Análise Comparativa de Métodos de Clusterização em Dados Geoespaciais

Eduardo Pedro Kuiava

Instituto de Tecnologia – Universidade de Passo Fundo (UPF) – Caixa Postal 611 – CEP 99001-970 – Passo Fundo – RS – Brasil

172945@upf.br

Abstract. *This study presents a comparative analysis of clustering methods applied to geospatial data from municipalities in Rio Grande do Sul, focusing on the performance and representativeness of the resulting groups. The evaluated methods include K-means, Weighted K-means, Agglomerative Clustering, and DBSCAN, using geographic and population data. The methodology incorporated fixed centroids, inspired by Aline Gomes's thesis, to align automatic methods with regional contexts. Results showed that K-means and Weighted K-means outperformed in terms of cluster cohesion and separation, particularly in densely populated regions. Conversely, DBSCAN was effective in detecting outliers but faced challenges when adapted to fixed centroids. The analysis emphasized that integrating automatic methods with manual adjustments can offer more effective hybrid solutions, combining mathematical efficiency with qualitative sensitivity. This study contributes to enhancing clustering techniques in geospatial analyses, providing valuable insights for public policy and regional planning.*

Keywords: *clustering, geospatial, K-means, DBSCAN, regional analysis.*

Resumo. *Este trabalho apresenta uma análise comparativa de métodos de clusterização aplicados a dados geoespaciais dos municípios do Rio Grande do Sul, com foco no desempenho e na representatividade dos agrupamentos formados. Foram avaliados os métodos K-means, Weighted K-means, Agglomerative Clustering e DBSCAN, utilizando como base dados geográficos e populacionais. A metodologia incluiu a definição de centroides fixos, inspirada na tese de Aline Gomes, para alinhar os métodos automáticos ao contexto regional. Os resultados mostraram que o K-means e o Weighted K-means tiveram melhor desempenho em termos de coesão e separação dos clusters, especialmente em regiões de alta densidade populacional. Por outro lado, o DBSCAN demonstrou eficácia na detecção de outliers, mas enfrentou desafios em sua adaptação aos centroides fixos. A análise destacou que a integração de métodos automáticos e ajustes manuais pode proporcionar soluções híbridas mais eficazes, combinando eficiência*

matemática com sensibilidade qualitativa. Este estudo contribui para o aprimoramento das técnicas de clusterização em análises geoespaciais, promovendo insights úteis para políticas públicas e planejamento regional.

Palavras-chave: *clusterização, geoespacial, K-means, DBSCAN, análise regional.*

1. INTRODUÇÃO

Os dados geoespaciais, que combinam informações geográficas, como coordenadas, com atributos relacionados, desempenham um papel crucial em análises avançadas em áreas como planejamento urbano, agricultura, saúde e transporte. Com o avanço das tecnologias de coleta e processamento de dados, esses dados têm se tornado ferramentas indispensáveis para a tomada de decisões estratégicas. Além de capturar as características físicas do espaço, eles possibilitam a identificação de padrões espaciais e demográficos, orientando ações em setores críticos e promovendo uma gestão mais eficiente e baseada em evidências.

Nos últimos anos, a popularização de ferramentas como sistemas de informação geográfica (GIS) e bibliotecas computacionais, como GeoPandas, scikit-learn e Folium, permitiu que a análise geoespacial fosse amplamente acessível a pesquisadores e tomadores de decisão. Com a disponibilidade crescente de dados, surgem desafios relacionados à identificação de padrões e agrupamentos que possam facilitar a interpretação e aplicação prática desses dados. Técnicas de clusterização têm se mostrado eficazes para lidar com esses desafios, permitindo identificar padrões latentes que podem orientar ações em áreas como planejamento urbano, distribuição de recursos e monitoramento ambiental.

O trabalho de Gomes (2021) propõe diretrizes para a elaboração de planos de regionalização destinados à gestão integrada de resíduos sólidos urbanos (RSU), contribuindo para a constituição de consórcios públicos intermunicipais. Esses consórcios foram pensados como uma alternativa sustentável e eficiente para atender aos princípios globais de sustentabilidade e às diretrizes normativas relacionadas à gestão de resíduos. Como resultado, foi proposta a escolha de 15 cidades polos e, de acordo com as diretrizes elaboradas, as demais cidades do Rio Grande do Sul foram agrupadas em torno de cada uma destas cidades. Para definir as cidades polo (ou centroides) em seus mapas, Gomes (2021) baseou-se em critérios técnicos e qualitativos que envolvem aspectos como:

1. Relevância econômica e populacional das cidades;

2. Localização estratégica para promover a acessibilidade entre os municípios consorciados;
3. Infraestrutura disponível para o tratamento e destinação de resíduos sólidos;
4. Distância geográfica entre os municípios.

