Saneamento e Educação: Explorando Padrões em Municípios Brasileiros através de Clusterização

Pablo Henrique da Silva Lima (Universidade de São Paulo) *lima.pablohs@gmail.com*
Miguel Ângelo Lellis Moreira (Universidade Federal Fluminense) *miguellellis@id.uff.br*

Resumo

O saneamento básico e a educação são fatores fundamentais para o desenvolvimento social e econômico, especialmente em países com desigualdades regionais significativas. Este estudo investiga a relação entre esses aspectos aplicando análise de clusters aos dados do Censo de 2010 do IBGE. Utilizando o algoritmo K-means, os municípios brasileiros foram agrupados em três clusters, considerando variáveis como acesso à água encanada, coleta de lixo e indicadores de escolaridade. Os resultados demonstram uma correlação positiva entre infraestrutura sanitária e desempenho educacional, evidenciando que municípios do Cluster 2, com melhores condições de saneamento, apresentam maiores taxas de frequência escolar e menores índices de atraso. A análise reforça que deficiências na infraestrutura sanitária podem comprometer o desempenho educacional e ampliar desigualdades regionais. O estudo recomenda políticas públicas integradas que abordem simultaneamente melhorias em saneamento e educação, priorizando municípios mais vulneráveis.

Palavras-Chaves: *Saneamento básico, educação, análise de clusters, K-means, desigualdades regionais*

1. Introdução

A educação é um direito fundamental garantido pela Constituição Federal do Brasil (1988) e essencial para o desenvolvimento socioeconômico. No entanto, o sistema educacional brasileiro continua apresentando desafios, refletidos nos baixos desempenhos escolares ao longo dos anos (OECD, 2020). Paralelamente, o saneamento básico também é um direito essencial, mas ainda inacessível para milhões de brasileiros. De acordo com o Instituto Trata Brasil (2023), cerca de 35 milhões de pessoas não possuem acesso à água tratada, e aproximadamente 100 milhões carecem de serviços adequados de esgoto.

O Marco Legal do Saneamento Básico (Lei nº 14.026/2020) estabelece metas para a universalização desses serviços até 2033. Evidências sugerem que melhorias em infraestrutura sanitária impactam positivamente a educação, reduzindo o absentismo escolar e favorecendo o desempenho acadêmico (Abanyie et al., 2021; Sharma et al., 2024). Apesar disso, a relação

entre saneamento e educação ainda é pouco explorada na literatura, que frequentemente prioriza aspectos socioeconômicos gerais (Valencio, Valencio e Baptista, 2023).

Este estudo investiga a correlação entre saneamento e indicadores educacionais em municípios brasileiros, aplicando análise de clusters aos dados do Censo de 2010. Utilizando o algoritmo *K-means*, os municípios foram agrupados em três clusters com base em variáveis como acesso à água encanada, coleta de lixo e escolaridade. A pesquisa busca evidenciar padrões regionais e fornecer subsídios para políticas públicas integradas que promovam melhorias simultâneas em saneamento e educação, contribuindo para a redução das desigualdades socioeconômicas no Brasil.

2. Referencial Teórico

O saneamento básico e a educação são direitos sociais garantidos pela Constituição Federal do Brasil (1988). No entanto, o país ainda enfrenta desafios significativos nesses setores. Segundo o Instituto Trata Brasil (2023), cerca de 35 milhões de brasileiros não têm acesso à água tratada, e aproximadamente 100 milhões carecem de serviços adequados de esgoto. A precariedade da infraestrutura sanitária pode comprometer o desempenho educacional, uma vez que condições inadequadas impactam a frequência e a qualidade do aprendizado (Sharma et al., 2024).

A relação entre saneamento e educação tem sido analisada sob diferentes perspectivas. Abanyie et al. (2021) destacam que investimentos em infraestrutura sanitária em escolas reduzem o absentismo e melhoram o desempenho acadêmico. Além disso, estudos apontam que a ausência desses serviços está associada à perpetuação da pobreza e a dificuldades estruturais nas regiões mais vulneráveis (Valencio, Valencio e Baptista, 2023).

A análise de clusters tem sido amplamente utilizada para estudar padrões socioeconômicos e desigualdades regionais. O algoritmo *K-means*, em particular, permite segmentar dados em grupos homogêneos, sendo aplicado em diversas áreas, como segmentação de mercado e agrupamento de dados sociais (Prakash, 2022). Fávero e Belfiore (2024) ressaltam que a escolha da métrica de agrupamento influencia diretamente a interpretação dos resultados, tornando essencial a definição criteriosa dos parâmetros do modelo.

Diante desse contexto, este estudo busca analisar padrões de saneamento e educação no Brasil por meio da clusterização de municípios, utilizando dados do Censo de 2010 do IBGE.

3. Metodologia

A desigualdade no acesso a serviços básicos, como saneamento e educação, é um desafio central para o desenvolvimento socioeconômico no Brasil. Estudos demonstram que a carência de infraestrutura afeta diretamente o desempenho educacional e contribui para o aprofundamento das desigualdades regionais. Dessa forma, investigar a relação entre esses fatores é crucial para a formulação de políticas públicas mais eficazes.

Para analisar essa dinâmica, a pesquisa utiliza a técnica de análise de clusters, que permite agrupar dados de forma a identificar padrões e relações entre variáveis. O método *K-means* foi escolhido por sua capacidade de segmentar os municípios brasileiros em grupos homogêneos, com base em indicadores de saneamento e educação.

As variáveis de saneamento básico analisadas foram:

Acesso à água encanada;

Banheiro em casa e acesso à água encanada;

Acesso à coleta de lixo.

As variáveis de educação incluíram:

IDHM de Frequência Escolar (proporção de crianças e jovens até 18 anos frequentando a escola);

IDHM de Escolaridade (percentual de pessoas com 18 anos ou mais com o ensino fundamental completo);

Atraso escolar no ensino fundamental (percentual de estudantes com idade superior à esperada para o ano escolar no ensino fundamental).

A pesquisa foi baseada nos dados do Censo de 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE), que forneceu informações sobre saneamento básico e indicadores educacionais. A opção por utilizar o Censo de 2010 se deve ao impacto da pandemia de COVID-19 no Censo de 2022, que poderia distorcer os dados de frequência escolar.

