M menor ou igual 49,9%

Fonte: TCESP (2023)

Após a realização dos cálculos dos índices setoriais, o IEG-M é determinado mediante a aplicação de uma ponderação aos resultados de cada setor, conforme pode ser verificado na form. (1):

$$IEGM = \frac{i-plan*20 + i-fiscal*20 + i-educ*20 + i-saúde*20 + i-amb*10 + i-cidade*5 + i-govTI*5}{100} \quad (1)$$

onde, os resultados setoriais são multiplicados pelos seus respectivos pesos.

Essa ponderação é estabelecida com base na relevância de cada setor no contexto do orçamento público. Especificamente, os setores de Educação, Saúde, Planejamento e Fiscal possuem uma ponderação maior, cada um representando 20% do total. A seguir, o Meio Ambiente é ponderado com 10%, enquanto os setores de Proteção dos Cidadãos/Defesa Civil e Governança em Tecnologia possuem uma ponderação de 5% cada.

### **Banco de dados**

Com o propósito fundamental de fornecer informações essenciais para embasar as atividades de fiscalização realizadas pelo controle externo e pela sociedade, o IEG-M teve sua primeira divulgação no ano de 2015 e, desde então, vem sendo publicado de forma anual. No entanto, para a realização deste trabalho, utilizou-se a base de dados do TCE-SP contendo resultados do IEG-M referentes ao ano de 2021, elaborados no ano de 2022. Esses dados são de livre acesso ao público e foram obtidos diretamente através do portal oficial do TCE-SP.

O banco de dados estava estruturado com variáveis qualitativas, onde cada resultado foi apresentado por meio de uma faixa de valores e classificado em cinco categorias: C, C+, B, B+ e A. Para dar início à análise e com o objetivo de evitar qualquer possibilidade de ponderação arbitrária, foram realizados alguns procedimentos preliminares de preparação de dados. O primeiro passo consistiu na transformação do banco de dados em uma matriz binária, para o qual as variáveis foram convertidas em zeros para o não evento e um para o evento. Em seguida, foram aplicados procedimentos de “Data Wrangling” para organizar e estruturar os dados de forma adequada para a análise subsequente.

Todos os procedimentos de manipulação e análise de dados, bem como a geração de gráficos, foram conduzidos fazendo uso da vasta biblioteca de funções estatísticas disponíveis no programa Rproject, na sua versão 4.2.2, lançada em 2022, utilizando-se da linguagem de programação R.

### **Análise de agrupamentos**

Conforme Fávero e Belfiore (2017), a análise de agrupamentos é uma técnica exploratória que identifica comportamentos semelhantes entre observações em relação a variáveis específicas, criando grupos, chamados de clusters, que se destacam pela homogeneidade interna. Isso auxilia na compreensão do comportamento conjunto das observações com base nessas variáveis, facilitando a identificação de padrões, a segmentação de grandes conjuntos de dados, a classificação automática com base em

características e a criação de visualizações. Ampla em suas aplicações, essa ferramenta é essencial para explorar e entender estruturas em dados complexos.

No contexto deste estudo, que destaca a necessidade de criar uma alternativa ao modelo de classificação influenciado pelo orçamento público do IEG-M, e considerando as vantagens apresentadas por Fávero e Belfiore (2017) em relação à análise de agrupamentos, foram testados diversos modelos de agrupamentos com o objetivo de determinar o modelo com maior representatividade e homogeneidade nos grupos formados.

Vale ressaltar dois pontos adicionais. Primeiro, esses modelos não têm caráter preditivo, e os resultados apresentados refletiram o período estudado, exigindo uma nova modelagem para testar outros períodos. Segundo a análise de agrupamentos não exige um conhecimento avançado em estatística, mas foi necessário escolher medidas de distância ou semelhança e esquemas de aglomeração com base nos objetivos da pesquisa, o que será abordado na sequência do estudo.

### **Medida de similaridade**

De relevância central na análise de agrupamentos, a métrica de distância dos dados é frequentemente empregada na análise de dados binários como uma medida de similaridade. Conforme mencionado por Fávero e Belfiore (2017), esse método é utilizado quando o enfoque se direciona à frequência de pares de respostas convergentes, ou seja, 1-1 ou 0-0, entre duas observações específicas. Nesse contexto, quanto maior a frequência de pares convergentes, maior a semelhança entre as observações.

Para o cálculo das medidas de similaridade, uma ampla gama de coeficientes foi proposta por vários autores. Essa diversidade de coeficientes surge em função das incertezas associadas à manipulação das correspondências 0-0 e da importância e pesos dados nas correspondências 1-1, 1-0 ou 0-1.

No intuito de calcular a matriz de distância do banco de dados, foco deste estudo, foram testados três coeficientes de similaridade: o coeficiente de emparelhamento simples, o coeficiente de Jaccard e o coeficiente de DICE. Ambas as métricas se baseiam na construção de uma tabela de frequências absolutas de respostas 0 e 1, como exemplificado na Tabela 2.

Tabela 2. Tabela de frequências absolutas de 0 e 1

Observação p / q	1	0	Total
1	a	b	a + b
0	c	d	c + d
Total	a + c	b + d	a + b + c + d

Fonte: Fávero e Belfiore (2017)



A medida de emparelhamento simples iguala os pesos nas correspondências, e considera as ausências simultâneas (0-0) informativas. Esta é a medida de similaridade mais amplamente utilizada para variáveis binárias e é expressa a partir da form. (2) seguinte:

$$S_{pq} = \frac{a + d}{a + b + c + d} \quad (2)$$

onde, a é o evento 1-1, b é o evento 1-0, c é o evento 0-1 e d representa o evento de ausência simultânea 0-0.

Por outro lado, o coeficiente de Jaccard propõe um cálculo no qual as correspondências 0-0 são consideradas irrelevantes, ou seja, não contribuem com informações úteis sobre a semelhança entre os objetos, como pode ser observado na form. (3), representada a seguir:

$$S_{pq} = \frac{a}{a + b + c} \quad (3)$$

onde, a é o evento 1-1, b é o evento 1-0, c é o evento 0-1.

O coeficiente de DICE, assim como o coeficiente de Jaccard, desconsidera as correspondências de ausência simultânea 0-0; entretanto, ele atribui o dobro do peso às correspondências 1-1. Essa medida é expressa através da fórmula form. (4), conforme apresentado abaixo:

$$S_{pq} = \frac{2.a}{2.a + b + c} \quad (4)$$

onde, a é o evento 1-1, b é o evento 1-0, c é o evento 0-1.

De acordo com Everitt et al. (2013), embora não exista uma regra estrita e rápida em relação à escolha do coeficiente, cada conjunto de dados deve ser avaliado pelo analista em seus méritos. Isso ressalta a importância fundamental do analista, uma vez que a utilização de diferentes coeficientes de similaridade pode levar a resultados bastante divergentes.

### **Método hierárquico aglomerativo**

Como parte do processo de análise dos dados, foi testado e apresentado de forma comparativa o método hierárquico aglomerativo. De acordo com Everitt et al. (2013), essa

abordagem consiste em um método no qual os dados são particionados passo a passo, e sua construção pode variar desde um único grupo contendo todos os indivíduos até  $n$  grupos, cada um contendo um único indivíduo.

As técnicas de agrupamento hierárquico podem ser subdivididas em métodos aglomerativos, que realizam fusões sucessivas dos  $n$  indivíduos em grupos, e métodos divisivos, que separam os  $n$  indivíduos sucessivamente em agrupamentos mais refinados. Ambos os tipos de agrupamento hierárquico podem ser interpretados como uma tentativa de encontrar o passo ideal, em algum sentido definido (Everitt et al. 2013).

No método hierárquico aglomerativo, uma etapa fundamental é a escolha do método de encadeamento das observações. Conforme afirmado por Fávero e Belfiore (2017), existem três principais métodos de encadeamento em esquemas aglomerativos: o método do tipo único (também conhecido como “nearest neighbor” ou “single linkage”), o método completo (também conhecido como “furthest neighbor” ou “complete linkage”) e o método médio (também conhecido como “between groups” ou “average linkage”). Neste trabalho, foram testados e avaliados os resultados utilizando os três métodos de encadeamento.

O método de encadeamento único, é uma abordagem que prioriza as menores distâncias entre observações. Ele é comumente empregado em situações onde as observações são distintas e a ênfase recai sobre a proximidade entre elas. Esse método é representado pela seguinte form. (5):

$$d_{(MN)W} = \min\{ d_{MW} ; d_{NW} \} \quad (5)$$

onde,  $d_{MW}$  e  $d_{NW}$  são distâncias entre as observações mais próximas.

Por outro lado, o método de encadeamento completo privilegia as maiores distâncias entre observações ou grupos, o que resulta na formação de novos agrupamentos. Esse método é recomendado em situações envolvendo observações semelhantes entre si, e pode ser observado na seguinte form. (6):

$$d_{(MN)W} = \max\{ d_{MW} ; d_{NW} \} \quad (6)$$

onde,  $d_{MW}$  e  $d_{NW}$  são distâncias entre as observações mais distantes.

