



ConBRepro

XIV CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO

I CIGELUBRA

Congresso Internacional de Gestão e Engenharia Luso-Brasileiro

04 a 06 de dezembro de 2024

Energias Limpas nas Engenharias

Segmentação Socioeconômica dos Municípios de São Paulo Utilizando *k-means*: Uma Abordagem de Clusterização para o Desenvolvimento Regional

Julia Agatha Carnieri

Universidade Federal do Paraná

Carla Regina Mazia Rosa

Universidade Federal do Paraná

Marcos Augusto Mendes Marques

Universidade Federal do Paraná

Mariana Kleina

Universidade Federal do Paraná

Resumo: Atualmente, novos tipos de divisões de estados estão sendo explorados, refletindo mudanças demográficas, econômicas e culturais que desafiam as divisões políticas tradicionais. Este estudo investiga a segmentação socioeconômica dos 645 municípios do estado de São Paulo por meio da técnica de clusterização *k-means*. O objetivo principal do projeto é identificar padrões e agrupamentos baseados em variáveis como população, Produto Interno Bruto (PIB) per capita, setores econômicos e localização geográfica. Os dados foram coletados de fontes como o Instituto Brasileiro de Geografia e Estatística (IBGE) e o SENATRAN, sendo preparados e analisados no ambiente estatístico R. Métodos de análise exploratória, incluindo correlação e detecção de *outliers* foram utilizados para preparar os dados antes da aplicação do algoritmo *k-means*. A determinação do número ideal de clusters foi realizada por meio dos métodos do cotovelo e da silhueta, resultando na identificação de sete clusters como ideal para a segmentação. Os resultados mostraram clusters heterogêneos em termos de perfil demográfico e econômico, destacando desde áreas urbanas altamente industrializadas até municípios rurais com economia agrícola predominante. A análise de silhueta indicou uma boa separação entre os clusters, reforçando a validade da segmentação obtida. As conclusões destacam a utilidade da clusterização para entender as dinâmicas regionais e a factibilidade de personalizar agrupamentos regionais para apoiar políticas públicas direcionadas ao desenvolvimento econômico sustentável e equitativo de uma região, no caso, o estado de São Paulo.

Palavras-chave: Segmentação socioeconômica, São Paulo, clusterização, *k-means*, desenvolvimento regional.

Socioeconomic Segmentation of São Paulo Municipalities Using *K-means*: A Clustering Approach for Regional Development

Abstract: Currently, new state divisions are being explored, reflecting demographic, economic, and cultural changes that challenge traditional political boundaries. This study investigates the socioeconomic segmentation of the 645 municipalities in the state of São Paulo using the *k-means* clustering technique. The main objective is to identify patterns and groupings based on variables such as population, Gross Domestic Product (GDP) per capita, economic sectors, and geographic

location. Data were collected from sources like the Brazilian Institute of Geography and Statistics (IBGE) and SENATRAN, and analyzed using the R statistical environment. Exploratory data analysis methods, including correlation and outlier detection, were employed to prepare the data before applying the *k-means* algorithm. The optimal number of clusters was determined through the elbow and silhouette methods, resulting in seven clusters being identified as the ideal segmentation. The results revealed heterogeneous clusters in terms of demographic and economic profiles, ranging from highly industrialized urban areas to rural municipalities with predominantly agricultural economies. The silhouette analysis indicated good separation between the clusters, reinforcing the validity of the segmentation obtained. The findings highlight the usefulness of clustering for understanding regional dynamics and the feasibility of customizing regional groupings to support public policies aimed at sustainable and equitable economic development in São Paulo.

Keywords: Socioeconomic segmentation, São Paulo, *k-means*, clustering, regional development.

1. Introdução

A globalização nas últimas décadas acelerou a reconfiguração espacial das atividades econômicas, resultando em uma transformação significativa em algumas regiões, enquanto outras enfrentaram estagnação ou declínio (HISSA, 2015). Esse processo é impulsionado pelo capital financeiro desregulado, políticas públicas mal planejadas e generalização nas tomadas de decisão (SANTOS, 2023). As disparidades regionais e sociais se acentuaram à medida que certas áreas se tornaram plataformas produtivas de grandes empresas, enquanto outras regiões são frequentemente deixadas à margem da dinâmica econômica global, intensificando os fluxos migratórios e as tensões sociais (HISSA, 2015; SANTOS, 2023).

No Brasil, segundo o governo federal, o estado de São Paulo destaca-se como o maior centro econômico do país, responsável por cerca de 30,18% do Produto Interno Bruto (PIB), com um valor de 2.719.751 bilhões de reais (IBGE, 2022). A diversidade socioeconômica e demográfica de São Paulo, com 645 municípios e uma densidade populacional de 178,92 habitantes por quilômetro quadrado, significativamente superior à média nacional de 23,86 habitantes por quilômetro quadrado, reflete as diferentes trajetórias de desenvolvimento dentro do estado (IBGE, 2022).

Neste cenário, a técnica de clusterização *k-means*, amplamente utilizada para agrupar entidades com base em características observadas, é uma ferramenta eficiente para identificar padrões (SANTOS et. al., 2020). Este estudo aplica essa metodologia para segmentar os municípios do estado de São Paulo, com o objetivo de identificar agrupamentos que forneçam subsídios mais específicos e direcionados para a formulação de políticas públicas, indo além das divisões políticas tradicionais. A escolha do estado de São Paulo se justifica por sua relevância socioeconômica no contexto nacional, e os resultados obtidos podem ser generalizados e aplicados a outros estados e regiões, proporcionando insights quanto a análise de dinâmicas regionais em diferentes contextos.