Tabela 1 – Dados brutos do Censo 2010 por município

Territorialidades	% da população em domicílios com água encanada 2010	% da população que vive em domicílios com banheiro e água encanada 2010	% de pessoas em domicílios urbanos com coleta de lixo 2010	Subíndice de frequência a escolar - IDHM Educação 2010	Subíndice de escolaridade - IDHM Educação 2010	% de 6 a 14 anos no ensino fundamental com 2 anos ou mais de atraso idade-série 2010	Renda per capita média do 4º quinto mais pobre 2010
Abadia de Goiás (GO)	93,06	99,01	99,83	0,702	0,489	22,55	611,3
Abadia dos Dourados (MG)	88,5	98,18	98,03	0,673	0,394	20,39	577,33
Abadiânia (GO)	94,5	94,7	98,49	0,669	0,433	14,92	536,24
Abaeté (MG)	98,4	97,43	98,42	0,689	0,363	15,83	567,51
Abaetetuba (PA)	68,86	44,71	97,86	0,599	0,432	24,47	295,64
Xique-Xique (BA)	84,97	74,2	93,94	0,547	0,368	29,67	270,66
Zabelê (PB)	80,03	80,38	99,75	0,755	0,355	6,78	328,83
Zacarias (SP)	96,25	100	100	0,802	0,476	2,9	635,81
Zé Doca (MA)	89,28	46,12	89,38	0,589	0,372	18,7	277,4
Zortéa (SC)	96,08	99,91	99,58	0,734	0,536	11,58	876,79

Fonte: IBGE, Censo 2010

A partir dos dados coletados, foram aplicadas técnicas de limpeza e transformação de dados utilizando a biblioteca **Pandas** (McKinney, 2010), uma ferramenta de código aberto amplamente adotada para a manipulação de dados relacionais de forma eficiente e prática (McKinney, 2018).

Para organizar os dados, foi utilizada a função *iloc* do **Pandas** (versão 2.2.2), que possibilitou a seleção e limitação das linhas correspondentes aos municípios a serem analisados. A conversão de dados não numéricos para o formato *string* foi realizada com o método *astype*. Para transformar os valores percentuais em formato decimal, as colunas foram divididas por 100 e, em seguida, os resultados foram arredondados para seis casas decimais utilizando a função round. Linhas contendo dados ausentes foram eliminadas por meio da função *dropna*.

Com os dados devidamente organizados por município, procedeu-se à normalização utilizando o método Z-score. Segundo Larson e Farber (2015), o Z-score indica o número de desvios padrão em que um valor 'X' se encontra em relação à média μ, conforme apresentado pela equação 1:

$$z = \frac{(x-\mu)}{\sigma} \quad (1)$$

Onde:

z é o valor normalizado;

x é o valor da variável original;

μ é a média da variável;

σ é o desvio padrão da variável.

Dentro do ambiente Python, a normalização dos dados foi realizada por meio da função StandardScaler, presente no pacote **Scikit-learn** (Cournapeau, 2010). Segundo Reitz e Schlusser (2017), o **Scikit-learn** é uma biblioteca de aprendizado de máquina amplamente utilizada, que oferece uma variedade de ferramentas, incluindo redução de dimensões, imputação de dados ausentes, modelos de regressão e classificação, algoritmos de árvore, agrupamento, ajuste automático de parâmetros do modelo, entre outros.

A técnica de clusterização de dados é um tema central na literatura de mineração de dados e aprendizado de máquina, dada sua ampla gama de aplicações, como segmentação de mercado, filtragem colaborativa e análise de redes sociais (Aggarwal e Reddy, 2013). Em sua essência, a clusterização busca particionar um conjunto de dados em grupos, ou clusters, de forma que os pontos de dados dentro de cada grupo sejam o mais semelhantes possível. A definição precisa do problema de clusterização pode variar de acordo com o modelo adotado, como métodos baseados em distância ou modelos probabilísticos. Além disso, a clusterização também pode ser utilizada como uma etapa intermediária em outras tarefas de mineração de dados, como classificação e detecção de outliers.

Bruce e Bruce (2019) destacam que a técnica de clusterização tem como objetivo organizar dados em grupos distintos, com base na similaridade entre os registros dentro de cada grupo. O objetivo principal dessa técnica é identificar conjuntos de dados que sejam relevantes e significativos, os quais podem ser usados diretamente em análises adicionais ou como características para modelos de regressão ou classificação. O método *K-means*, um dos mais conhecidos na área de agrupamento, continua sendo amplamente aplicado devido à sua simplicidade e capacidade de lidar com grandes volumes de dados.

Fávero e Belfiore (2024) enfatizam que a escolha das medidas de distância ou semelhança é crucial na análise de agrupamentos, uma vez que depende do tipo de variável em estudo (métrica ou binária). Após essa escolha, o pesquisador deve decidir entre diversos métodos de aglomeração, que podem ser hierárquicos ou não hierárquicos. Embora a tarefa de agrupar observações em clusters homogêneos possa parecer simples, a complexidade aumenta devido à diversidade de combinações possíveis entre as medidas e os métodos de aglomeração. Portanto, é fundamental que o pesquisador defina critérios claros para a alocação das observações nos grupos, fundamentados na teoria, nos objetivos da pesquisa e em sua própria experiência.

O algoritmo escolhido para a realização da análise de agrupamentos foi o *K-means*. Fávero e Belfiore (2024) explicam que os métodos de aglomeração não hierárquicos, como o *K-means*, requerem que o número de clusters seja estipulado previamente. A função objetivo do *K-means* é representada pela equação (2):

$$J = \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2 \quad (2)$$

Onde:

J é o valor total da função objetivo;

k é o número de clusters;

Cj é o conjunto de pontos de dados no cluster j;

xi é um ponto de dados no cluster *j*;

μ_j é o centroide do cluster *j*;

$||x_i - \mu_j||^2$ é a distância euclidiana ao quadrado entre o ponto *xi* e μ_j.

Com os dados normalizados e o algoritmo definido, foi necessário determinar o número ideal de clusters. Segundo Géron (2019), uma abordagem precisa para selecionar o número ideal de clusters é o uso do Método da Silhueta, que é a média do coeficiente de silhueta sobre todas as instâncias. Este coeficiente é calculado através da equação 3, onde a diferença entre a distância média para o cluster mais próximo (b) e a distância média para as outras instâncias no mesmo cluster (a), dividido pelo máximo entre (a) e (b). Essa métrica varia entre -1 e +1, onde valores próximos de +1 indicam instâncias bem dentro de seu próprio cluster, valores próximos de 0 indicam instâncias próximas aos limites do cluster e valores próximos de −1 indicam possíveis atribuições incorretas de cluster.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$