O método de encadeamento completo é aplicado quando se deseja enfatizar a dissimilaridade entre grupos ou observações, destacando as diferenças significativas entre eles. Isso pode ser particularmente útil em cenários onde a distinção entre os grupos é um aspecto crítico da análise de agrupamento.

O método de encadeamento médio é uma técnica que envolve a junção de dois grupos com base na distância média entre todos os pares de observações pertencentes a esses grupos. Esse método é representado pela seguinte form. (7):

$$d_{(MN)W} = \frac{\sum_{p=1}^{m+n} \sum_{q=1}^w d_{pq}}{(m+n) \cdot (w)} \quad (7)$$

onde,  $d_{(MN)W}$  é a média de todas as distâncias (MN) entre os pares de observações, m e n são o número de observações nos grupos M e N, respectivamente, e w é o número de observações no grupo W.  $d_{pq}$  representa a distância entre a observação p no grupo M ou N e a observação q no grupo W.

O método de encadeamento médio é aplicado quando se deseja avaliar a similaridade entre grupos com base na média das distâncias entre seus membros. Ele é útil em situações onde a variabilidade dentro dos grupos é uma consideração importante, e a análise procura destacar agrupamentos que apresentam uma média de distância entre membros relativamente pequena. A seguir, a Tabela 3 ilustra a aplicação dos três métodos de encadeamento:

Tabela 3. Métodos de encadeamento

Método de encadeamento	Ilustração	Distância (Dissimilaridade)
Único ("nearest neighbor" ou "single linkage")		$d_{23}$
Completo ("furthest neighbor" ou "complete linkage")		$d_{15}$
Médio ("between groups" ou "average linkage")		$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$

Fonte: Adaptado de Fávero e Belfiore (2017)

Além dos três principais métodos de encadeamento, também foi verificado o método de Ward. Nesse método, a fusão de dois grupos é determinada através do cálculo do erro da soma dos quadrados. Em cada etapa, o objetivo é minimizar o aumento total na soma dos quadrados dos erros dentro do grupo (Everitt et al. 2013).

Essa abordagem visa diminuir a soma dos quadrados das diferenças entre as variáveis dentro de cada grupo. Ela tende a produzir grupos de tamanhos semelhantes e é menos sensível a variações na dispersão dos dados.

Como mencionado anteriormente, o método de Ward é um dos vários métodos de encadeamento empregados na análise de agrupamento hierárquico aglomerativo. Ele é reconhecido por gerar grupos com tamanhos mais uniformes em comparação com outros métodos. Embora não seja comumente utilizado para matrizes binárias, quando aplicado a esse tipo de matriz, o modelo de Ward calcula a dissimilaridade entre grupos com base na redução da soma dos quadrados das diferenças nas entradas binárias. A form. (8) correspondente a esse modelo é a seguinte:

$$D(UV, W) = \frac{n_U n_V}{n_U + n_V} d(U, W) + \frac{n_W n_V}{n_W + n_V} d(V, W) - \frac{n_V}{n_U + n_V + n_W} d(U, V) + \frac{n_U n_W}{(n_U + n_V)(n_U + n_W)} d(V, W) \quad (8)$$

onde,  $D(UV, W)$  é a dissimilaridade entre o cluster resultante de combinar os clusters U e V e o cluster W.  $n_U$ ,  $n_V$ , e  $n_W$  são os números de observações nos clusters U, V e W, respectivamente. E,  $d(U, W)$ ,  $d(V, W)$  e  $d(U, V)$  são as distâncias entre os clusters U e W, V e W, e U e V, respectivamente.

### **Método aglomerativo não hierárquico**

Ao contrário dos métodos hierárquicos de agrupamento, que seguem um caminho passo a passo na criação de agrupamentos com estrutura hierárquica, os métodos não hierárquicos utilizam algoritmos para criar agrupamentos mais homogêneos dentro de si, sem a obrigação de seguir um processo hierárquico (Fávero e Belfiore, 2017).

Neste estudo, adotou-se o método k-Means para análise e comparação. O k-Means é um dos procedimentos mais amplamente empregados por pesquisadores em várias disciplinas. Nesse método, a quantidade de grupos é definida inicialmente pelo pesquisador. Além disso, essa abordagem pode ser aplicada após a execução de um método hierárquico, especialmente quando não há uma clara noção de quantos grupos devem ser formados. Nessas circunstâncias, os resultados obtidos pelo método hierárquico podem ser usados como entrada para o método não hierárquico (Fávero e Belfiore, 2017).

Outra abordagem para determinar o número ideal de grupos a ser usado como entrada para o método K-Means é aplicar o método do Cotovelo (Elbow). Este método é uma técnica de análise de agrupamento baseada em representação gráfica. O nome deriva da semelhança do gráfico da variância explicada em relação ao número de grupos com um cotovelo, em que o cotovelo representa o ponto no gráfico em que o aumento na variância explicada começa a diminuir. Isso sugere que a adição de mais grupos não contribui significativamente para a

melhoria do desempenho do modelo. Assim como as entradas a partir dos modelos hierárquicos, também se utilizou o modelo do cotovelo como parte deste trabalho. Essa abordagem visual e heurística fornece uma maneira eficaz de determinar o número adequado de clusters e é uma contribuição valiosa para a avaliação do desempenho do método K-Means.

Após a definição do número de grupos, o método empregado pelo algoritmo K-Means começa selecionando inicialmente K pontos representativos como centroides iniciais. Em seguida, cada ponto é associado ao centroide mais próximo, utilizando uma métrica de proximidade específica que é escolhida. Uma vez que os agrupamentos são estabelecidos, os centroides de cada grupo são ajustados. O processo algorítmico continua de maneira iterativa por meio dessas duas etapas até que os centroides não sofram mais alterações ou até que critérios alternativos de convergência relaxada sejam satisfeitos. Isso está representado na Figura 2.

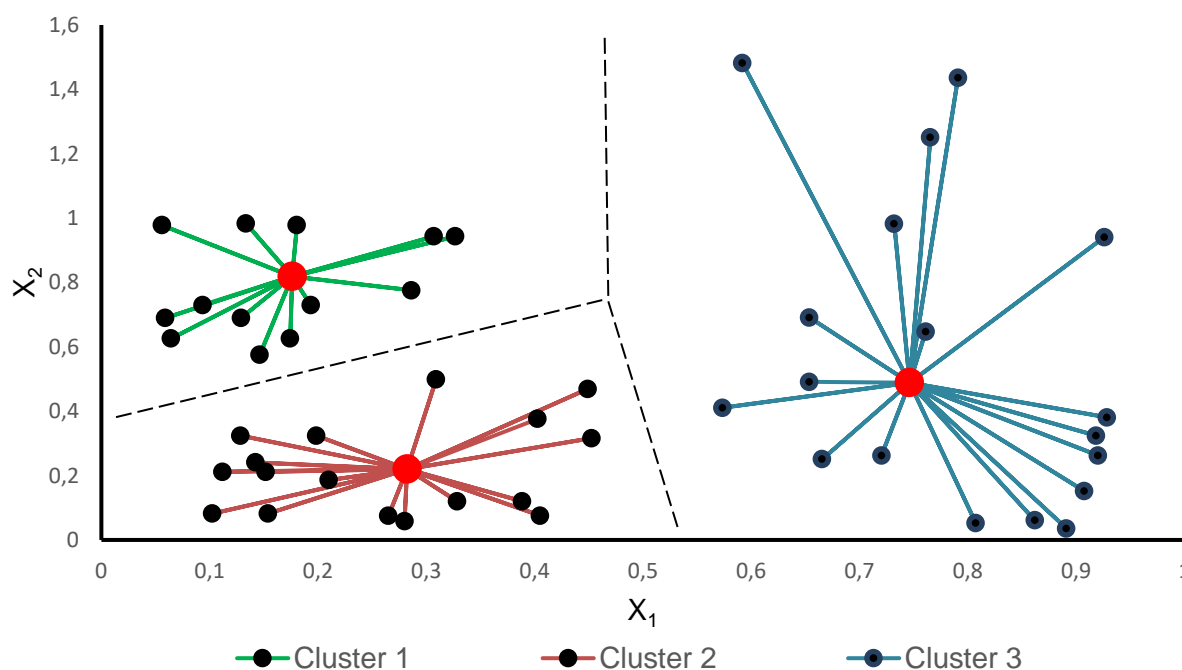


Figura 2. Representação do método K-means

Fonte: Adaptada de Fávero e Belfiore (2017)

Uma informação importante e perceptível durante a análise pelo método K-Means é a necessidade de fixação de sementes aleatórias, chamadas de "set. seeds", para a classificação dos grupos pelo algoritmo. Essa necessidade surge devido ao processo do próprio algoritmo que começa com uma inicialização aleatória dos centroides dos grupos, o

que significa que cada execução pode levar a um resultado ligeiramente diferente, mesmo para o mesmo número de grupos. Isso é normal e é uma característica do algoritmo.

### **Coeficiente de silhueta**

O coeficiente de silhueta é uma métrica de qualidade usada para avaliar a homogeneidade dos grupos em uma análise de agrupamento. Neste trabalho, ele foi aplicado com o intuito de criar um parâmetro de comparação entre os modelos apresentados, permitindo a avaliação da eficácia dos diferentes métodos de agrupamento utilizados.