Esses critérios permitiram identificar cidades capazes de atuar como pólos regionais, facilitando a formação de consórcios públicos e contribuindo para a gestão sustentável e integrada dos RSU. Além disso, os arranjos territoriais propostos levaram em conta fatores como a hierarquia dos municípios e suas condições geográficas e socioeconômicas, promovendo uma abordagem inclusiva e estratégica. Este agrupamento obtido pelo trabalho de Gomes (2021) será utilizado como base de comparação aos resultados que este trabalho obteve com a aplicação de métodos de clusterização.

O objetivo deste trabalho é realizar uma análise comparativa entre diferentes métodos de clusterização aplicados aos municípios do estado do Rio Grande do Sul com o agrupamento obtido por Gomes (2021). Além do K-means com centroides fixos, este estudo inclui métodos alternativos, como Weighted K-means, Agglomerative Clustering e DBSCAN (com e sem centroides fixos). A proposta é avaliar as vantagens e limitações de cada método em termos de desempenho, representatividade visual e coerência dos agrupamentos, considerando dados heterogêneos e contextos densamente povoados. O estudo busca fornecer uma visão abrangente sobre como essas técnicas podem ser aplicadas para identificar padrões geoespaciais, contribuindo para análises mais robustas e alinhadas às realidades regionais.

Este artigo está estruturado da seguinte forma: na seção 2 apresenta-se uma revisão bibliográfica sobre métodos de clusterização e sua aplicação em dados geoespaciais. Na seção 3 descreve-se a metodologia utilizada, abrangendo o pré-processamento dos dados, ferramentas empregadas e critérios de avaliação. Na seção 4 detalham-se a execução e os resultados obtidos para cada método de clusterização, com análises comparativas. Por fim, na seção 5, são discutidos os principais achados, limitações e possíveis trabalhos futuros, encerrando-se com as conclusões que destacam as contribuições e implicações práticas deste estudo.

2. REVISÃO DA LITERATURA

A análise geoespacial combina dados espaciais e atributos descritivos para identificar padrões e relações, desempenhando um papel fundamental em áreas como planejamento urbano, gestão ambiental e desenvolvimento socioeconômico. Esta seção aborda os fundamentos dos métodos de clusterização aplicados a dados geoespaciais, destacando sua aplicabilidade prática e as ferramentas computacionais utilizadas, como Python, Scikit-learn, GeoPandas e Folium, que viabilizaram as análises realizadas neste trabalho. Inspirado pela abordagem apresentada na tese de Gomes (2021), que utilizou centroides fixos para estudar os municípios do Rio Grande do Sul, este estudo amplia a análise para métodos como K-means, Weighted K-means, Agglomerative Clustering e DBSCAN. A contribuição da tese destaca-se por explorar as vantagens e limitações de cada técnica, fornecendo uma base sólida para comparações metodológicas e análises práticas em dados geoespaciais.

2.1. Dados Geoespaciais e Clusterização

Os dados geoespaciais combinam informações geográficas, como latitude e longitude, com atributos específicos, como população, infraestrutura, condições ambientais e outras variáveis relevantes. Essa integração permite uma análise aprofundada de padrões espaciais e contextuais, sendo amplamente utilizada em áreas como geografia, planejamento urbano, análise ambiental e economia regional. No entanto, esses dados apresentam desafios específicos, como a alta dimensionalidade, a presença de lacunas e a necessidade de considerar as relações espaciais inerentes aos dados. Isso exige o uso de técnicas especializadas capazes de processar e interpretar essas complexidades.

A clusterização, ou análise de agrupamento, é uma dessas técnicas e tem como objetivo identificar padrões ou agrupamentos de elementos semelhantes dentro de um conjunto de dados. Em análises geoespaciais, ela é especialmente útil para segmentar regiões com características compartilhadas, como áreas com densidades populacionais semelhantes, padrões econômicos ou características ambientais específicas. Os métodos de clusterização permitem explorar e organizar dados de forma sistemática, fornecendo insights valiosos para tomadas de decisão e desenvolvimento de políticas públicas.