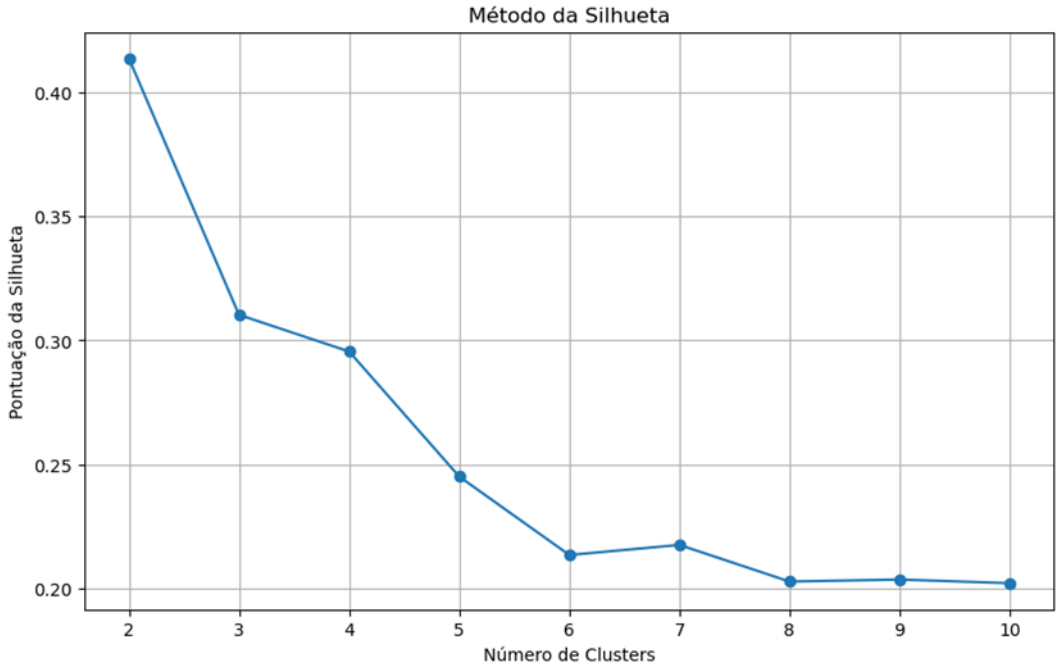
Onde:

a(i) é a distância média entre o ponto *i* e todos os pontos dentro do mesmo cluster;

b(i) é a distância média entre o ponto *i* e todos os pontos do cluster mais próximo ao qual *i* não pertence.

A Figura 1 apresenta os valores médios do coeficiente de silhueta para diferentes quantidades de clusters, auxiliando na escolha do número mais adequado para a análise

Figura 1 – Método da silhueta



Fonte: Elaborado pelo autor

Outro método amplamente utilizado para determinar o número ideal de clusters é o Método do Cotovelo. Segundo Nainggolan et al. (2019), o método envolve a análise da variação do Erro Quadrático Total (SSE) com diferentes valores de k, conforme representado pela equação 4. Com isso, identifica-se o ponto onde a taxa de diminuição do SSE se torna menos acentuada, formando um "cotovelo" no gráfico. A escolha do número de clusters é baseada na localização deste ponto de inflexão, onde a redução do SSE começa a desacelerar significativamente.

$$\mathrm{SSE}(k) = \sum_{i=1}^{n} \sum_{j=1}^{k} 1_{\{x_i \in C_j\}} ||x_i - \mu_j||^2 \quad (4)$$

Onde:

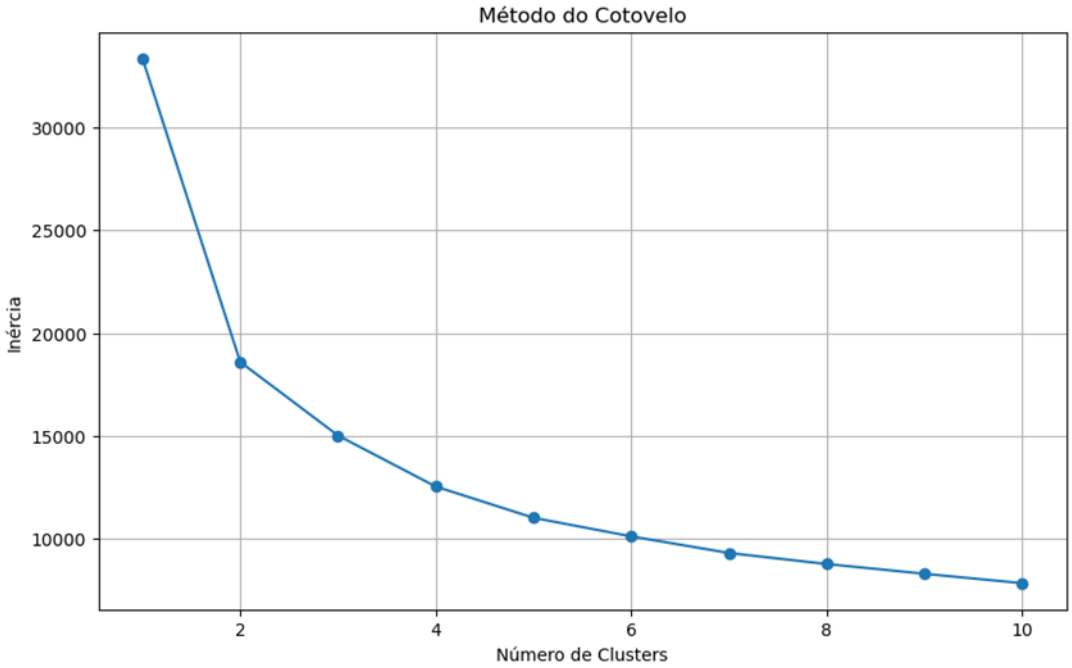
SSE(k) é a soma dos erros quadráticos para o número de clusters k;

xi é o ponto de dados i;

Cj é o conjunto de pontos de dados no cluster j;

$||\mathrm{x_i} - \mu_\mathrm{j}||^2$ é a distância euclidiana ao quadrado entre o ponto xi e μ_j.

Figura 1 – Método do Cotovelo



Fonte: Elaborado pelo autor

Com a definição de 3 clusters, o algoritmo K-means foi executado por meio da biblioteca Python Scikit-Learn (versão 1.5), agrupando os municípios em diferentes clusters e proporcionando uma abordagem coletiva na análise dos dados. Além disso, para investigar a relação linear entre as variáveis selecionadas, aplicou-se a correlação de Pearson. Essa técnica estatística quantifica o grau de associação linear entre duas variáveis métricas, apresentando valores que variam de -1 (correlação negativa perfeita) a 1 (correlação positiva perfeita), com 0 indicando a ausência de correlação linear (Montgomery & Runger, 2014).