Conforme descrito por Aggarwal e Reddy (2014), o coeficiente de silhueta é derivado considerando as distâncias dentro e entre os grupos. Para um ponto específico “ $x_i$ ”, inicialmente, calcula-se a média das distâncias para todos os outros pontos no mesmo grupo, denominando esse valor como “ $a_i$ ”. Em seguida, para cada cluster que não inclui “ $x_i$ ”, calcula-se a média das distâncias entre  $x_i$  e todos os pontos de dados em cada um desses grupos, chamando esse valor de “ $b_i$ ”. Com base em “ $a_i$ ” e “ $b_i$ ”, calcula-se o coeficiente de silhueta para um ponto. A média de todas as silhuetas no conjunto de dados é referida como a “largura média das silhuetas” para todos os pontos no conjunto de dados. Para avaliar a qualidade de um agrupamento, é possível calcular o coeficiente de silhueta médio para todos os pontos. Essa métrica oferece uma visão sobre a qualidade dos agrupamentos obtidos. A form. (9) a seguir expressa o conceito:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (9)$$

onde,  $S_i$  representa o coeficiente de silhueta para o ponto de dados  $i$ ,  $a_i$  é a média das distâncias do ponto  $i$  para todos os outros pontos no mesmo grupo (distância intra-grupo),  $b_i$  é a menor média das distâncias do ponto  $i$  para todos os pontos em qualquer cluster diferente do seu próprio (distância intergrupo). O coeficiente de silhueta  $S_i$  é calculado como a diferença entre  $b_i$  e  $a_i$ , dividida pelo máximo entre  $a_i$  e  $b_i$ .

O resultado do coeficiente de silhueta varia de -1 a 1 e é interpretado da seguinte forma: um valor próximo de 1 indica que o ponto está bem classificado dentro de seu grupo, refletindo uma ótima separação, onde os pontos estão claramente associados ao grupo correto; um valor próximo de 0 indica que o ponto está na fronteira entre dois grupos, sugerindo uma ambiguidade na classificação, indicando que o ponto pode estar igualmente próximo de grupos diferentes; e um valor próximo de -1 indica que o ponto pode ter sido atribuído ao grupo errado, sugerindo uma má separação ou sobreposição com outros grupos, onde a atribuição do ponto ao grupo é inadequada.

## Resultados e Discussão

A análise de agrupamento no contexto deste estudo implica uma maior compreensão dos resultados dos índices setoriais do IEG-M. Ao agrupar os municípios com resultados semelhantes, pode-se facilitar ações do setor público, uma vez que o agrupamento resume muitos municípios em um número menor e representativo. Essa facilidade na compreensão proporcionada pela análise permite, por exemplo, a identificação de informações ocultas sobre movimentos regionais que não seriam visíveis em uma análise individualizada. Além disso, pode facilitar algumas ações do governo estadual, como campanhas publicitárias que podem ser direcionadas de forma mais assertiva de acordo com o perfil de cada grupo, ou ainda direcionar campanhas de fiscalização com ênfase nos setores críticos de cada grupo. No contexto geral, a análise de agrupamento pode ajudar na tomada de decisões, permitindo que os setores envolvidos compreendam melhor os resultados do IEG-M e identifiquem tendências, oportunidades e desafios.

Como ponto de partida deste estudo, procedeu-se à coleta de dados descritivos do banco de dados subjacente, a fim de melhor compreender sua estrutura e características. Na sequência, conforme apresentado na Tabela 4, estão os resultados consolidados em uma tabela de frequência para cada setor, incluindo o próprio IEG-M.

Tabela 4. Resultado consolidado IEG-M 2022

	C	C+	B	B+	A
IEG-M	447	164	33	0	0
i-Educ	338	182	118	6	0
i-Saúde	231	235	173	5	0
i-Plan	548	54	35	7	0
i-Fiscal	26	136	437	45	0
i-Amb	493	77	60	14	0
i-Cidade	440	77	78	41	8
i-Gov TI	444	93	61	38	8

Fonte: Resultados originais da pesquisa

Conforme evidenciado na primeira linha dos resultados do IEG-M, o agrupamento construído pela análise enviesada pelo orçamento público e proposta pelo TCE-SP, divide as prefeituras em três grupos, sendo que 69,40% delas agrupadas no grupo com nota C - Baixo nível de adequação. Ou seja, um agrupamento que dificulta ações em um contexto mais amplo, uma vez que a massiva maioria dos dados é agrupada em uma única categoria, o que limita a capacidade de distinguir nuances nas avaliações e torna a tomada de decisões eficazes mais desafiadora.

Esses resultados são também visualizados no gráfico representado na Figura 3, que foi incluída com o intuito de proporcionar uma visualização mais clara e compreensível das informações da Tabela 4, tornando mais acessível a interpretação dos resultados.

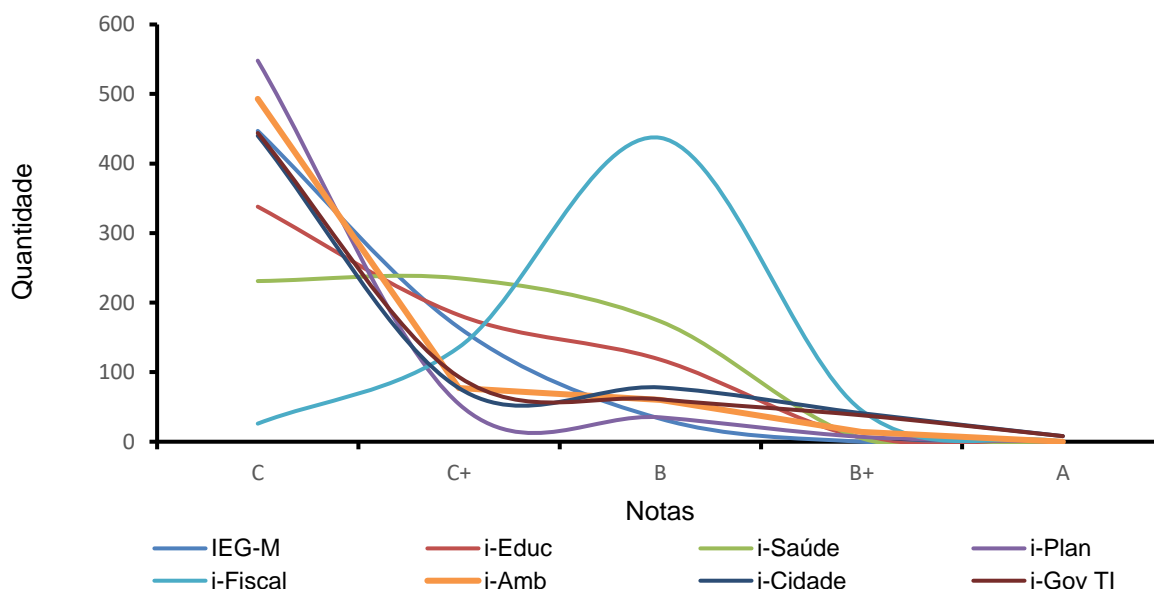


Figura 3. Resultados acumulados IEG-M  
Fonte: Resultados originais da pesquisa

Vale destacar, analisando o gráfico da Figura 3, o desempenho das prefeituras no resultado do i-Fiscal, onde 67,86% delas obtiveram a nota B – Efetiva, um resultado muito positivo quando comparado aos demais setores. Outro ponto de destaque são os resultados do i-Saúde, que se mostram mais equilibrados em relação aos demais setores.

Conforme mencionado anteriormente, as variáveis apresentadas no banco de dados são categóricas, exibidas em intervalos de valores de distâncias desiguais e por notas, as quais são classificadas com base no percentual atingido no questionário do IEG-M. Na Tabela 5, pode ser verificado um recorte dos primeiros dez resultados do IEG-M setorial.

Tabela 5. Recorte dos dez primeiros resultados do IEG-M Setorial 2022

Cidade	i-Saúde	i-Plan	i-Fiscal	i-Educ	i-Amb	i-Cidade	i-Gov
Adamantina	C+	C	B	C	C+	C	B+
Adolfo	C+	C+	C+	C+	C	C	C
Aguaí	C+	C+	B	C+	C	B	C
Águas da Prata	C	C	B	C+	B	C	C
Águas de Lindóia	B	C	B	C+	C	C+	C+
Águas de Santa Bárbara	C+	C	B	C	C+	C	C
Águas de São Pedro	B	C	B	B	B	C	C
Agudos	C	C	B	C	C	B	C
Alambari	C+	C	B+	C	C	C	C
Alfredo Marcondes	B	C	B	C	C	C	C

Fonte: TCESP (2023)



Diante dos dados apresentados na Tabela 5 e com o objetivo de evitar ponderações arbitrárias, o que resultaria na atribuição subjetiva de pesos, importância ou critérios, a análise dos dados se deu a partir de uma matriz binária. Na Tabela 6, apresenta-se um recorte dos dez primeiros resultados para o setor da saúde. Nessa matriz, os resultados são organizados entre zeros e uns, sendo zero para a ausência de evento e um para a ocorrência do evento.