A escolha do método *k-means* é justificada por sua eficiência, escalabilidade, pela capacidade de identificar padrões de similaridade entre os municípios e facilidade de implementação. Para avaliar a qualidade dos agrupamentos, foram aplicados métodos como o cotovelo e a análise de silhueta.

A pergunta de pesquisa que orienta este estudo é: Como variáveis socioeconômicas e geográficas podem influenciar o agrupamento eficiente dos municípios no estado de São Paulo utilizando a técnica de clusterização *k-means*? Ao responder à questão, espera-se contribuir tanto para o entendimento das disparidades regionais em São Paulo quanto demonstrar a factibilidade de se construir agrupamentos específicos para análises direcionadas à gestão pública, capazes de fornecer informações relevantes para políticas de desenvolvimento regional.

2. Revisão de Literatura

2.1 Clusterização: Conceitos Fundamentais

Clusterização é uma técnica amplamente utilizada para agrupar objetos com características similares em grupos denominados clusters, de acordo com informações observadas nos dados (EVERITT et al., 2011). A clusterização encontra aplicações em diversas áreas, incluindo geografia, ciência de dados, planejamento urbano e economia regional (HISSA, 2015). Ela tem o objetivo de identificar padrões ocultos e estruturas subjacentes, que frequentemente não são aparentes à primeira vista, facilitando a segmentação e análise de fenômenos complexos, permitindo previsões e otimizações de processos em contextos variados, desde análises de mercado até planejamentos regionais.

Conforme observado por Kassambara (2017), a clusterização é um método de aprendizado de máquina não supervisionado, que não depende de uma variável preditora previamente definida. O processo visa inferir padrões ocultos nos dados não rotulados. Existem diferentes tipos de algoritmos de clusterização, como os métodos particionais, hierárquicos, baseados em densidade, em grade e em modelos (Mount e Zumel, 2019). Cada um desses métodos é empregado conforme a natureza dos dados e os objetivos da análise, sendo útil para revelar estruturas que, de outra forma, poderiam passar despercebidas.

Entre os métodos de clusterização, os chamados algoritmos particionais são reconhecidos pela simplicidade de sua implementação. Kaufman e Rousseeuw (2009) observam que esses algoritmos dividem os dados em k clusters distintos, com base em medidas de similaridade ou dissimilaridade, onde cada grupo deve conter pelo menos um objeto, e cada objeto é alocado ao grupo mais próximo, geralmente com base em uma métrica de distância.

2.2 Métodos de Clusterização *k-means*

O algoritmo *k-means* é reconhecido e utilizado na análise de clusterização devido à sua eficiência computacional e simplicidade. Proposto por MacQueen (1967), o *k-means* visa particionar um conjunto de dados em k clusters de forma que a variância intra-cluster seja minimizada. Inicialmente, o algoritmo seleciona aleatoriamente k centroides, que atuam como centros iniciais dos clusters. Cada objeto do conjunto de dados é então atribuído ao cluster cujo centroide está mais próximo, utilizando geralmente a distância euclidiana como métrica de proximidade.

Durante a etapa de atribuição, todos os objetos são agrupados nos clusters com base nas distâncias deles até os centroides. Após essa atribuição inicial, os centroides são recalculados como a média dos objetos pertencentes a cada cluster. Esse processo de atribuição e atualização dos centroides é repetido iterativamente até que não haja mais mudanças significativas nas atribuições ou até que um critério de convergência seja alcançado, como um número máximo de iterações (MACQUEEN, 1967).

Ao final do processo, obtém-se um conjunto de clusters onde os objetos dentro de cada cluster são mais similares entre si do que com os objetos de outros clusters. O algoritmo *k-means* é amplamente utilizado em aplicações que requerem segmentação rápida de grandes conjuntos de dados, como análises geoespaciais, agrupamento de consumidores e segmentação de mercado. No entanto, é relevante mencionar que o desempenho do *k-means* pode ser influenciado pela escolha inicial dos centroides, do número de clusters

utilizado e pelos métodos empregados, tais quais o método do cotovelo e o da silhueta (THORNDIKE, 1953; ROUSSEEUW, 1987).

2.4 Método do Cotovelo

O Método do Cotovelo, também conhecido como *Elbow Method*, é amplamente utilizado na análise de clusterização para determinar o número ideal de clusters, k . Proposto inicialmente por Thorndike (1953), esse método traça a soma dos erros quadráticos (SSE – *Sum of Squared Errors*) em função do número de clusters. A SSE mede a variabilidade dentro dos clusters; quanto menor o SSE, mais homogêneos são os grupos. No gráfico gerado, o objetivo é identificar o ponto onde a redução da SSE torna-se marginal, ou seja, onde a adição de mais clusters não resulta em uma diminuição significativa do erro. Esse ponto de inflexão é visualmente identificado como o "cotovelo" da curva, indicando o número de clusters mais apropriado.