Neste estudo, a correlação de Pearson foi utilizada para identificar quais variáveis apresentam forte associação, justificando, assim, a escolha dos indicadores para a formação dos clusters. Valores próximos a 1 ou -1 sugerem que as variáveis estão altamente correlacionadas e, possivelmente, podem ser combinadas em um único fator, enquanto valores próximos a 0 indicam uma relação fraca, o que corrobora a necessidade de considerá-las de forma individual no processo de agrupamento. A correlação de Pearson é calculada conforme a equação (5):

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \quad (5)$$

Onde:

r é o coeficiente de correlação;

n é o número de observações;

$\sum XY$ é a soma do produto de cada par de observações correspondentes das duas variáveis;

$\sum X$ é a soma das observações da primeira variável;

$\sum Y$ é a soma das observações da segunda variável;

$\sum X^2$ é a soma dos quadrados das observações da primeira variável;

$\sum Y^2$ é a soma dos quadrados das observações da segunda variável.

Para a criação de gráficos e trazer uma melhor visibilidade do resultado dos clusters, foi utilizada as bibliotecas Seaborn (Waskom, 2012) e Matplotlib (Hunter, 2002). Segundo McKinney (2018), a geração de visualizações informativas é uma etapa crucial na análise de dados, podendo ser parte do processo exploratório para identificar outliers ou necessidades de transformações nos dados, além de ser uma forma de gerar insights para modelos.

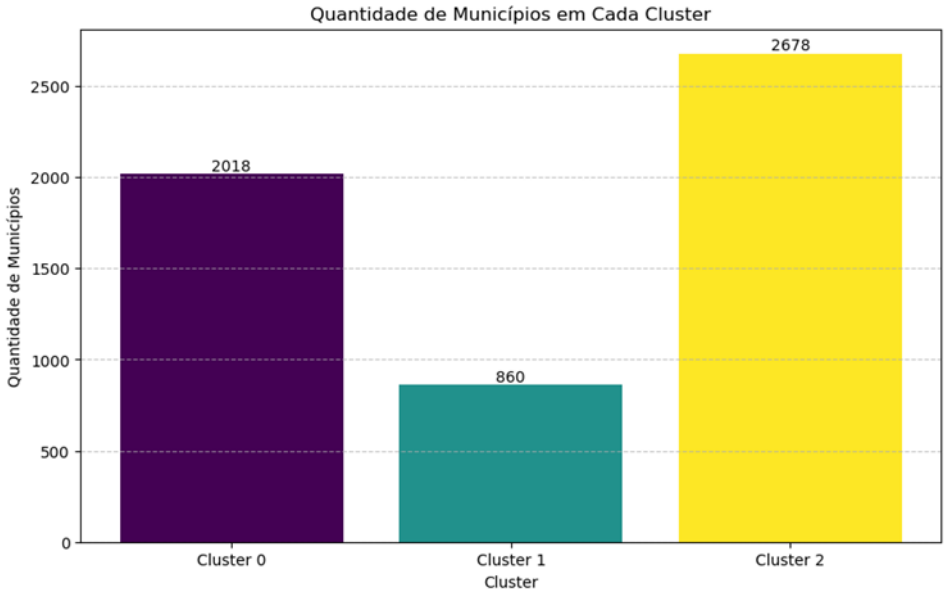
4. Resultados e Discussão

Após a exclusão dos municípios sem dados, a amostra consistiu em 5.556 municípios distribuídos em três clusters:

- Cluster 0: 2.018 municípios;
- Cluster 1: 860 municípios;
- Cluster 2: 2.678 municípios.

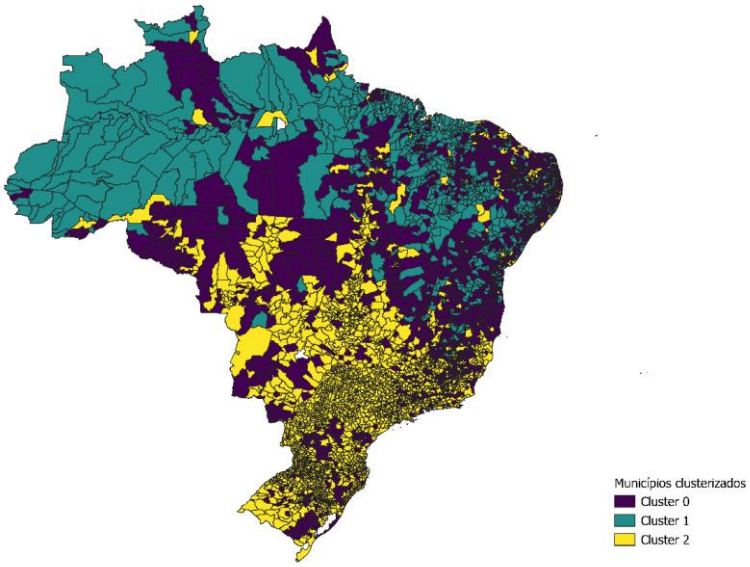
A Figura 3 apresenta essa divisão, enquanto a Figura 4 ilustra a distribuição geográfica dos clusters. Nota-se que o Cluster 2 predomina nas regiões Sul e Sudeste, sugerindo maior homogeneidade socioeconômica, enquanto os Clusters 0 e 1 concentram-se nas regiões Norte e Nordeste, evidenciando desigualdades históricas (Bação, Mazeron e Simões, 2023).

Figura 2 – Divisão de municípios por cluster



Fonte: Elaborado pelo autor

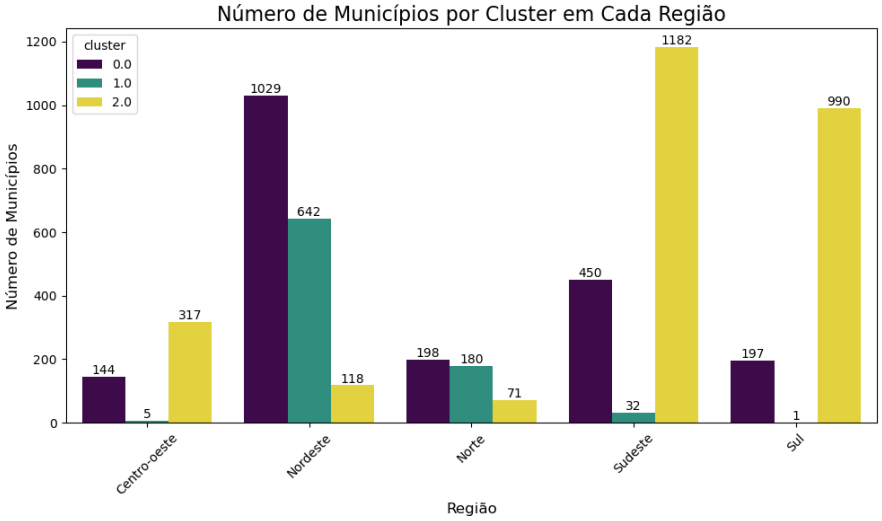
Figura 3 – Mapa do Brasil com os clusters



Fonte: Elaborado pelo autor

O gráfico de barras da Figura 5 detalha a distribuição dos clusters por região: o Cluster 0 é predominante no Nordeste, o Cluster 1 na região Norte, e o Cluster 2 no Sul e Sudeste. Essa segmentação reforça a necessidade de políticas regionais que considerem as particularidades locais.