No trabalho de Gomes (2021), a clusterização foi realizada de forma manual, seguindo critérios adotados pela autora, como relevância econômica, conectividade regional e características culturais para definir os agrupamentos. Essa abordagem destaca uma das

principais limitações dos métodos automáticos: sua incapacidade de considerar fatores qualitativos e contextuais. Ao integrar esses fatores manuais à análise automatizada, é possível produzir resultados mais alinhados às realidades locais, uma lição importante trazida pelo estudo comparativo realizado neste trabalho.

2.2. Trabalhos Relacionados

O trabalho de Metz (2024) explora como o aprendizado de máquina transforma a análise geoespacial, destacando o uso de bibliotecas como scikit-learn e GeoPandas. Ele mostra exemplos práticos de aplicação de K-means e Random Forest em geociências, com foco em segmentação de áreas urbanas e previsões de densidade populacional. A integração de dados de sensores remotos e satélites é um diferencial para enriquecer as análises. Ferramentas como QGIS são mencionadas como suporte na visualização de padrões.

O guia, elaborado por Aya Data (2024), enfatiza o uso de Python como linguagem principal para análise de dados geoespaciais, destacando bibliotecas como GeoPandas, Shapely e Folium para manipulação e visualização de dados espaciais. Além disso, aborda o DBSCAN para identificar anomalias em dados geográficos, sugerindo ajustes criteriosos de parâmetros (eps e min_samples) para maximizar a detecção de padrões em regiões de densidades variadas. A importância do pré-processamento, como a correção de valores ausentes, também é amplamente discutida.

O estudo realizado por Urban Spatial (2024) explora a aplicação prática de métodos como K-means e Agglomerative Clustering em análises geoespaciais, utilizando bibliotecas robustas como scikit-learn e matplotlib. Além de apresentar a implementação técnica dessas abordagens, o trabalho destaca o uso de ferramentas de visualização, como Folium, para criar mapas interativos que facilitam a interpretação dos padrões identificados. Essa combinação de técnicas é apresentada como uma solução prática para análises espaciais detalhadas, sendo especialmente útil em cenários que exigem uma compreensão clara da distribuição geográfica de atributos.

O estudo conduzido por Rey et al (2024) concentra-se na aplicação da biblioteca PySAL para realizar clusterizações espacialmente restritas, um diferencial importante em análises que precisam respeitar limites administrativos. O trabalho enfatiza a regionalização de áreas geográficas, integrando variáveis espaciais e demográficas para criar agrupamentos

que atendam a critérios práticos e políticos. Essa abordagem é destacada como essencial para análises espaciais multivariadas e para a alocação eficiente de recursos em regiões com grande variação geográfica.

O texto de ArcGIS (2024) explora as capacidades do software em realizar clusterizações multivariadas, com ênfase em respeitar limites geográficos predefinidos. Ferramentas avançadas do ArcGIS são apresentadas como essenciais para aplicações em planejamento urbano e políticas públicas, permitindo a integração de variáveis espaciais e demográficas em análises regionais. O trabalho destaca como o uso de tais ferramentas possibilita a geração de agrupamentos que respeitam restrições espaciais, oferecendo suporte robusto para decisões estratégicas em ambientes urbanos e regionais.

Gomes (2021) explorou o uso de centroides fixos para criar agrupamentos consistentes, facilitando comparações e análises espaciais. Apesar de não utilizar métodos de machine learning, sua abordagem foi robusta ao demonstrar que os centroides fixos permitem uma interpretação clara e direta dos agrupamentos.

3. METODOLOGIA

Este estudo utilizou um conjunto de dados geoespaciais dos municípios do Rio Grande do Sul, abrangendo informações cruciais como latitude, longitude, população e nomes dos municípios. Para garantir uma análise mais robusta e abrangente, foi realizada a integração de shapefiles ao conjunto de dados, permitindo uma representação visual detalhada das áreas de interesse. Essa integração possibilitou não apenas uma avaliação espacial precisa, mas também a criação de mapas interativos, ferramentas indispensáveis para a interpretação clara e acessível dos resultados. Ao longo desta seção, detalham-se os procedimentos de pré-processamento, aplicação dos métodos de clusterização e as estratégias empregadas para avaliar e comparar os resultados obtidos.