Segundo Thorndike, o objetivo da *cluster analysis* é dividir um conjunto de dados em grupos que compartilham características similares, maximizando a homogeneidade dentro de cada cluster e a heterogeneidade entre eles. O Método do Cotovelo oferece uma abordagem objetiva para a escolha de k , ajudando a evitar a supersegmentação, em que muitos clusters são formados, ou a subsegmentação, em que poucos clusters ocultam importantes nuances dos dados.

Esse método é frequentemente aplicado em conjunto com algoritmos como 12 *k-means* e outras técnicas de clusterização, contribuindo para uma escolha fundamentada do número de clusters a ser utilizado. Ao reduzir a subjetividade na determinação de k , o Método do Cotovelo melhora a confiabilidade e a interpretabilidade dos resultados em análises que envolvem dados espaciais, demográficos ou socioeconômicos.

2.5 Método da Silhueta

O Método da Silhueta, introduzido por Rousseeuw (1987), é uma métrica empregada na análise de clusterização para avaliar a qualidade dos agrupamentos gerados pelo algoritmo *k-means* e outros métodos de clusterização. Ele atribui um valor de silhueta a cada objeto, quantificando o quão similar o objeto é ao seu próprio cluster em comparação com outros. O cálculo é feito para cada objeto i dentre os m elementos do conjunto, $i = 1, \dots, m$:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \quad (1)$$

Os valores de silhueta variam de -1 a +1, onde valores próximos de +1 indicam que o objeto está bem ajustado ao seu cluster, enquanto valores próximos de -1 sugerem que o objeto pode ter sido atribuído ao cluster errado (ROUSSEEUW, 1987). Essa métrica proporciona uma análise da estrutura dos clusters, auxiliando os pesquisadores na interpretação e validação da consistência dos agrupamentos formados.

Segundo Rousseeuw, ao utilizar o Método da Silhueta, além de avaliar a qualidade dos clusters formados, os pesquisadores também podem ajustar parâmetros, como o número de clusters (k), no algoritmo *k-means*, otimizando a segmentação.

2.6 Aplicações em Geografia e Desenvolvimento Regional

A clusterização é um potencial aliado na análise espacial e no desenvolvimento regional. Em contextos geográficos, ela permite identificar padrões naturais em dados como distribuição populacional e características socioeconômicas, o que facilita a formulação de políticas públicas e estratégias de desenvolvimento localizadas e eficazes (STORPER; SCOTT, 2009; ZHANG et al., 2024). Além disso, a clusterização segmenta áreas geográficas que compartilham características socioeconômicas e demográficas similares, o que pode tornar a alocação de recursos públicos eficiente e identificar necessidades específicas de desenvolvimento.

Ao contrário das divisões administrativas convencionais, como mesorregiões e regiões administrativas existentes, que são definidas por critérios político-administrativos e úteis para o planejamento em nível estadual, a clusterização baseia-se em critérios socioeconômicos e geográficos dos municípios ou regiões. A consideração destes dados permite identificar agrupamentos que compartilham características semelhantes, mesmo que ultrapassem fronteiras administrativas convencionais, capturando dinâmicas socioeconômicas e interações entre diferentes áreas de forma mais completa (ZHANG et al., 2024).

Ao aplicar técnicas como o *k-means* aos dados de São Paulo, é possível identificar clusters que representam as características locais de maneira mais precisa em comparação a divisões que consideram apenas o parâmetro geográfico, uma vez que se ateste a correlação das variáveis analisadas. Isso pode ser particularmente útil para projetos que exigem análise socioeconômica, fornecendo uma base para estratégias de desenvolvimento regional adaptadas e eficazes. Esses métodos podem auxiliar no planejamento urbano e rural, identificando áreas adequadas para crescimento urbano ou expansão agrícola, além de orientar investimentos específicos para setores industriais ou agrícolas, estimulando a criação de empregos locais (SCOTT; STORPER, 2009). Adicionalmente, essa abordagem facilita a alocação eficiente de recursos para infraestrutura básica, como transporte e serviços públicos, atendendo às diversas necessidades de cada agrupamento de municípios.

3. Metodologia

Este artigo adota uma abordagem aplicada, utilizando análise quantitativa para investigar padrões socioeconômicos e geográficos dos 645 municípios de São Paulo. A pesquisa é descritiva, visando delinear e interpretar agrupamentos de municípios com base em características socioeconômicas escolhidas para ilustrar um cenário específico. O processo metodológico segue uma sequência lógica que inclui coleta de dados, pré-processamento, análise exploratória, aplicação de técnicas de clusterização, avaliação dos resultados e interpretação dos agrupamentos.

3.1 Coleta e Preparação de Dados

A seleção das variáveis foi realizada de maneira ilustrativa, com o intuito de demonstrar um exemplo de agrupamento potencialmente relevante para o estado de São Paulo. Reconhecendo a diversidade de possíveis abordagens, futuras pesquisas podem explorar diferentes combinações de variáveis, conforme as necessidades específicas de pesquisadores, municípios ou outras partes interessadas. Para assegurar a integridade e a confiabilidade da análise, os dados foram obtidos de fontes amplamente reconhecidas. Embora as fontes de dados apresentem diferentes anos de referência, optou-se por utilizar as versões mais recentes disponíveis, preservando a atualidade da análise.