Tabela 6. Matriz binária i-Saúde (recorte dos dez primeiros resultados)

Cidade	i-Saúde_C	i-Saúde_C+	i-Saúde_B	i-Saúde_B+	i-Saúde_A
Adamantina	0	1	0	0	0
Adolfo	0	1	0	0	0
Aguai	0	1	0	0	0
Águas da Prata	1	0	0	0	0
Águas de Lindóia	0	0	1	0	0
Águas de Santa Bárbara	0	1	0	0	0
Águas de São Pedro	0	0	1	0	0
Agudos	1	0	0	0	0
Alambari	0	1	0	0	0
Alfredo Marcondes	0	0	1	0	0

Fonte: Resultados originais da pesquisa

Prosseguindo, foram testadas combinações entre os métodos de encadeamento e as medidas de similaridade propostas, para identificar as melhores combinações para o estudo. Os resultados se mostraram mais consistentes nas combinações entre os métodos de encadeamento Complete Linkage e Ward e as medidas de similaridades propostas, como ilustrado nos dendrogramas a seguir. A Figura 4 demonstra o resultado da combinação do coeficiente de Jaccard e o método de encadeamento Complete Linkage.

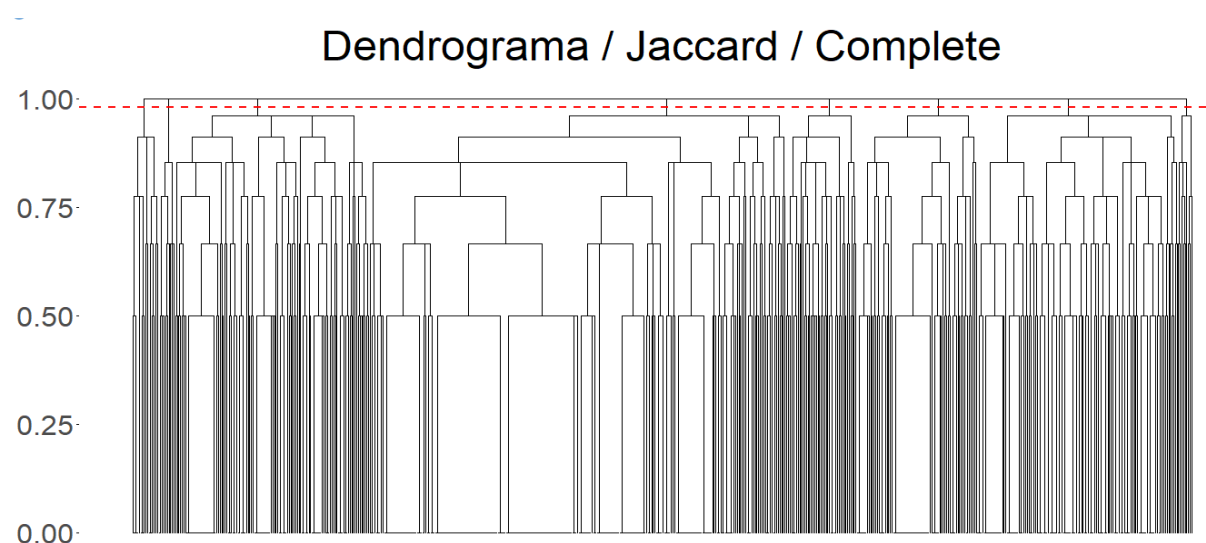


Figura 4. Dendrograma IEG-M/2022, coeficiente de Jaccard e método de encadeamento Complete Linkage

Fonte: Resultados originais da pesquisa

Nessa combinação, exposta na Figura 4, é possível visualizar claramente as estruturas dos agrupamentos, e a linha tracejada em vermelho na parte superior do gráfico demonstra um agrupamento mínimo de oito grupos. Na Figura 5, o dendrograma contém o resultado da combinação do método de emparelhamento simples com o método Complete Linkage.

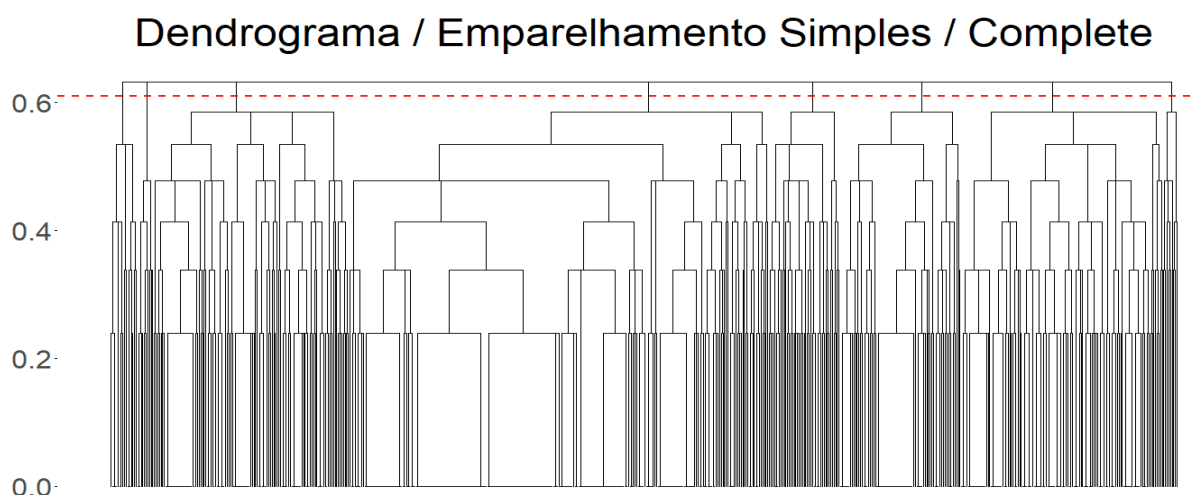


Figura 5. Dendrograma IEG-M/2022, método de emparelhamento simples e método de encadeamento Complete Linkage

Fonte: Resultados originais da pesquisa

Assim como no resultado anterior, as estruturas dos agrupamentos se apresentaram muito claras, tendo como primeira possibilidade a formação de oito grupos. A Figura 6, ilustra o resultado da combinação do coeficiente DICE com o método Complete Linkage.

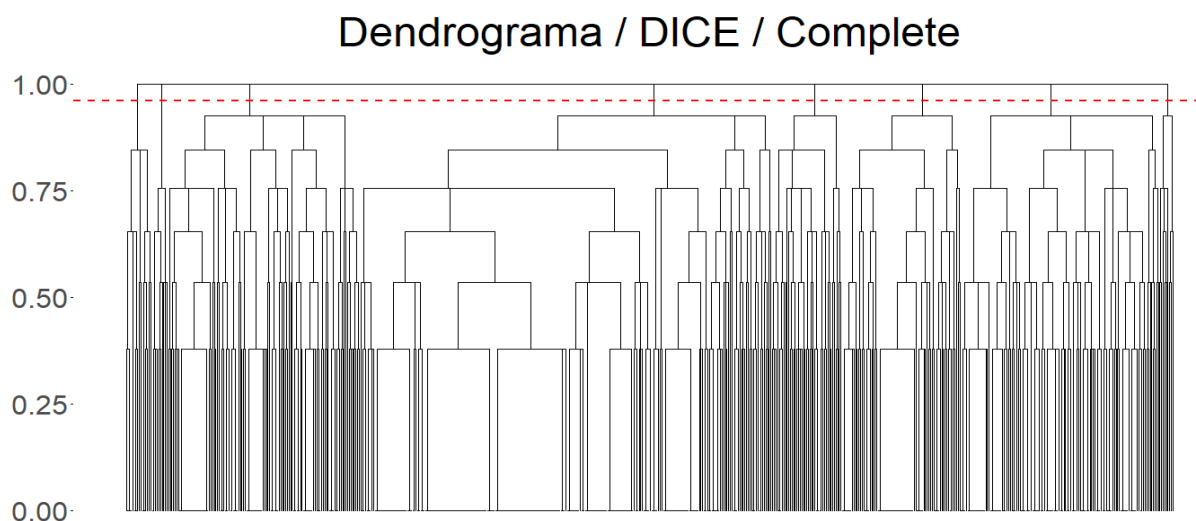


Figura 6. Dendrograma IEG-M/2022, coeficiente DICE e método de encadeamento Complete Linkage

Fonte: Resultados originais da pesquisa

Assim como todas as combinações feitas com o método de emparelhamento Complete Linkage, a combinação representada na Figura 6 apresenta uma clareza na formação dos grupos e uma primeira possibilidade de agrupamento com oito grupos.

Comparando as primeiras possibilidades de agrupamento visualizadas nos dendrogramas anteriores, nota-se uma indicação de quantidades de grupos muito semelhantes e uma constituição idêntica em termos de quantidade de prefeituras em cada grupo. Isso sugere que a medida de similaridade não representou uma variável significativa para essas combinações. No entanto, a situação é diferente nas combinações entre as medidas de similaridade propostas e o método de encadeamento Ward, conforme observado nos dendrogramas a seguir. A Figura 7 demonstra a combinação entre a medida de similaridade proposta pelo coeficiente de Jaccard e o método de encadeamento Ward Linkage.