3.1. Pré-processamento dos Dados

O pré-processamento dos dados desempenhou um papel essencial para assegurar a integridade e compatibilidade com os métodos de clusterização aplicados neste estudo.

Inicialmente, realizou-se a correção de valores decimais nos dados de latitude e longitude, substituindo vírgulas por pontos, de modo a padronizar o formato das coordenadas geográficas e evitar inconsistências durante os cálculos. Posteriormente, tratou-se os valores ausentes no conjunto de dados, com estratégias como preenchimento ou remoção de entradas incompletas, garantindo que esses dados não comprometessem os resultados das análises.

Para métodos que dependem de métricas de distância, como o K-means e o Agglomerative Clustering, aplicou-se a normalização das variáveis, ajustando os valores de latitude e longitude para uma escala padrão. Essa etapa foi crucial para garantir que as métricas de similaridade refletissem de maneira equilibrada a relação espacial entre os municípios. No caso específico do DBSCAN, as coordenadas foram convertidas para radianos, permitindo o cálculo de distâncias geodésicas, que são mais adequadas para dados geoespaciais devido à curvatura da Terra. Além disso, uma etapa diferenciada foi a definição de centroides fixos, inspirada no trabalho de Gomes (2021). Foram selecionados 15 municípios polos estratégicos do Rio Grande do Sul como centroides fixos, com base em sua relevância geográfica e econômica. Essa abordagem permitiu uma base consistente de comparação entre os métodos, mantendo um padrão que refletisse os padrões regionais já estudados previamente.

3.2. Ferramentas, Linguagens e Bibliotecas Utilizadas

A análise e visualização de dados geoespaciais neste estudo foram realizadas com o uso da linguagem Python e uma seleção de bibliotecas amplamente reconhecidas por sua eficiência e versatilidade. Python foi escolhido pela sua sintaxe acessível, robustez e grande suporte a ferramentas especializadas em manipulação de dados e cálculos geoespaciais.

Entre as bibliotecas de análise de dados, destacaram-se o Pandas e o Numpy. O Pandas foi essencial para a manipulação dos dados tabulares, como a organização de coordenadas e informações populacionais, enquanto o Numpy facilitou operações matemáticas e o cálculo de distâncias geográficas, fundamentais para métodos como o Weighted K-means.

Os algoritmos de clusterização foram implementados com a biblioteca Scikit-learn, que forneceu as ferramentas para K-means, Weighted K-means, Agglomerative Clustering e

DBSCAN, além de métricas de avaliação como o Silhouette Score. Complementando a análise, GeoPandas e Haversine permitiram integrar e calcular distâncias geodésicas, cruciais para métodos baseados em densidade e na curvatura da Terra.

Para a visualização, Matplotlib e Seaborn foram utilizadas na criação de gráficos comparativos, com ajustes nas cores para tornar os resultados mais intuitivos e atraentes. A geração de mapas interativos foi realizada com a biblioteca Folium, que possibilitou sobrepor os resultados dos métodos de clusterização em mapas detalhados, destacando os centroides fixos e os clusters resultantes. Todo o trabalho foi executado no ambiente de desenvolvimento Google Colab, que ofereceu recursos computacionais baseados na nuvem e integração direta com o Google Drive para armazenamento dos resultados e dados. A sinergia dessas ferramentas garantiu precisão nos cálculos, clareza nas análises e resultados visualmente acessíveis.

3.3. Aplicação dos Métodos

Para a análise e comparação, foram aplicados quatro métodos de clusterização amplamente utilizados em dados geoespaciais: K-means, Weighted K-means, Agglomerative Clustering e DBSCAN. Cada método foi configurado e adaptado às características dos dados e aos objetivos do estudo, destacando suas vantagens, limitações e particularidades na identificação de padrões espaciais.

O K-means, um dos métodos mais clássicos e amplamente utilizados, foi configurado para gerar 15 clusters, utilizando latitude e longitude como variáveis principais. Esse método baseia-se na minimização da soma das distâncias dos pontos aos centróides dentro de cada cluster, sendo particularmente eficiente em dados bem distribuídos e de densidade uniforme. A simplicidade do K-means e sua velocidade de execução o tornam ideal para aplicações em larga escala. No entanto, sua limitação está na dificuldade de lidar com clusters de formas irregulares e na sensibilidade a outliers.