Os dados populacionais dos municípios de São Paulo foram extraídos do “Censo Demográfico 2022, com Malha Territorial 2023”, que fornece as delimitações municipais e os dados demográficos mais atualizados. Informações sobre a frota de veículos por município foram obtidas do sistema do SENATRAN, com ano de referência em 2024, garantindo a atualização mais recente dessas variáveis. As coordenadas geográficas dos 645 municípios de São Paulo foram coletadas diretamente do Google Maps, também em 2024, assegurando a precisão geoespacial necessária. Por fim, os indicadores econômicos, como o PIB per capita, o Valor Adicionado Bruto da Agropecuária e o Valor Adicionado Bruto da Indústria, foram extraídos do IBGE, com dados atualizados até 2021, conforme as diretrizes das Contas Nacionais de 2010.

A coleta de dados de fontes distintas e com diferentes anos de referência reflete a variação natural no ritmo de atualização dessas bases de dados. Cada conjunto de dados foi selecionado com base em sua relevância e confiabilidade no contexto desta pesquisa. Todos os dados foram convertidos para um formato compatível com o ambiente estatístico R, onde foram processados e analisados. Durante a preparação, duplicatas foram removidas para assegurar a consistência dos dados, e os nomes dos municípios foram transformados em variáveis categóricas, utilizando uma codificação numérica ordinal de 1 a 645.

3.2 Análise Exploratória de Dados

Com o objetivo de compreender a estrutura e as inter-relações entre as variáveis socioeconômicas dos 645 municípios do estado de São Paulo, foi realizada uma análise exploratória de dados para identificar redundâncias e otimizar o esforço computacional. Utilizando o ambiente estatístico R, juntamente com a biblioteca *corrplot* para a visualização gráfica da matriz de correlação, aplicou-se o método de correlação a fim de examinar a intensidade e a direção das associações entre as variáveis. Variáveis consideradas redundantes foram excluídas, reduzindo a multicolinearidade e simplificando o modelo.

Além disso, foi aplicada uma transformação logarítmica às variáveis com valores não negativos (como população, PIB per capita, valor adicionado bruto da agropecuária e valor adicionado bruto da indústria) para suavizar a variação entre os dados e reduzir a influência de valores extremos. A fórmula utilizada para essa transformação foi:

$$\log(X) = \log(X + 1) \quad (2)$$

onde X representa o valor original da variável. Esse ajuste é necessário para evitar problemas com valores nulos e negativos, garantindo que todos os valores sejam válidos para o cálculo do logaritmo.

Após a transformação logarítmica, todos os dados numéricos passaram pelo processo de normalização *Z-score*, utilizando a função *scale()* do R, que calcula a padronização dos dados em relação à média e ao desvio padrão da população. Esse processo garante que todas as variáveis tenham média zero e desvio padrão unitário, padronizando as escalas e evitando que variáveis com diferentes magnitudes influenciem de maneira desproporcional a análise. A partir destes resultados, a clusterização seguiu a fim de manter a representatividade da amostra estudada.

3.3 Determinação do Número Ideal de Clusters

Para definir o número ideal de clusters do estado de São Paulo, foram aplicados métodos consagrados de análise, assegurando uma estrutura consistente de agrupamento. O algoritmo *k-means*, conhecido por sua eficiência na segmentação de grandes volumes de

dados, foi o método escolhido. O processo de determinação do número de clusters utilizou duas abordagens: o método do cotovelo e o coeficiente de silhueta.

O método do cotovelo foi aplicado inicialmente. Esse método consiste em calcular a soma das distâncias quadradas internas dentro dos clusters (*within-cluster sum of squares*) para diferentes valores de k , variando de 1 a 12. O ponto em que a redução dessa soma começa a se estabilizar no gráfico, formando o que visualmente se assemelha a um “cotovelo”, indica o número ideal de clusters. Esse valor sugere o momento em que a inclusão de mais clusters deixa de reduzir significativamente a variabilidade interna, equilibrando complexidade e qualidade dos agrupamentos.

Adicionalmente, utilizou-se o coeficiente de silhueta, uma métrica que avalia tanto a coesão quanto a separação dos clusters, fornecendo uma indicação quantitativa da qualidade dos agrupamentos. O número de clusters que maximiza o valor médio do coeficiente de silhueta foi selecionado como o mais adequado.

3.4 Utilização do Algoritmo e *k-means*

Após a determinação do número ideal de clusters, o algoritmo e *k-means* foi implementado para segmentar os 645 municípios do estado de São Paulo em grupos homogêneos. Inicialmente, os centroides dos clusters foram atribuídos aleatoriamente, e cada município foi associado ao cluster cujo centroide estivesse mais próximo, utilizando a distância euclidiana como métrica.

Com a atribuição inicial concluída, os centroides foram recalculados com base na média das variáveis de todos os municípios pertencentes a cada cluster. Esse processo de atribuição e atualização dos centroides foi repetido iterativamente, refinando progressivamente os grupos, até que a variação na posição dos centroides fosse mínima, ou o número máximo de 100 iterações fosse atingido. Essa abordagem garantiu a minimização da variabilidade dentro de cada cluster e a maximização da separação entre os grupos.

O algoritmo e *k-means* foi executado no ambiente estatístico R, e as variáveis utilizadas foram previamente normalizadas para garantir que todas contribuíssem de forma equilibrada para a segmentação. Esse procedimento assegurou que o processo de clusterização fosse eficaz, resultando em agrupamentos que refletiam com precisão as características socioeconômicas e geográficas dos municípios analisados.