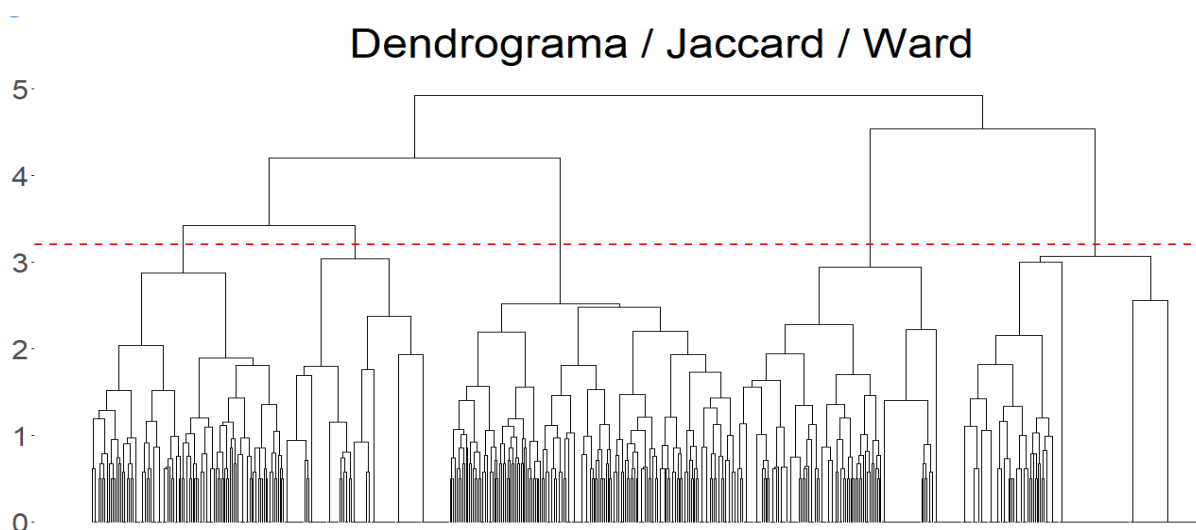


Figura 7. Dendrograma IEG-M/2022, coeficiente de Jaccard e método de encadeamento Ward  
Fonte: Resultados originais da pesquisa

Nessa combinação, é possível visualizar diversas possibilidades de formação de grupos e de forma bastante clara, começando a partir de uma formação mínima com dois grupos. Para fins de uma posterior comparação com os demais modelos, neste trabalho optou-se por identificar a formação com cinco grupos, conforme destacado pela linha tracejada vermelha na Figura 7. Na Figura 8, a seguir, apresentam-se os resultados com a combinação da medida de emparelhamento simples e a medida de encadeamento no método Ward Linkage.

## Dendrograma / Emparelhamento Simples / Ward

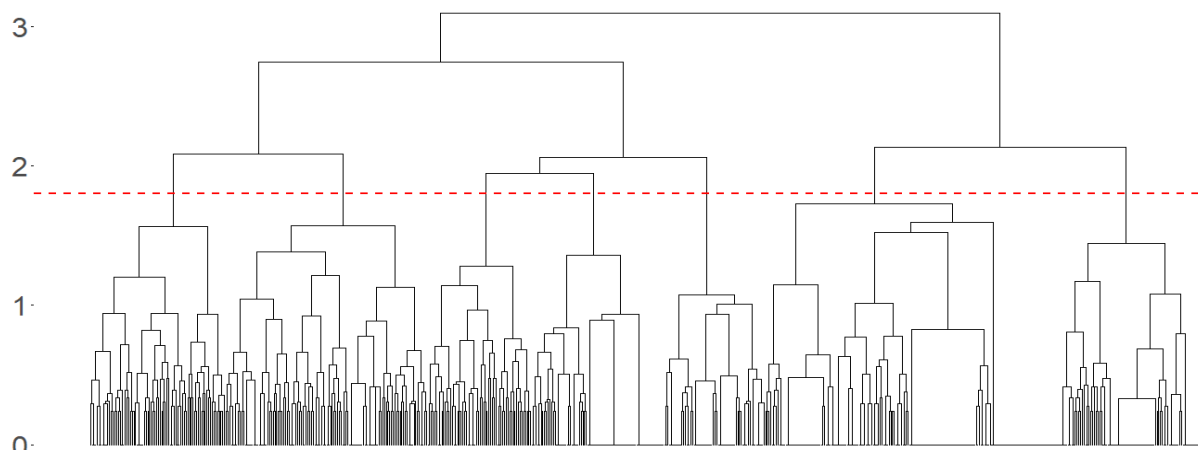


Figura 8. Dendrograma IEG-M/2022, método de emparelhamento simples e método de encadeamento Ward

Fonte: Resultados originais da pesquisa

Com a combinação do método Ward e o método de emparelhamento simples, conforme a Figura 8, as estruturas também se apresentaram de forma bem clara, e optou-se por demarcar a formação de sete grupos para fins didáticos e comparativos com outros modelos. Na Figura 9, apresentam-se os resultados com a combinação do método Ward Linkage e o coeficiente DICE.

## Dendrograma / DICE / Ward

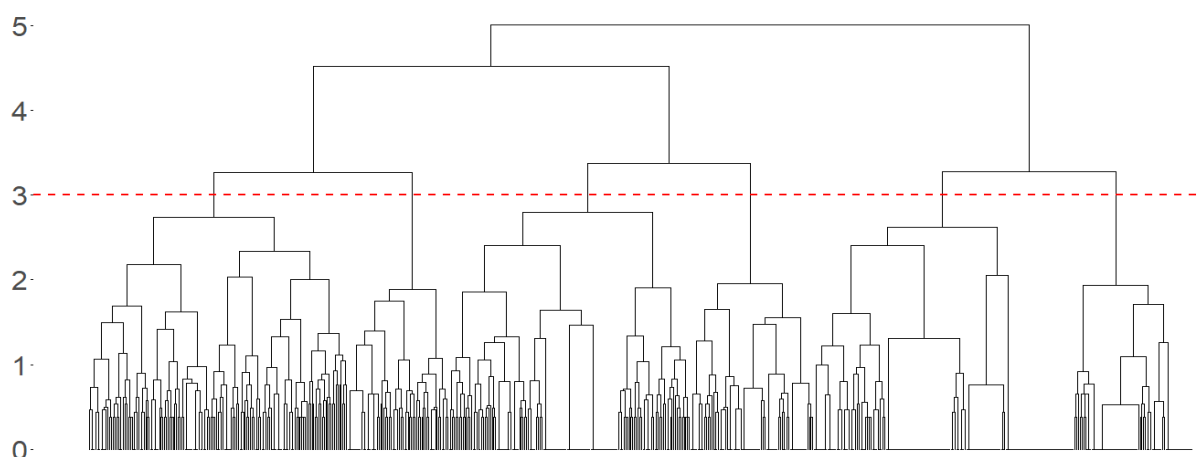


Figura 9. Dendrograma IEG-M/2022, coeficiente DICE e método de encadeamento Ward

Fonte: Resultados originais da pesquisa

Nessa combinação, a estrutura dos grupos também está bem definida, e optou-se por delimitar a formação de seis grupos. Vale ressaltar neste ponto que a escolha da quantidade de grupos nestes métodos se deu por um critério de facilitar comparação com os demais

resultados, permitindo uma avaliação qualitativa da formação desses grupos entre os diferentes métodos, como será evidenciado na sequência do estudo. No entanto, como é possível verificar nos dendrogramas, existem outras possibilidades de agrupamentos, podendo essas serem definidas de acordo com a necessidade do tomador de decisão ou do analista responsável pela análise. A seguir, na Tabela 7, pode-se verificar a formação de cada grupo em sua magnitude, para cada modelo estudado até aqui.

Tabela 7. Formação dos grupos dentro de cada medida de similaridade para os métodos Ward e Complete

Grupo	Complete Linkage			Ward Linkage		
	Jaccard	Emp. simples	DICE	Jaccard	Emp. simples	DICE
1	16	16	16	109	78	151
2	10	10	10	127	60	141
3	118	118	118	162	170	152
4	253	253	253	153	58	72
5	43	43	43	93	86	69
6	73	73	73		116	59
7	122	122	122		76	
8	9	9	9			

Fonte: Resultados originais da pesquisa

Conforme mencionado anteriormente, a formação dos grupos resultantes das combinações usando o método de encadeamento Complete Linkage gerou grupos idênticos. Além disso, é notável que esses grupos são bastante diversificados em relação ao tamanho, variando desde o maior grupo, com 253 prefeituras, até o menor, com apenas 9 prefeituras. Essa variação pode indicar a presença de possíveis “outliers” e ressalta a importância de considerar a incorporação desses grupos menores por grupos maiores, dependendo dos objetivos da análise e das necessidades do tomador de decisão. Em contraste, os grupos gerados pelo método de encadeamento Ward Linkage apresentaram uma maior diversidade em termos da quantidade de grupos formados, além de uma distribuição mais equânime na quantidade de municípios por grupos.