Figura 4 – Número de Município em cada Cluster por Região



Fonte: Elaborado pelo autor

Os municípios se agruparem em clusters com números tão distintos evidencia a desigualdade regional no Brasil. Para identificar o padrão de cada cluster, foram gerados Boxplots de cada variável, para assim entender como estão postulados os municípios de cada cluster.

A análise dos indicadores por meio de boxplots (Figuras 6) revela o seguinte padrão:

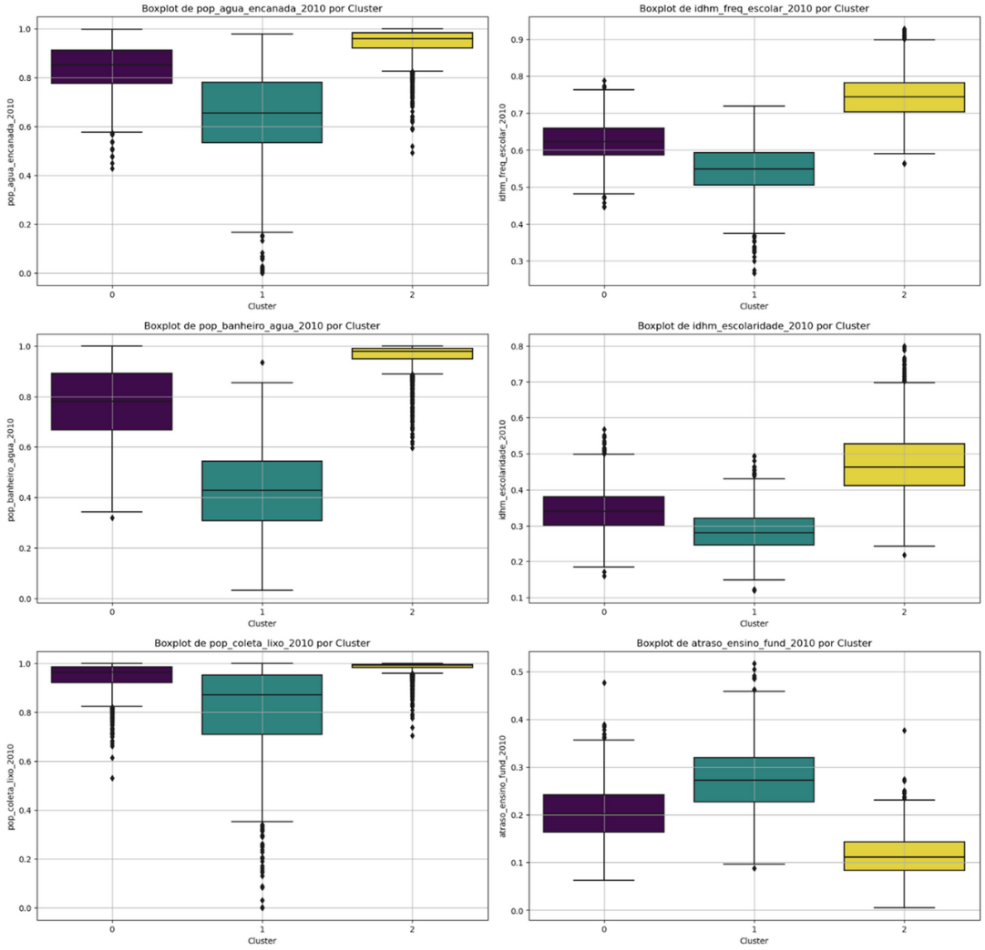
Saneamento:

- O Cluster 2 concentra municípios com elevado acesso à água encanada e coleta de lixo, apresentando baixa variabilidade;
- O Cluster 1 agrupa municípios com os índices mais baixos e maior dispersão, indicando precariedade em infraestrutura;
- O Cluster 0 exibe valores intermediários e variabilidade acentuada.

Educação:

- Municípios do Cluster 2 apresentam melhores índices de frequência escolar, maior percentual de ensino fundamental completo e menores níveis de atraso escolar;
- O Cluster 1 tem os piores indicadores educacionais, enquanto o Cluster 0 fica em uma posição intermediária.