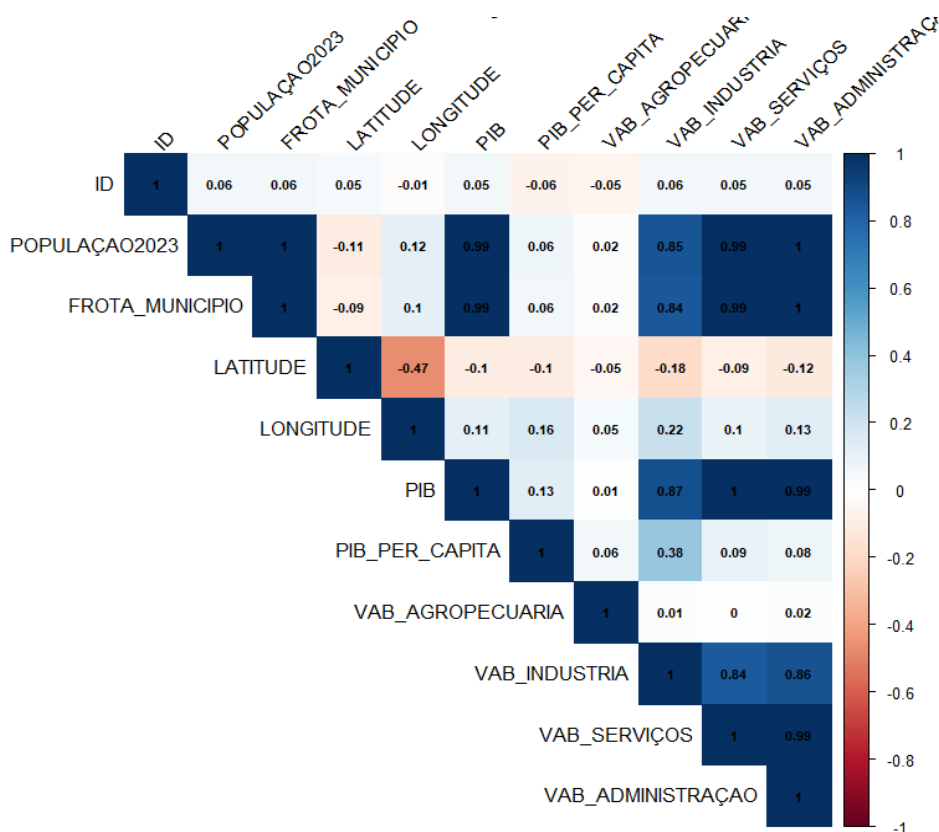
4. Resultados e Discussões

4.1 Análise de Correlação

A análise elaborada revelou uma correlação significativa entre a população dos municípios, a frota de veículos atualmente registrada em São Paulo, o Produto Interno Bruto (PIB) a preços correntes, e os valores adicionados brutos (VAB) de serviços de administração, defesa, educação, saúde pública e seguridade social, tanto no setor privado quanto público. Sabido desta constatação, optou-se por eliminar as variáveis redundantes, minimizando assim o processamento computacional necessário na clusterização.

Além da identificação das correlações, a análise exploratória também abrangeu a exploração das distribuições estatísticas e a detecção de possíveis valores atípicos. Tais procedimentos são utilizados para garantir a integridade dos dados e a robustez das conclusões derivadas da clusterização.

Figura 1 - Análise de Correlação das Variáveis Levantadas



Fonte: Os Autores (2024).

Após a eliminação das variáveis com alta correlação, foi elaborada uma nova matriz de correlação para as variáveis restantes: "POPULAÇÃO2023", "LATITUDE", "LONGITUDE", "PIB_PER_CAPITA", "VAB_AGROPECUARIA" e "VAB_INDUSTRIA". Esta etapa visa confirmar a influência das variáveis escolhidas para a clusterização dos municípios de São Paulo.

A análise, como apresentada na Tabela 1, mostrou que a correlação entre "POPULAÇÃO2023" e "VAB_INDUSTRIA" é alta (0.8519), sugerindo que municípios com maiores populações tendem a ter uma contribuição industrial mais significativa. A correlação negativa entre "LONGITUDE" e "LATITUDE" (-0.4675) reflete a distribuição geográfica dos municípios, com os valores indicando uma dispersão longitudinal e latitudinal relevante. As demais variáveis apresentaram correlações baixas, indicando independência mútua, o que é desejável para a clusterização. Com esses resultados, manteve-se o conjunto de dados inalterado para a aplicação do *k-means*.