A seguir, com o propósito de demonstrar os resultados internos obtidos a partir dos agrupamentos e apresentar uma possível aplicação, verifica-se uma análise detalhada das combinações entre os métodos de encadeamento Complete Linkage e Ward, juntamente com a medida de similaridade de Jaccard. Na Tabela 8, são exibidos os resultados obtidos ao combinar os métodos de encadeamento Complete Linkage com a medida de similaridade de Jaccard.

**Tabela 8. Resultados internos dos grupos para o método de encadeamento Complete Linkage e o coeficiente de Jaccard**

K	N	i-Saúde	i-Ambiente	i-Cidades	i-Educação	i-Fiscal	i-GOV TI	i-Planej.	Repres.
1	16	C+ 50,00%	C+ 100,00%	C 37,50%	C 93,75%	B 56,25%	C+ 50,00%	C 81,25%	66,96%
2	10	C+ 50,00%	C 100,00%	C 40,00%	C+ 100,00%	C+ 100,00%	C 70,00%	C+ 60,00%	74,29%
3	118	C+ 44,92%	C 86,44%	C 93,22%	C+ 42,37%	B 64,41%	C 80,51%	C 38,98%	64,41%
4	253	C+ 49,41%	C 78,26%	C 88,54%	C 60,08%	B 85,77%	C 72,73%	C 98,81%	76,23%
5	43	B 97,67%	C 44,19%	B 37,21%	C 37,21%	B 88,37%	B 37,21%	C 88,37%	61,46%
6	73	C 100,00%	C 80,82%	C 49,32%	C 69,86%	C+ 52,05%	C 56,16%	C 97,26%	72,21%
7	122	C 36,89%	C 86,07%	C 43,44%	C 49,18%	B 53,28%	C 88,52%	C 97,54%	64,99%
8	9	B 77,78%	B 77,78%	B+ 33,33%	B 66,67%	C+ 55,56%	B+ 77,78%	C 77,78%	66,67%

Fonte: Resultados originais da pesquisa

Na Tabela 8, é possível observar a estrutura interna dos grupos, onde a coluna K representa a quantidade de grupos, a coluna N indica a quantidade de prefeituras em cada grupo K, e nas demais colunas, encontram-se as notas majoritárias em cada setor e a quantidade percentual de prefeituras para aquela nota. Para ilustrar, no setor i-Ambiental, pertencente ao grupo 1, todas as 16 prefeituras desse grupo receberam a nota C+, representando 100% de suas prefeituras com essa classificação, e assim por diante. Dessa maneira, é possível destacar as características distintivas de cada grupo. Por fim, a coluna Repres. é o percentual de representatividade das notas majoritárias em relação ao total de cada grupo, indicando um grau de eficiência nos agrupamentos.

Exemplificando com uma possível aplicação prática, e tomando como ponto de partida o grupo 4, que neste caso é o maior grupo para este modelo, um ente público que queira propor uma ação de fiscalização poderia focar seus esforços no setorial i-Planejamento, onde 98,81% das prefeituras obtiveram nota C indicando um baixo nível de adequação. A seguir na Tabela 9, os resultados da combinação entre o método de encadeamento Ward e o coeficiente de Jaccard.

**Tabela 9. Resultados internos dos grupos para o método de encadeamento Ward e o coeficiente de Jaccard**

K	N	i-Saúde	i-Ambiente	i-Cidades	i-Educação	i-Fiscal	i-GOV TI	i-Planej.	Repres
1	109	C+ 37,61%	C+ 36,70%	C 84,40%	C 42,20%	B 86,24%	C 84,40%	C 47,71%	59,90%
2	127	C 37,01%	C 95,28%	C 85,04%	C 49,61%	C+ 74,02%	C 90,55%	C 82,68%	73,45%
3	162	C+ 37,65%	C 56,17%	C 30,86%	C 48,77%	B 75,31%	C+ 40,74%	C 90,12%	54,23%
4	153	C 42,48%	C 98,04%	C 69,93%	C 98,04%	B 85,62%	C 96,73%	C 99,35%	84,31%
5	93	C 41,94%	C 98,92%	C 89,25%	C+ 59,14%	B 95,70%	C 86,02%	C 100,00%	81,57%

Fonte: Resultados originais da pesquisa

Um ponto destaque nesses resultados são os percentuais de representatividade dos grupos 4 e 5, que alcançam 84,31% e 81,57%, respectivamente. Esses números indicam uma aderência significativa das notas majoritárias apresentadas com o total de cada grupo, sugerindo que ações aplicadas nesses grupos seriam mais assertivas e aderentes.

Para ilustrar uma aplicação prática desses resultados, considerando o grupo 1 como referência e no contexto de publicidade institucional, um órgão público poderia buscar uma

maior divulgação para suas campanhas publicitárias, priorizando as áreas de Proteção dos Cidadãos/Defesa Civil (i-Cidade) e Governança em Tecnologia (i-Gov TI), ambas com um baixo nível de adequação. Para fins de comparação, segue na Tabela 10, os resultados do IEG-M aplicando a regra estabelecida pelo TCE-SP, ponderada pela importância no orçamento público.

**Tabela 10. Resultados internos IEG-M 2022**

K	N	i-Saúde	i-Ambiente	i-Cidades	i-Educação	i-Fiscal	i-GOV TI	i-Planej.	Repres.
1	447	C 45,41%	C 87,02%	C 76,51%	C 64,65%	B 65,77%	C 78,52%	C 94,85%	73,25%
2	164	B 41,46%	C 55,49%	C 53,66%	C+ 39,63%	B 75,61%	C 54,27%	C 65,24%	55,05%
3	33	B 69,70%	C 39,39%	C 30,30%	B 54,55%	B 57,58%	B 27,27%	C 51,52%	47,19%

Fonte: Resultados originais da pesquisa

Os resultados apresentados na Tabela 10 destacam a dificuldade de aplicar qualquer ação com base na classificação proposta pelo TCE-SP, devido à discrepância nas quantidades de prefeituras em cada grupo, bem como à notável baixa variabilidade das notas predominantes, que se limitam principalmente às notas C e B, com uma única nota C+, não permitindo um delineamento claro do grupo. Além disso, é importante mencionar a baixa representatividade dos grupos 2 e 3, que correspondem apenas a 55,05% e 47,19%, respectivamente, do total de cada grupo correspondente. Neste contexto, é relevante destacar que a análise alternativa proposta por meio dos modelos de agrupamentos hierárquicos aglomerativos baseados em uma classificação dos dados sem ponderação, revelou grupos com uma representatividade maior e uma distribuição mais equilibrada das prefeituras, ressaltando um caráter de homogeneidade e um delineamento mais claro do perfil de cada grupo, em contraste com o resultado apresentado pela classificação enviesada do IEG-M.

Continuando e expandindo o estudo sobre métodos de agrupamento, também foi testado o modelo de agrupamento não hierárquico K-Means. Esse modelo é amplamente adotado por analistas de dados e se destaca pela eficácia no agrupamento de dados. No entanto, diferentemente dos modelos hierárquicos, onde o número de grupos é determinado passo a passo, nos modelos não hierárquicos, é necessário definir previamente a quantidade desejada de grupos.

Como um dos objetivos deste trabalho é comparar diferentes modelos de agrupamento, adotou-se duas abordagens para gerar os "inputs" para o modelo K-Means. A primeira abordagem foi utilizar os resultados dos modelos hierárquicos que foram anteriormente testados. A segunda abordagem envolveu a utilização do método do Cotovelo. Essa técnica implica a criação de um gráfico que representa a variância explicada em relação ao número de grupos, apresentando uma semelhança visual com um cotovelo. O ponto de virada no gráfico é onde o aumento na variância explicada começa a diminuir. Isso sugere

que adicionar mais grupos não resulta em melhorias significativas no desempenho do modelo (Everitt et al., 2013)

A premissa subjacente ao método do cotovelo é que, à medida que aumentamos o número de grupos em um modelo de agrupamento, a variância explicada (ou a soma dos quadrados das distâncias dos pontos de dados aos centroides) tende a diminuir. Isso ocorre porque, com mais grupos, é mais provável que cada ponto esteja mais próximo do seu centroide, o que resulta em uma redução na variância explicada. No entanto, chega um ponto em que o ganho na redução da variância explicada diminui significativamente, e é nesse ponto que podemos identificar o número ideal de grupos. Essa representação pode ser visualizada no gráfico da Figura 10 a seguir:

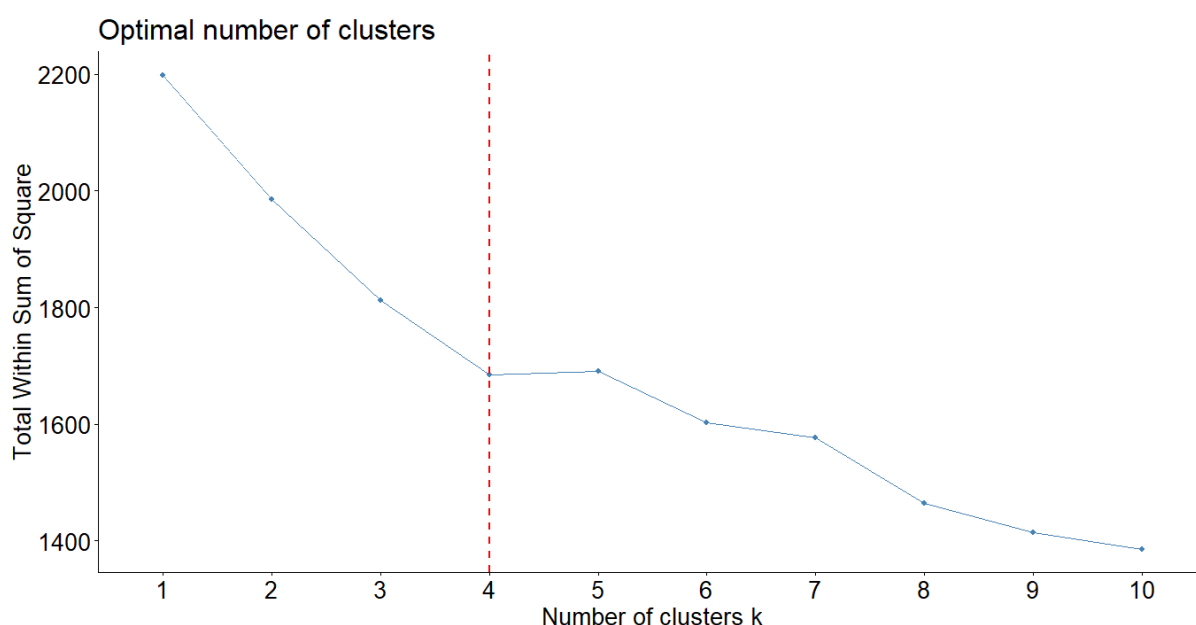


Figura 10. Gráfico de cotovelo  
Fonte: Resultados originais da pesquisa

No gráfico, é perceptível a visualização do ponto de inflexão, conhecido como o 'cotovelo', onde a curva se torna menos acentuada. Para facilitar a identificação desse ponto, ele foi destacado por uma linha tracejada em vermelho. A partir do quarto grupo, tornou-se evidente que o ganho no desempenho do modelo já não era tão significativo em termos de redução de variância.

Utilizando como “input” a primeira sugestão de agrupamento obtida pelo método hierárquico aglomerativo para oito grupos resultantes das combinações efetuadas com o método de encadeamento Complete Linkage, observa-se a seguir os resultados internos dos grupos gerados pelo método K-Means, conforme demonstrado na Tabela 11.



**Tabela 11. Resultados internos para oito grupos no método K-Means**

K	N	i-Saúde	i-Ambiente	i-Cidades	i-Educação	i-Fiscal	i-GOV TI	i-Planej.	Repres
1	68	C+ 51,47%	C+ 42,65%	B 33,82%	C 92,65%	B 77,94%	C+ 55,88%	C 89,71%	63,45%
2	55	C+ 40,00%	C+ 70,91%	C 67,27%	C+ 76,36%	B 81,82%	C 74,55%	C 74,55%	69,35%
3	193	C 50,78%	C 95,34%	C 100,00%	C 81,35%	B 62,69%	C 96,37%	C 89,12%	82,24%
4	63	B 100,00%	C 87,30%	C 100,00%	C 77,78%	B 69,84%	C 92,06%	C 82,54%	87,07%
5	63	C 55,56%	C 96,83%	C+ 50,79%	C 90,48%	B 69,84%	C 82,54%	C 88,89%	76,42%
6	58	B 60,34%	C 55,17%	B 31,03%	B 89,66%	B 68,97%	B 29,31%	C 93,10%	61,08%
7	100	C+ 43,00%	C 100,00%	C 83,00%	C+ 100,00%	B 63,00%	C 92,00%	C 77,00%	79,71%
8	44	C 70,45%	C 90,91%	C 70,45%	C+ 77,27%	B 61,36%	C+ 84,09%	C 79,55%	76,30%

Fonte: Resultados originais da pesquisa

Nota: Foi utilizado o set.seed 07061951 para gerar os grupos

Em uma primeira análise, podemos observar uma distribuição dos grupos mais eficaz com o método K-Means, especialmente quando comparado com os resultados apresentados pelo modelo hierárquico aglomerativo utilizando o método de encadeamento Complete Linkage, conforme mostrado na Tabela 8. Outro ponto notável ao comparar esses métodos é a melhoria no percentual de representatividade, conforme indicado na Tabela 11 pela coluna P, o que sugere uma maior homogeneidade na formação dos grupos. A seguir na Tabela 12, os resultados para um agrupamento de 5 grupos, também sugeridos pelo método hierárquico aglomerativo, utilizando o método K-Means.

**Tabela 12. Resultados internos para 5 grupos no método K-Means**

K	N	i-Saúde	i-Ambiente	i-Cidades	i-Educação	i-Fiscal	i-GOV TI	i-Planej.	Repres
1	193	C+ 81,35%	C 96,37%	C 88,08%	C 45,60%	B 66,84%	C 95,34%	C 83,42%	79,57%
2	89	C 56,18%	C 89,89%	C 47,19%	C 47,19%	B 66,29%	C+ 74,16%	C 86,52%	66,77%
3	74	C+ 39,19%	C+ 98,65%	C 56,76%	C 43,24%	B 81,08%	C 51,35%	C 77,03%	63,90%
4	201	C 78,61%	C 97,01%	C 80,60%	C 74,63%	B 62,69%	C 97,01%	C 87,56%	82,59%
5	87	B 68,97%	B 52,87%	C 27,59%	B 40,23%	B 72,41%	C 31,03%	C 88,51%	54,52%

Fonte: Resultados originais da pesquisa

Nota: Foi utilizado o set.seed 07061951 para gerar os grupos

Em comparação com os resultados do método hierárquico aglomerativo utilizando o método de encadeamento Ward Linkage e a medida de similaridade Jaccard para um agrupamento com 5 grupos, conforme a Tabela 9, o método K-Means não demonstrou uma melhoria na distribuição dos grupos. Pelo contrário, as prefeituras se concentraram nos grupos 1 e 4, sendo que a soma desses dois grupos é responsável por 61,18% do total das prefeituras. No entanto, houve uma melhora no percentual de representatividade, o que sugere que, embora os grupos tenham se formado de maneira menos equilibrada, eles estão se agrupando de forma mais eficiente. Por fim, na Tabela 13, são apresentados os resultados para um agrupamento com 4 grupos, a fim de analisar a sugestão obtida através do método Cotovelo.

**Tabela 13. Resultados internos para 4 grupos no método K-Means**

K	N	i-Saúde	i-Ambiente	i-Cidades	i-Educação	i-Fiscal	i-GOV TI	i-Planej.	Repres
1	201	C+ 82,09%	C 99,50%	C 88,56%	C 44,28%	B 66,17%	C 88,56%	C 82,59%	78,82%
2	127	B 43,31%	C 73,23%	B 32,28%	C 33,07%	B 69,29%	C+ 37,01%	C 88,98%	53,88%
3	113	C+ 38,94%	C+ 66,37%	C 60,18%	C 46,02%	B 80,53%	C 56,64%	C 79,65%	61,19%
4	203	C 78,82%	C 98,52%	C 81,77%	C 76,35%	B 61,58%	C 94,09%	C 88,18%	82,76%

Fonte: Resultados originais da pesquisa

Nota: Foi utilizado o set.seed 07061951 para gerar os grupos

Os resultados da Tabela 13 não trouxeram dados substancialmente diferentes dos anteriores testados pelo método K-Means. Os índices de representatividade permaneceram muito próximos, indicando que a escolha de um agrupamento com quatro grupos neste estudo poderia ser influenciada principalmente pela preferência do analista ou do tomador de decisão, e não pela identificação de uma melhora na qualidade do agrupamento.

Um dos pontos relevantes na análise desses resultados foi a percepção da flexibilidade que o método K-Means oferece na formação dos grupos, juntamente com a facilidade e eficiência na execução do algoritmo, resultando em uma simplificação dos processos computacionais. Isso tornou o método facilmente comparável com outros métodos, além de apresentar uma melhora na qualidade dos agrupamentos.

Uma outra forma de comparar os resultados obtidos nos diferentes métodos de agrupamento é através do uso do coeficiente de Silhueta. Esse coeficiente é uma métrica de avaliação da qualidade dos grupos e mede o quão bem os objetos estão agrupados em seus respectivos grupos. Ele atribui uma pontuação a cada ponto de dados, indicando o quão similar ele é ao seu próprio grupo em comparação com outros grupos próximos. A Tabela 14 apresenta os resultados obtidos através do coeficiente de Silhueta aplicado a cada método utilizado até este ponto, na formação de até dez grupos.