Figura 6 – Boxplots das variáveis analisadas por Cluster



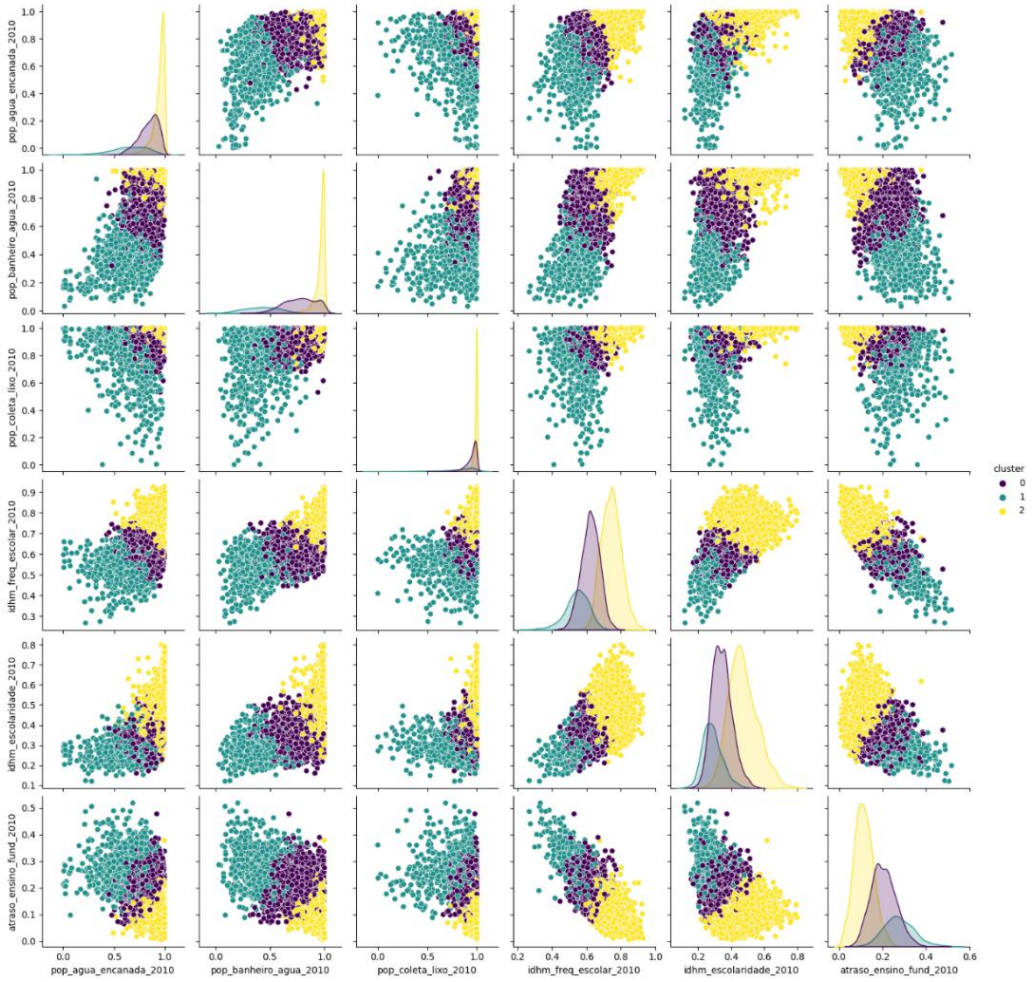
Fonte: Elaborado pelo autor

A análise das disparidades entre os municípios brasileiros tornou-se evidente ao agrupar os municípios em clusters com base em variáveis como o acesso à água encanada e acesso à educação de base. Essa organização permitiu identificar com clareza as diferenças significativas nos níveis de infraestrutura básica e no desenvolvimento educacional entre as regiões.

Municípios com melhor infraestrutura de saneamento tendem a apresentar também indicadores educacionais mais elevados, enquanto aqueles com acesso limitado a esses serviços estão associados a menores níveis de escolaridade, reforçando a interdependência entre saneamento e educação como fatores determinantes para o desenvolvimento socioeconômico.

A Figura 7 traz o gráfico de pares entre as variáveis, o qual revela uma forte associação entre saneamento e educação. Esse resultado evidencia que os municípios com melhor infraestrutura de saneamento tendem a apresentar também melhores indicadores educacionais, reforçando a interdependência entre o acesso a serviços básicos e o desempenho escolar sendo possível identificar a distribuição dos três grupos de municípios.

Figura 7 – Gráfico de pares entre as variáveis analisadas



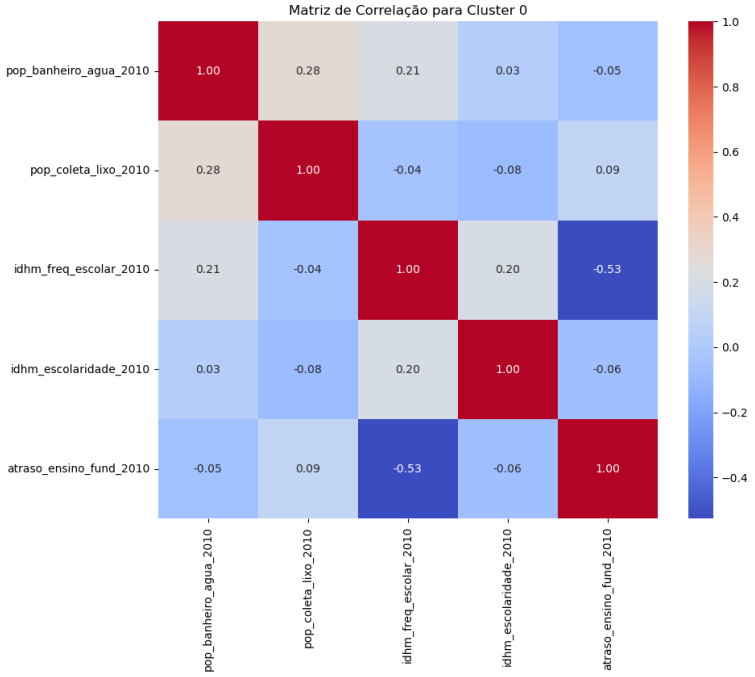
Fonte: Elaborado pelo autor

Ao aplicar a matriz de correlação de Pearson das variáveis nos clusters, é possível verificar relações lineares dentro de cada cluster, com isso pode-se identificar como as variáveis estão correlacionadas dentro de cada agrupamento.

Dentro do Cluster 0, conforme a Figura 8, pode-se observar uma correlação baixa, porém positiva entre a população com acesso a água encanada e banheiro com coleta de lixo. Destaca-

se a correlação negativa entre frequência escolar e atraso no ensino fundamental, mostrando que a frequência escolar é um fator relevante para evitar reprovação e atraso escolar.

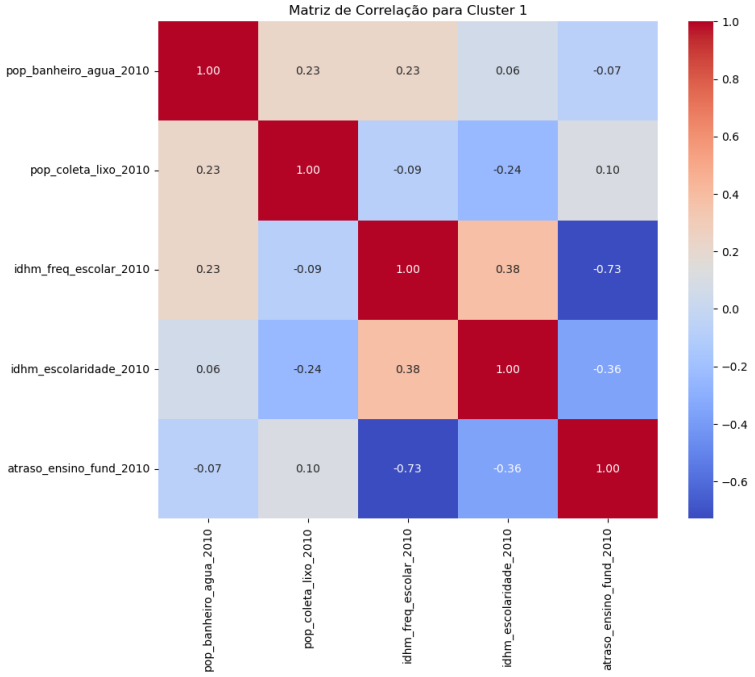
Figura 8 – Matriz de correlação para o Cluster 0



Fonte: Elaborado pelo autor

No Cluster 1, conforme a Figura 9, observa-se uma correlação negativa ainda maior do que no Cluster 0 entre a frequência escolar e o atraso no ensino fundamental, mais uma vez apontando que a frequência escolar desempenha um papel importante para evitar o atraso escolar.