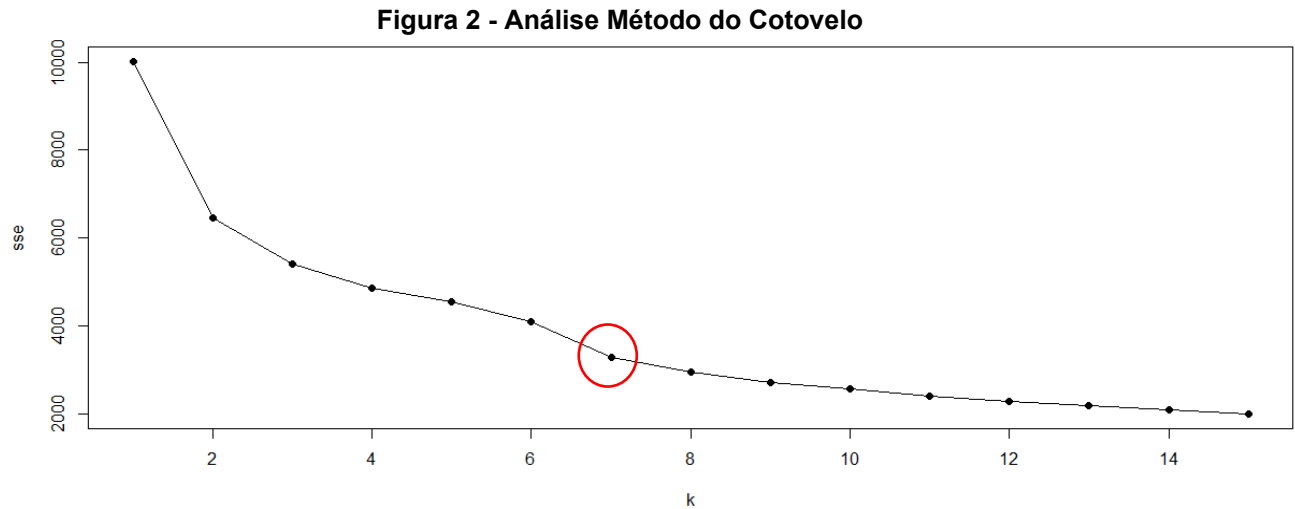
Tabela 1 - Matriz de Correlação das Variáveis para a Clusterização

	POPUL	LAT	LONG	PIB_PCAP	VAB_AGRO	VAB_INDU
POPUL	100.000.000	-0.10820817	0.11945614	0.06064834	0.01609088	0.85192624
LATITUDE	-0.10820817	100.000.000	-0.46752034	-0.10347235	-0.04809505	-0.1827399
LONGIT	0.11945614	-0.46752034	100.000.000	0.16305046	0.04683825	0.22250445
PIB_PCAP	0.06064834	-0.10347235	0.16305046	100.000.000	0.0560778	0.38037862
VAB_AGRO	0.01609088	-0.04809505	0.04683825	0.0560778	100.000.000	0.01462908
VAB_INDUS	0.85192624	-0.1827399	0.22250445	0.38037862	0.01462908	100.000.000

Fonte: Os Autores (2024)

4.2 Método do Cotovelo

Para determinar o número ideal de clusters, foi utilizado o método do cotovelo. Este método envolve a plotagem da soma das distâncias quadráticas (SSE) dentro dos clusters para diferentes valores de k e a identificação do ponto onde a taxa de diminuição se torna menos acentuada, formando um "cotovelo" no gráfico. A Figura 2 apresenta o gráfico do cotovelo gerado para este estudo.



Fonte: Os Autores (2024).

No gráfico apresentado, observa-se uma queda acentuada na SSE até o número de 7 clusters, o que indica que esse pode ser o ponto ideal em muitos contextos. A partir desse ponto, a melhoria na homogeneidade entre os grupos se torna significativamente menor. Essa escolha é especialmente relevante ao considerar a diversidade socioeconômica dos municípios de São Paulo, que são frequentemente agrupados em 15 mesorregiões ou 11 regiões intermediárias. Portanto, a decisão sobre o número de clusters não se baseia apenas na análise do gráfico do cotovelo, mas também na necessidade de capturar as nuances das características dos municípios, como PIB, frota de veículos e população, refletindo com precisão as disparidades entre eles, evitando generalizações inadequadas.

4.4 Utilização do Algoritmo K-means

Após a determinação do número de clusters $k=7$, o algoritmo *k-means* foi usado para segmentar os municípios de São Paulo. O resultado da execução do algoritmo *k-means* mostrou a formação de 7 clusters com os seguintes tamanhos: 34, 8, 21, 93, 47, 1, e 441. As médias das variáveis de cada cluster são apresentadas na Tabela 2.

Tabela 2 - Resultado da Clusterização por K-means

k	TAM k	POP 2023	LAT	LONG	PIB/PER_CAPITA	VAB_AGRÔ	VAB_INDÚSTRIA
1	34	161.599,65	-22.80656	-47.25608	84.118,03	90.589,35	2.445.427,85
2	8	666.064,62	-23.32297	-46.74721	133.357,34	64.212,20	15.022.755,63
3	21	327.318,24	-23.11284	-46.66850	106.441,32	94.345,60	6.127.405,47
4	93	46.357,31	-22.13746	-48.35863	53.919,97	117.657,15	485.660,12
5	47	119.568,28	-22.73777	-47.62852	55.816,77	77.970,22	1.197.194,29
6	1	11.451.999,00	-23.55050	-46.63330	66.872,84	61.624,74	62.513.855,11
7	441	12.090,05	-21.95528	-49.05436	30.818,21	58.843,96	50.662,24