O Weighted K-means, uma variação do K-means, foi utilizado para incorporar pesos aos dados, representando a população de cada município. Essa adaptação permite que clusters sejam influenciados por fatores demográficos, refletindo a importância relativa de cada região. O método foi configurado para considerar 15 clusters, com os pesos ajustando a

influência dos pontos na formação dos agrupamentos. Essa abordagem se mostrou eficiente para análises em que fatores populacionais desempenham um papel importante, gerando clusters que representam com maior fidelidade as concentrações demográficas do estado. Embora apresente resultados mais representativos em cenários com dados ponderados, o Weighted K-means mantém as limitações do K-means em relação a formas irregulares e outliers.

O Agglomerative Clustering, um método hierárquico, foi aplicado utilizando a métrica de ligação Ward para formar 15 clusters. Essa técnica agrupa iterativamente os dados com base na similaridade, começando com cada ponto como um cluster individual e fundindo os mais próximos até atingir o número desejado de clusters. Essa abordagem permitiu identificar relações hierárquicas entre os municípios, fornecendo insights detalhados sobre subgrupos naturais nos dados. Apesar de sua capacidade de gerar uma visão hierárquica, o Agglomerative Clustering é computacionalmente mais intenso e, assim como o K-means, não lida bem com outliers ou formas irregulares de clusters.

O DBSCAN (Density-Based Spatial Clustering of Applications with Noise) foi configurado com os parâmetros eps = 0.003 (aproximadamente 300 metros) e min_samples = 5, utilizando uma matriz de distâncias geodésicas calculada a partir das coordenadas radianas. Este algoritmo se destaca por sua capacidade de detectar clusters de forma arbitrária e identificar outliers, sendo ideal para dados espaciais com densidade variável. No entanto, os resultados do DBSCAN dependem fortemente da parametrização adequada, e sua aplicação com centroides fixos apresentou desafios, já que a natureza do algoritmo é incompatível com essa abordagem. Enquanto a versão padrão do DBSCAN detectou outliers e padrões densos, a versão com centroides fixos precisou de ajustes significativos para alinhar os agrupamentos aos centroides predefinidos, resultando em uma redução na capacidade de detecção de anomalias.

Esses quatro métodos fornecem uma base diversificada para a análise de dados geoespaciais, cada um com seus méritos e limitações. A aplicação combinada e a comparação entre eles oferecem uma visão abrangente sobre como diferentes técnicas podem ser utilizadas para interpretar padrões espaciais e demográficos em um contexto regional.

3.4. Critérios de Avaliação

Os métodos de clusterização aplicados neste estudo foram avaliados com base em critérios rigorosos que permitiram mensurar sua eficácia, eficiência e adequação aos dados geoespaciais utilizados. Os critérios adotados incluem o Silhouette Score, o tempo de execução e a detecção de outliers, oferecendo uma análise detalhada e comparativa das técnicas empregadas.

O Silhouette Score foi utilizado como uma métrica fundamental para avaliar a qualidade dos agrupamentos formados por cada método. Essa métrica analisa simultaneamente a separação entre os clusters e a coesão interna dos grupos, oferecendo uma medida numérica que varia de -1 a 1. Valores próximos a 1 indicam que os dados estão bem agrupados e os clusters são bem separados, enquanto valores negativos sugerem que os pontos podem estar alocados em clusters errados. Essa métrica é especialmente relevante em análises geoespaciais, pois reflete a eficácia dos métodos em identificar padrões espaciais significativos.

O tempo de execução foi outro critério essencial, medindo a eficiência computacional de cada método. Esse indicador é particularmente importante ao considerar a aplicação de técnicas de clusterização em conjuntos de dados maiores e mais complexos, como aqueles frequentemente encontrados em análises geoespaciais. Métodos mais rápidos podem ser mais adequados para análises em tempo real ou em cenários com recursos computacionais limitados, enquanto métodos mais lentos, mas com resultados robustos, podem ser preferidos para análises detalhadas.

A detecção de outliers foi avaliada como um aspecto crítico, especialmente no contexto de dados espaciais complexos, onde anomalias podem representar informações valiosas, como regiões geográficas atípicas ou dados com densidade reduzida. Métodos como o DBSCAN, por exemplo, são particularmente eficazes na identificação de outliers devido à sua abordagem baseada em densidade, enquanto técnicas como o K-means e o Agglomerative Clustering, por sua natureza, tendem a ignorar ou suavizar a influência desses dados atípicos.