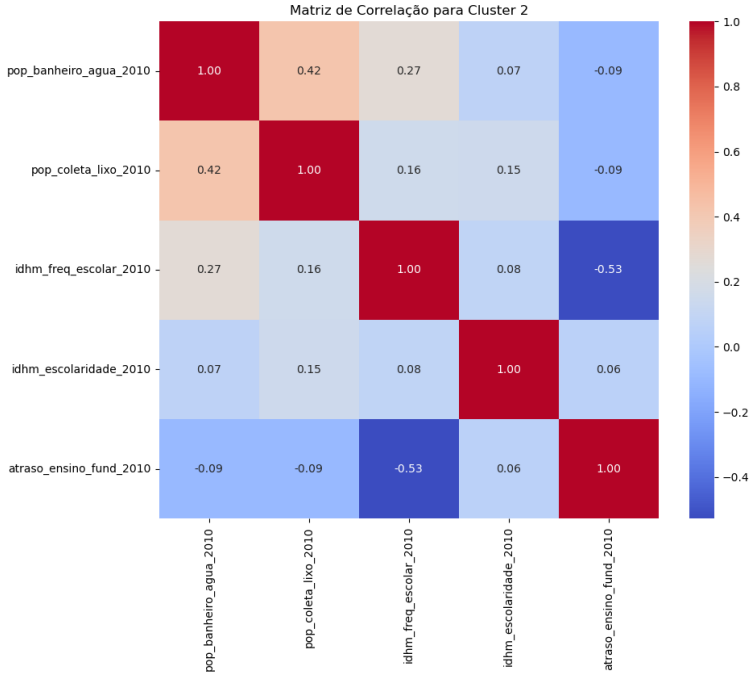
Figura 9 – Matriz de Correlação para o Cluster 1



Fonte: Elaborado pelo autor

Dentro do Cluster 2, conforme a Figura 10, observou-se novamente a alta correlação negativa entre frequência escolar e atraso no ensino fundamental. Também ocorre uma maior correlação entre a coleta de lixo e o acesso a água encanada e banheiro na residência. Porém o que destoou mais no Cluster 2 em relação aos outros é a presença de correlação positiva entre o acesso a coleta de lixo e a escolaridade.

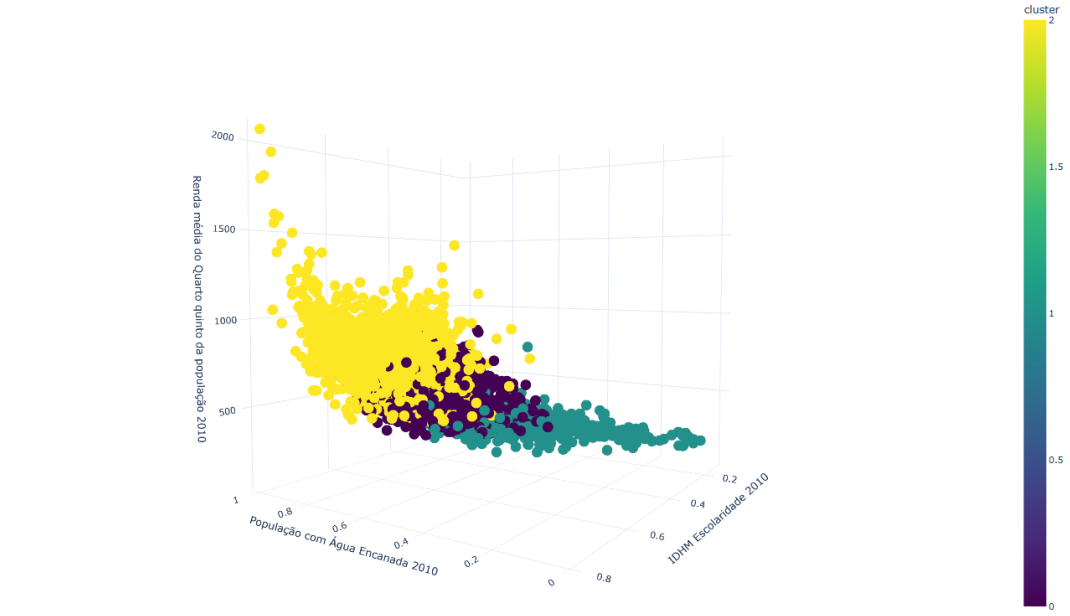
Figura 10 – Matriz de Correlação para o Cluster 2



Fonte: Elaborado pelo autor

Para complementar a análise, foi incluída uma variável de renda — definida como a renda do quarto quinto mais pobre da população — conforme ilustrado na Figura 11. Os resultados indicam que municípios com renda média mais elevada tendem a apresentar melhores condições de saneamento e índices educacionais superiores, corroborando a hipótese de que renda, saneamento e escolaridade estão positivamente correlacionados.

Figura 11 – Relação entre Saneamento, Escolaridade e Renda (2010)



Fonte: Elaborado pelo autor

Em síntese, a aplicação da análise de clusters permitiu segmentar os municípios de forma eficaz, identificando padrões que podem orientar a implementação de políticas públicas direcionadas para reduzir as disparidades regionais. Embora os dados do Censo de 2010 e a seleção específica de variáveis representem limitações, os achados fornecem uma base robusta para a compreensão das interações entre infraestrutura básica e desenvolvimento educacional, aspectos cruciais para a Engenharia de Produção na formulação de estratégias integradas de desenvolvimento regional.

5. Considerações Finais

A análise de clusters aplicada aos municípios brasileiros revelou padrões significativos de desigualdade no acesso a saneamento e educação, evidenciando as disparidades regionais do país. Os resultados demonstraram que os municípios agrupados no Cluster 2, que apresentam elevados índices de desenvolvimento, possuem melhores indicadores de acesso a água encanada, coleta de lixo e escolaridade. Em contrapartida, o Cluster 1 concentra os municípios mais vulneráveis, com baixos indicadores de infraestrutura básica e desempenho educacional, enquanto o Cluster 0 abrange localidades em condições intermediárias, com maior variabilidade nos dados analisados.

Os achados deste estudo evidenciam uma forte correlação positiva entre saneamento e educação, ressaltando a interdependência desses fatores no desenvolvimento regional. A

inclusão da variável de renda reforçou a influência dos fatores econômicos na melhoria dos indicadores de saneamento e educacionais, confirmando que municípios com maior renda média tendem a oferecer melhores condições aos seus cidadãos.