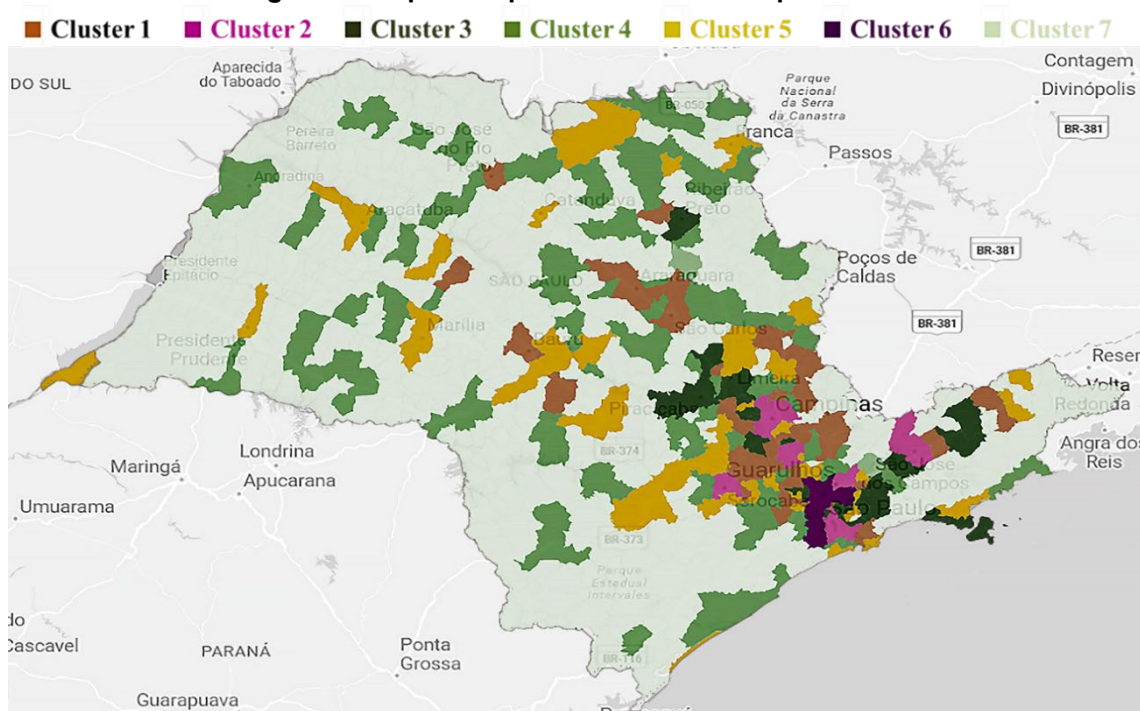
Fonte: Os Autores (2024)

A avaliação da clusterização utilizando o índice de silhueta médio resultou em um valor de 0.6379, indicando uma boa qualidade de clusterização com uma separação clara entre os clusters formados.

4.5 Resultado da Clusterização

A análise de clusterização pelo método *k-means* revelou uma diversidade substancial entre os municípios do estado de São Paulo, evidenciando variações significativas nas características demográficas e econômicas. A divisão dos municípios em sete clusters distintos fornece uma visão detalhada das disparidades regionais e das particularidades econômicas e sociais de cada grupo, como ilustrado na Figura 3.

Figura 3 - Mapa Coroplético de São Paulo por Cluster



Fonte: Os Autores (2024).

A clusterização dos municípios de São Paulo revelou padrões importantes sobre a dinâmica socioeconômica das regiões do estado, agrupando os municípios com base em variáveis como população, PIB per capita, VAB agropecuário e industrial. Esses clusters oferecem uma visão detalhada das diferentes realidades regionais, destacando áreas mais desenvolvidas e outras com potencial econômico menos expressivo.

A análise resultou em sete clusters, com grande diversidade tanto em tamanho quanto em características econômicas. O Cluster 6, composto por um único município, é o município de São Paulo, Capital. Este cluster destaca-se pela maior população entre todos (11.451.999 habitantes) e uma indústria altamente significativa, com um VAB industrial de R\$ 62,5 bilhões, o que o diferencia drasticamente dos demais grupos. A capital, portanto, reflete áreas de grande concentração urbana e industrial, sendo uma exceção no conjunto de dados devido ao seu peso econômico e populacional.

Por outro lado, o Cluster 7, o maior em número de municípios (441), apresenta uma média populacional bastante baixa (12.090 habitantes por município) e o menor PIB per capita entre os grupos, com apenas R\$ 30.818,21. Esse cluster representa municípios pequenos e rurais, com menor participação na economia estadual, especialmente no setor industrial, cujo VAB é o menor, com R\$ 50.662,24.

Os clusters intermediários mostram uma distribuição variada. O Cluster 2, com 8 municípios, chama a atenção pelo PIB per capita elevado, de R\$ 133.357,34, impulsionado por um VAB industrial robusto de R\$ 15 milhões, justificado pela presença de polos industriais de médio porte. Já o Cluster 4, com 93 municípios, exibe um perfil de regiões mais

agropecuárias, com o maior VAB agropecuário, chegando a R\$ 117.657,15, mas com um PIB per capita mais modesto em relação aos demais (R\$ 53.919,97).

Outros grupos, como o Cluster 1 e o Cluster 5, também mostram diversidade. O Cluster 1, com 34 municípios, tem um PIB per capita de R\$ 84.118,03 e um VAB industrial significativo (R\$ 2,4 milhões), refletindo uma economia mista, com forte presença tanto do setor industrial quanto agropecuário. O Cluster 5, com 47 municípios, apresenta características semelhantes, mas com um PIB per capita mais baixo (R\$ 55.816,77) e VAB industrial de R\$ 1,2 milhão.