A combinação desses critérios permitiu uma avaliação abrangente dos métodos de clusterização aplicados, destacando suas forças e limitações em diferentes contextos. Essa análise comparativa fornece insights valiosos para selecionar a técnica mais apropriada de

acordo com os objetivos específicos da análise geoespacial e as características dos dados disponíveis.

4. EXECUÇÃO E RESULTADOS

A aplicação dos métodos de clusterização gerou agrupamentos que refletem padrões geoespaciais nos municípios do Rio Grande do Sul, baseados em proximidade geográfica e densidade populacional. Esses clusters foram visualizados em mapas, com cores distintas para cada agrupamento, permitindo uma análise clara das divisões regionais. Os centroides fixos, definidos em municípios estratégicos como Porto Alegre e Caxias do Sul, foram utilizados como referência em métodos como K-means e Weighted K-means, replicando a abordagem adotada por Gomes (2021) em seu trabalho. Essa escolha garantiu uma conexão com realidades regionais importantes, mantendo um padrão comparativo entre os métodos automáticos e a abordagem manual. O resultado de seu agrupamento pode ser visto na Figura 1.

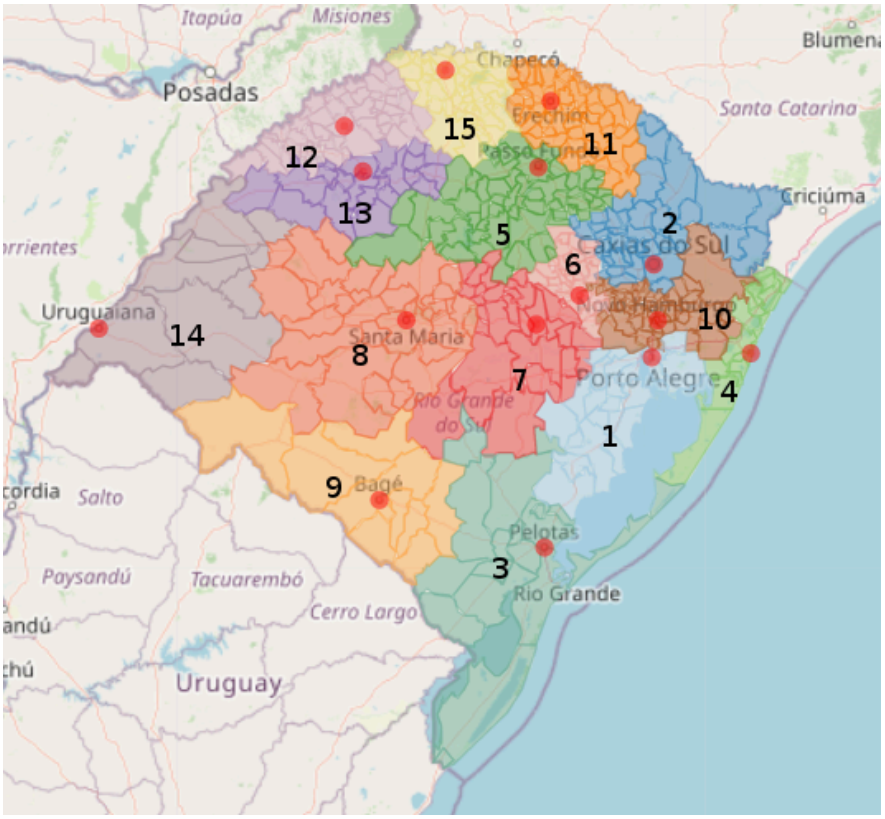

Figura 1. Mapa resultado de Gomes (2021)

Essa análise comparativa evidencia como diferentes métodos interpretam os dados espaciais, destacando vantagens e limitações em relação à abordagem manual. Os mapas e tabelas comparativas reforçam que, enquanto métodos automáticos como K-means garantem rapidez e replicabilidade, eles podem falhar em capturar nuances regionais e contextuais que a abordagem manual de Gomes (2021) priorizou. Os resultados sugerem que a combinação de métodos automáticos e ajustes manuais pode oferecer soluções mais robustas e alinhadas às realidades locais em estudos geoespaciais.