Diante disso, conclui-se que os objetivos deste estudo foram amplamente atendidos, uma vez que foi possível identificar a existência de grupos de municípios com características diferenciadas, possibilitando a formulação de recomendações direcionadas. Em particular, os municípios pertencentes ao Cluster 1 demandam intervenções específicas voltadas à ampliação e melhoria dos serviços básicos, o que pode, consequentemente, contribuir para a elevação dos índices educacionais e reduzir as desigualdades regionais.

Como continuidade deste trabalho, sugerem-se as seguintes direções para futuras pesquisas:

- A incorporação de dados mais recentes, como os do Censo de 2022, para uma análise temporal que permita acompanhar a evolução dos indicadores de saneamento e educação;
- A inclusão de variáveis socioeconômicas adicionais, como saúde, acesso à tecnologia e emprego, para uma compreensão mais abrangente das disparidades regionais;
- A aplicação de métodos avançados de machine learning, como redes neurais, que possam refinar a segmentação dos municípios e revelar padrões mais complexos de inter-relação entre os indicadores;
- A realização de simulações preditivas para avaliar o impacto potencial de políticas públicas integradas na melhoria dos serviços básicos e, consequentemente, nos indicadores educacionais.

Essas propostas podem fornecer subsídios valiosos para a elaboração de intervenções governamentais mais eficazes, promovendo um desenvolvimento socioeconômico mais equilibrado e uma distribuição equitativa de oportunidades em todo o território brasileiro.

REFERÊNCIAS

ABANYIE, S. K.; AMUAH, E. E. Y.; DOUTI, N. B.; OWUSU, G.; AMADU, C. C.; et al. **WASH in Selected Basic Schools and Possible Implications on Health and Academics: An Example of the Wa Municipality of Ghana, West Africa.** *American Journal of Environmental Science and Engineering*, v. 5, n. 1, p. 15-20, 2021. DOI: 10.11648/j.ajese.20210501.13.

AGGARWAL, C. C.; REDDY, C. K. **Data Clustering: Algorithms and Applications.** 1. ed. Boca Raton: CRC Press, 2013.

ATLAS DO DESENVOLVIMENTO HUMANO NO BRASIL. **Elaboração: Atlas do Desenvolvimento Humano no Brasil.** PNUD Brasil, IPEA e FJP, 2022. Fontes: dados do IBGE e de registros administrativos. Disponível em: http://atlasbrasil.org.br/acervo/biblioteca. Acesso em: 07 set. 2024.

BAÇÃO, P.; MAZERON, D.; SIMÕES, M. **The Financialization of Health and Education and Inequality in Twenty-First Century Brazil.** *Latin American Perspectives*, v. 50, n. 5, p. 47-66, 2023.

BRASIL. **Constituição da República Federativa do Brasil.** Brasília, DF: Presidência da República, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 16 set. 2024.

BRASIL. **Proposta de Emenda à Constituição nº 2, de 2016.** Altera o art. 206 da Constituição Federal. Brasília, 2016. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/124779. Acesso em: 16 set. 2024.

BRASIL. **Lei nº 14.026, de 15 de julho de 2020.** Disponível em: https://normas.leg.br/?urn=urn:lex:br:federal:lei:2020-07-15;14026. Acesso em: 10 jun. 2024.

BRUCE, A.; BRUCE, P. **Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais.** Rio de Janeiro, RJ: Alta Books, 2019.

COURNAPEAU, D. **Scikit-learn.** Disponível em: https://scikit-learn.org/stable/. Acesso em: 10 jun. 2024.

FÁVERO, L. P.; BELFIORE, P. **Manual de Análise de Dados.** 2. ed. Brasil: GEN LTC, 2024.

GÉRON, A. **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.** Sebastopol, CA: O'Reilly Media, Inc., 2019.

HUNTER, J. D. **Matplotlib.** Disponível em: https://matplotlib.org/. Acesso em: 10 jun. 2024.

IBGE. COORDENAÇÃO DE ESTRUTURAS TERRITORIAIS. **Malha Municipal Digital e Áreas Territoriais 2023: Notas Metodológicas 01/2024.** Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?=&t=acesso-ao-produto. Acesso em: 16 set. 2024.

INSTITUTO TRATA BRASIL. **Guia do Saneamento 2023.** São Paulo: [s.n.], 2023. Disponível em: https://tratabrasil.org.br/wp-content/uploads/2024/04/Guia-do-Saneamento-2023_V20_12.11_Digital.pdf. Acesso em: 16 set. 2024.

LARSON, R.; FARBER, B. **Estatística Aplicada.** Tradução de José Fernando Pereira Gonçalves; revisão técnica de Manoel Henrique Salgado. São Paulo, SP: Pearson Education do Brasil, 2015.

MCKINNEY, W. W. **Pandas.** Disponível em: https://pandas.pydata.org/. Acesso em: 10 jun. 2024.

MCKINNEY, W. W. **Python para Análise de Dados.** 2. ed. São Paulo, SP: Pearson Education do Brasil, 2018.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros.** 6. ed. Wiley, 2014.

NAINGGOLAN, R.; et al. **Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) Optimized by Using the Elbow Method.** *Journal of Physics: Conference Series*, v. 1361, p. 012015, 2019. IOP Publishing. DOI: 10.1088/1742-6596/1361/1/012015.

OECD. **How's Life? 2020: Measuring Well-being.** Paris: OECD Publishing, 2020. Disponível em: https://doi.org/10.1787/9870c393-en. Acesso em: 16 set. 2024.

PRAKASH, K. B. **Data Science Handbook: A Practical Approach.** John Wiley & Sons, 2022. DOI: 10.1002/9781119858010.

REITZ, K.; SCHLUSSER, T. **O Guia do Mochileiro Python: Melhores Práticas Para Desenvolvimento.** Brasil: Novatec Editora, 2017.

SHARMA, M. K.; ADHIKARI, R.; KHANAL, S. P.; ACHARYA, D.; TEIJLINGEN, E. v. **Do School Water, Sanitation, and Hygiene Facilities Affect Students' Health Status, Attendance, and Educational Achievements?** *Health Science Reports*, v. 7, p. e2293, 2024. DOI: 10.1002/hsr2.2293.

VALENCIO, N. A.; BAPTISTA, M. S. **The Interface of Disasters, Sanitation, and Poverty in Brazil: A Sociological Perspective.** *Frontiers in Sustainable Cities*, v. 5, p. 1184532, 2023. DOI: 10.3389/frsc.2023.1184532.

WASKOM, M. **Seaborn.** Disponível em: https://seaborn.pydata.org/. Acesso em: 10 jun. 2024.