**Tabela 14. Resultados coeficiente silhueta**

Nº de grupos	K-Means	Jaccard Complete Linkage	Emp. Simples Complete Linkage	DICE Complete Linkage	Jaccard WARD	Emp. Simples WARD	DICE WARD
2	0,14868	0,10029	0,13454	0,13454	0,06806	0,07994	0,07108
3	0,14433	0,06340	0,07853	0,07853	0,05607	0,09051	0,08557
4	0,12806	0,03576	0,03897	0,03897	0,07429	0,06483	0,08788
5	0,14077	0,03720	0,03496	0,03496	0,07047	0,07777	0,07494
6	0,14448	0,04481	0,04578	0,04578	0,05758	0,09239	0,09024
7	0,12597	0,04859	0,04818	0,04818	0,06507	0,10065	0,09920
8	0,16256	0,05080	0,05147	0,05147	0,07252	0,09563	0,10880
9	0,15965	0,05066	0,05011	0,05011	0,09512	0,07348	0,08985
10	0,16006	0,06159	0,06212	0,06212	0,10468	0,08359	0,09240

Fonte: Resultados originais da pesquisa

Nota: Foi utilizado o set.seed 07061951 para gerar os grupos

O resultado do coeficiente de silhueta varia de -1 a 1. Uma proximidade a 1 indica que o ponto está bem dentro do seu próprio grupo e distante de outros grupos, o que representa uma boa separação. Por outro lado, uma proximidade a -1 indica que o ponto pode ter sido atribuído ao grupo errado, enquanto valores próximos a 0 indicam que os pontos estão próximos da fronteira entre dois clusters ou em uma região onde não está claro a qual grupo pertencem.

Neste estudo, é notável que todos os resultados se aproximam de zero, como os observados através da Tabela 14. Isso ocorreu devido à natureza dos dados binários e à forma como o coeficiente de Silhueta é calculado.

O coeficiente de Silhueta baseia-se na distância entre os pontos de dados e na definição dos grupos. No caso de dados binários, onde cada elemento da matriz é 0 ou 1, a distância entre os pontos muitas vezes não é suficientemente discriminativa para criar grupos bem definidos. Isso ocorre porque a distância entre os pontos de dados é calculada com base na diferença binária, o que pode não refletir com precisão a similaridade ou dissimilaridade entre os pontos. Além disso, a natureza discreta dos dados binários significa que os pontos de dados geralmente estão muito próximos uns dos outros em termos de distância, o que pode resultar em coeficientes de Silhueta próximos de zero, indicando que os pontos estão próximos do limite de decisão entre grupos.

Em uma análise individualizada dos métodos, isso tornaria o coeficiente de Silhueta ineficiente para definir a qualidade de um agrupamento com base em uma matriz binária. No entanto, no objetivo deste estudo, que é trazer uma nova perspectiva na análise dos resultados setoriais do IEG-M, sem o viés orçamentário, e através de técnicas não supervisionadas de aprendizado de máquina para análises de agrupamento e a comparação dos resultados dessas técnicas, o coeficiente se mostrou eficiente. Isso ocorre porque ele corrobora a análise anterior, na qual os agrupamentos efetuados pelo método K-Means obtiveram um melhor resultado qualitativo, baseado no percentual de representatividade dos grupos, indicando uma maior homogeneidade. É possível observar na Tabela 14 que, para todos os modelos de agrupamentos apresentados, o coeficiente para o método K-means apresentou um resultado qualitativo superior. Isso demonstra ser o método mais adequado para agrupar os 644 municípios paulistas, excluindo a capital, de forma homogênea internamente e heterogênea em relação aos grupos.

Ainda é válido propor um exemplo de aplicação que, embora seja hipotético, ajuda a ilustrar e demonstrar como este estudo pode facilitar a formulação de ações mais abrangentes, assertivas e direcionadas, a partir de uma visão mais completa e equilibrada dos resultados do IEG-M. Suponha-se que um governante do estado de São Paulo deseje distribuir verbas para investimentos de forma equitativa entre os municípios, considerando as

defasagens em cada setor municipal. Poderia-se fazer isso de forma individual, analisando cada uma das 644 prefeituras, ou partir do último resultado do modelo K-means na Tabela 13, que, como visto, se mostrou mais representativo e de maior qualidade.

Nesse caso, dividiria-se os municípios em quatro grupos, distribuindo a verba proporcionalmente ao número de prefeituras em cada grupo. Por exemplo, o Grupo 1 (201 prefeituras) receberia 31,21% da verba total, enquanto o Grupo 2 (127 prefeituras) ficaria com 19,72%. Além disso, poderia-se priorizar os setores críticos por meio desses grupos. No Grupo 1, áreas como meio ambiente, proteção aos cidadãos e governança em tecnologia da informação, onde a grande maioria das prefeituras apresenta um baixo nível de adequação (nota C), seriam prioridade.

Isso permitiria que cada prefeitura desse grupo recebesse proporcionalmente mais recursos para ações relacionadas ao ecossistema e serviços que impactam a qualidade de vida dos cidadãos, como a coleta de resíduos sólidos e a educação ambiental, além de iniciativas voltadas para a segurança dos munícipes diante de acidentes e desastres naturais, e investimentos tecnológicos em áreas como capacitação de pessoal, transparência e segurança da informação.

Por outro lado, o Grupo 2 poderia-se priorizar os setores de meio ambiente e planejamento, recebendo mais recursos para criação de ações para o meio ambiente, e para a implementação de um possível sistema de acompanhamento do planejado em relação ao efetivamente implementado pela prefeitura, além da implementação de métodos para verificar a coerência entre as metas e os recursos empregados.

Neste exemplo, podem-se observar como a análise de agrupamento facilita a tomada de decisão. Nesse caso, o tomador de decisões deixa de analisar individualmente 644 prefeituras e passa a considerar apenas 4 grupos, simplificando significativamente o processo de decisão.

## **Conclusão**

Este estudo comparou diferentes técnicas de aprendizado de máquina não supervisionado para a formação de grupos e demonstrou que, a partir de uma avaliação da qualidade da homogeneidade interna e do nível de representatividade na formação dos grupos, o método não hierárquico K-Means apresentou resultados mais robustos. Também mostrou que a aplicação deste método pode simplificar a tomada de decisões, a administração e a organização do estado, uma vez que o agrupamento se mostrou eficaz na redução do número de observações com as quais o tomador de decisão precisa lidar, facilitando a interpretação dos dados e tornando a visualização dos dados mais intuitiva. Usando exemplos

de agrupamentos, a análise proporcionou uma compreensão mais profunda dos resultados dos índices setoriais do IEG-M, oferecendo não apenas uma nova perspectiva isenta de viés relacionado ao orçamento público, mas também um agrupamento mais equilibrado e homogêneo das prefeituras, com um delineamento claro dos perfis de cada grupo em relação ao próprio IEG-M. Por fim, é importante destacar que os agrupamentos foram realizados sem considerar dados de georreferenciamento, o que poderia proporcionar uma abordagem adicional para a análise da distribuição geográfica dos grupos e, possivelmente, uma compreensão mais completa das tendências subjacentes.

## Referências

Aggarwal, C.C.; Reddy, C.K. 2014. Data Clustering: Algorithms and Applications. Chapman and Hall Book/CRC Press. Boca Raton. Florida, EUA.

Everitt, B.S.; Landau, S.; Leese, M.; Stahl, D. 2011. Cluster Analysis. 5ed. John Wiley & Sons. Chichester, West Sussex, Reino Unido.

Fávero, L.P; Belfiore, P. 2017. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. GEN LTC. Rio de Janeiro, RJ, Brasil.

Instituto Brasileiro de Geografia e Estatística [IBGE]. Censo 2022 –Panorama. 2023. Disponível Sound Parkway NW em :<[https://censo2022.ibge.gov.br/panorama/?utm\\_source=ibge&utm\\_medium=home&utm\\_campaign=portal](https://censo2022.ibge.gov.br/panorama/?utm_source=ibge&utm_medium=home&utm_campaign=portal)>. Acesso em: 20 jul. 2023.

Organização das Nações Unidas [ONU]. Objetivos de Desenvolvimento Sustentável. 2023. Disponível em :<<https://brasil.un.org/pt-br/sdgs>>. Acesso em 20 jul. 2023

Tribunal de Contas do Estado de São Paulo [TCESP]. Manual Índice de Efetividade da Gestão Municipal. 2022. Disponível em :<<https://www.tce.sp.gov.br/publicacoes/manual-ieg-m-2022>>. Acesso em: 19 jul. 2023.

Tribunal de Contas do Estado de São Paulo [TCESP]. Planejamento Estratégico 2022-2026. 2022. Disponível em :<<https://www.tce.sp.gov.br/competencia>>. Acesso em: 19 jul. 2023.

Tribunal de Contas do Estado de São Paulo [TCESP]. Índice de Efetividade da Gestão Municipal. 2022. Disponível em :<[https://painel.tce.sp.gov.br/pentaho/api/repos/%3Apublic%3Aieg\\_m%3Aiegm.wcdf/generatedContent?userid=anony&password=zero#anchor-indices](https://painel.tce.sp.gov.br/pentaho/api/repos/%3Apublic%3Aieg_m%3Aiegm.wcdf/generatedContent?userid=anony&password=zero#anchor-indices)>. Acesso em: 19 jul. 2023.

Tribunal de Contas do Estado de São Paulo [TCESP]. Questionário Índice de Efetividade da Gestão Municipal. 2022. Disponível em :<<https://www.tce.sp.gov.br/sites/default/files/publicacoes/QuestionarioPrincipal-IEGM%202022-AnoBase2021.pdf>>. Acesso em: 19 jul. 2023.