4.1. Aplicação dos métodos de clusterização

Os métodos K-means e Weighted K-means demonstraram resultados superiores, especialmente em cenários onde os centroides fixos foram utilizados. O K-means apresentou agrupamentos bem organizados ao redor dos centroides predefinidos, com boa separação espacial e coesão interna, sendo eficiente tanto em tempo de execução quanto em simplicidade de aplicação Figura 2.

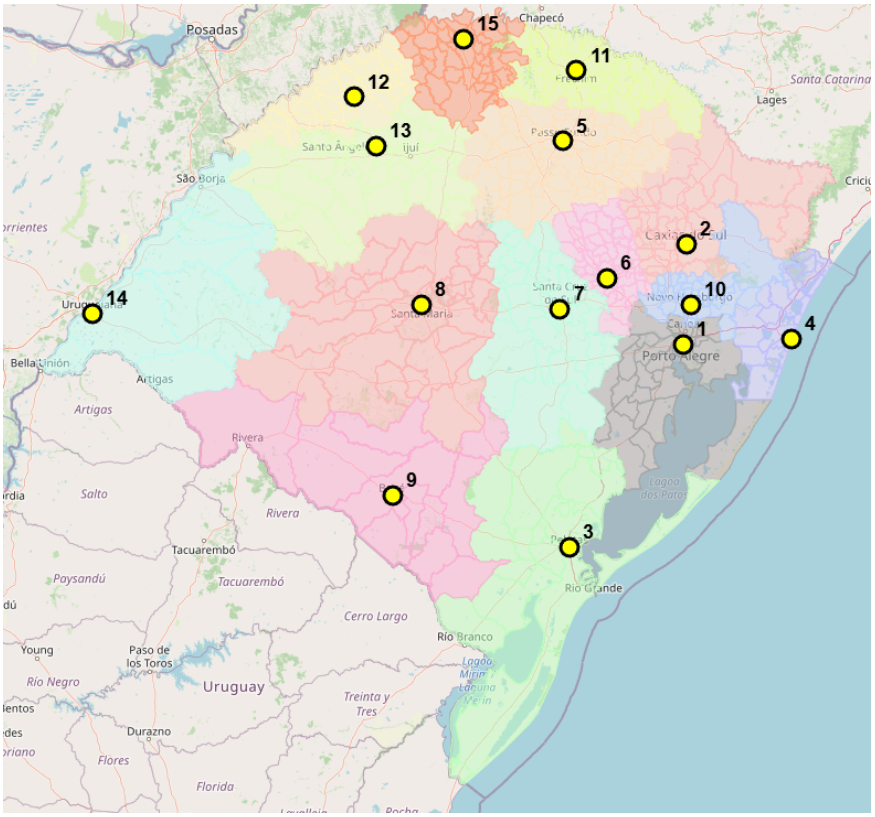

Figura 2. Resultado da clusterização obtida pelo método K-means

Já o Weighted K-means, por sua vez, se destacou pela capacidade de incorporar variáveis ponderadas, como a população, tornando os agrupamentos mais representativos da realidade demográfica dos municípios Figura 3. Essa versão ponderada conseguiu ajustar os clusters de forma a refletir variações populacionais, aumentando sua aplicabilidade em análises voltadas para políticas públicas e planejamento regional.