Esses resultados ressaltam a diversidade econômica do estado de São Paulo, com clusters que vão desde áreas altamente industrializadas e densamente povoadas até regiões predominantemente agrícolas ou menos desenvolvidas economicamente. A análise pode servir como base para políticas públicas e estratégias de desenvolvimento regional que levem em conta as especificidades de cada cluster, buscando uma melhor alocação de recursos e incentivos econômicos conforme as necessidades e potencialidades de cada grupo.

5. Considerações Finais

A clusterização realizada neste estudo mostrou-se diferente das divisões tradicionais, como micro e mesorregiões ou, as comumente utilizadas, regiões intermediárias e imediatas do estado de São Paulo, evidenciando que as divisões baseadas no *k-means* podem ser eficazes para diferentes tipos de análises. Através da aplicação dessa técnica, foi possível observar que, ao ajustar os dados conforme o contexto, podemos identificar agrupamentos que oferecem uma nova perspectiva para a formulação de políticas públicas mais precisas e direcionadas. Essa abordagem é particularmente útil na gestão de políticas específicas, pois permite que decisões sejam tomadas com base em dados mais detalhados e ajustados conforme o foco das intervenções necessárias.

Adicionalmente, a clusterização oferece a possibilidade de personalizar a análise com diferentes pesos e variáveis, de acordo com o tipo de política que se pretende implementar, tornando a gestão mais eficiente. Por exemplo, é possível utilizar variáveis demográficas com maior peso em políticas de saúde e educação, enquanto fatores econômicos e produtivos podem ser priorizados em decisões ligadas ao desenvolvimento industrial. A flexibilidade proporcionada pela clusterização torna essa ferramenta poderosa para a segmentação e a gestão regional, permitindo adaptar as divisões conforme as necessidades específicas de cada área ou setor.

Essas segmentações, geradas por dados atualizados e específicos, abrem novas possibilidades para uma gestão regional mais eficiente, em que as particularidades de cada município são mais bem compreendidas e aproveitadas. A identificação de clusters com perfis socioeconômicos distintos, diferente das divisões geográficas tradicionais, proporciona uma visão mais detalhada das disparidades e potencialidades regionais, permitindo intervenções mais eficazes e estratégias de alocação de recursos mais precisas para o desenvolvimento sustentável do estado de São Paulo.

REFERÊNCIAS

EVERITT, B.; LANDAU, S.; LESE, M. **Cluster Analysis**. 2011. John Wiley & Sons.

GOVERNO FEDERAL. **Frota de veículos**. 2023. Disponível em: <https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2023>. Acesso em: 2024.

HISSA, Hélio. **Clusterização: uma alternativa de desenvolvimento regional para o Brasil?** Saarbrücken: Novas Edições Acadêmicas, 2015. 200 p.

IBGE. **Panorama de São Paulo.** 2022. Disponível em: <https://cidades.ibge.gov.br/brasil/sp/panorama>. Acesso em: 2024.

IBGE. **Panorama do Brasil.** 2022. Disponível em: <https://cidades.ibge.gov.br/brasil/panorama>. Acesso em: 2024.

IBGE. **Produto Interno Bruto dos Municípios.** 2021. Disponível em: <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/pesquisa/38/46996>. Acesso em: 2024.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. **Data Clustering: A Review.** *ACM Computing Surveys (CSUR)*, v. 31, n. 3, p. 264-323, 1999.

KASSAMBARA, A. **Practical guide to cluster analysis in R: unsupervised machine learning.** 2017. STHDA. ISBN 978-1-5424-6262-9.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: An Introduction to Cluster Analysis.** 1990. ISBN: 0-471-87876-6.

MACQUEEN, J. B. **Some Methods for Classification and Analysis of Multivariate Observations.** *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, p. 281-297, 1967.

MOUNT, J.; ZUMEL, N. **Practical Data Science with R.** Manning Publications, 2019. ISBN 978-1-61729-587-4.

THORNDIKE, R. L. **Who Belongs in the Family?** *Psychometrika*, v. 18, n. 4, p. 267-276, 1953.

ROUSSEEUW, P. J. **Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.** *Journal of Computational and Applied Mathematics*, v. 20, p. 53-65, 1987.

SANTOS, C. R.; SILVA, M. J.; PEREIRA, S. R. **Cluster Analysis of Socioeconomic and Demographic Variables in São Paulo State, Brazil.** *Regional Studies*, v. 54, n. 6, p. 789-801, 2020.

SANTOS, Leon. **Entenda por que a clusterização pode alavancar a economia brasileira.** *Revista RBA*, São Paulo, 17 out. 2023. Disponível em: <https://revistarba.org.br/entenda-por-que-a-clusterizacao-pode-alavancar-a-economia-brasileira/>. Acesso em: 19 set. 2024.

STORPER, M.; SCOTT, A. J. **Repensando o capital humano, a criatividade e o crescimento urbano.** *Journal of Economic Geography*, v. 9, n. 2, 2009.

ZHANG, Zhongqian; LIU, Yaqun; SHENG, Shuangqing; LIU, Xu; XUE, Qiuli. **Evolving Urban Expansion Patterns and Multi-Scenario Simulation Analysis from a Composite Perspective of “Social–Economic–Ecological”:** A Case Study of the Hilly and Gully Regions of Northern Loess Plateau in Shaanxi Province. *Sustainability*, v. 16, p. 2753, 2024. Disponível em: <https://doi.org/10.3390/su16072753>.