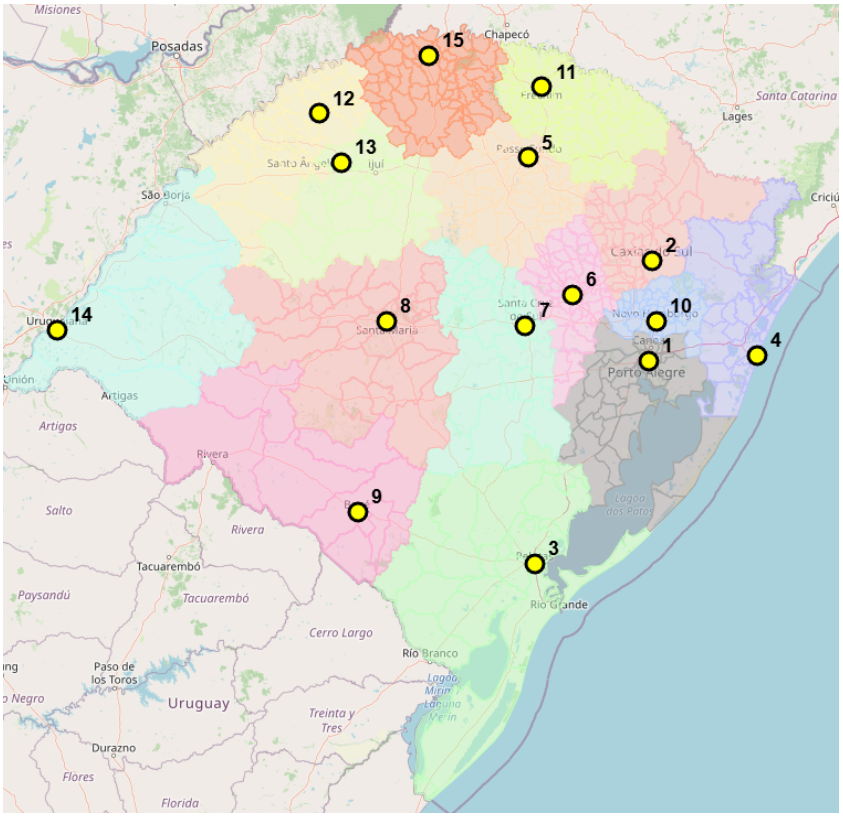

Figura 3. Resultado da clusterização obtida pelo método K-means Weighted

O Agglomerative Clustering evidenciou seu potencial ao identificar subgrupos hierárquicos naturais entre os municípios, fornecendo uma visão mais detalhada das relações internas dos agrupamentos Figura 4. Contudo, sua incapacidade de respeitar os centroides fixos dificultou comparações diretas com os métodos baseados em centroides predefinidos. Essa característica, embora seja uma limitação para algumas aplicações, pode ser uma vantagem em análises exploratórias ou em cenários onde hierarquias são mais relevantes do que a proximidade geográfica direta.

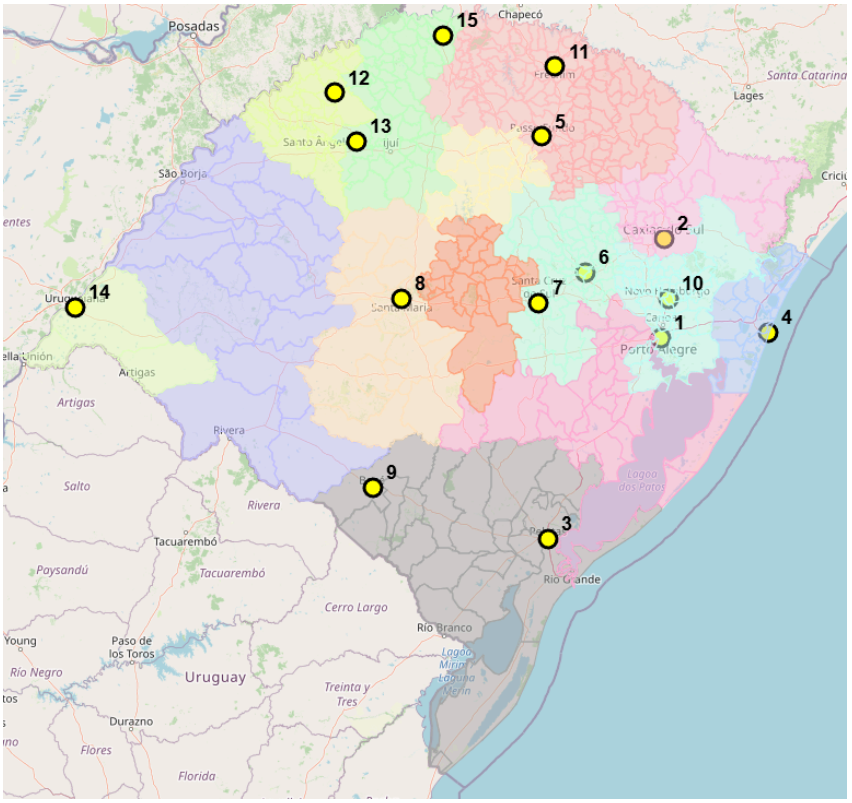

Figura 4. Resultado da clusterização obtida pelo método Agglomerative Clustering

O DBSCAN, em sua versão sem centroides fixos, identificou regiões de alta densidade e destacou 158 outliers, evidenciando sua robustez em relação a dados ruidosos e de densidade variável. No entanto, o método apresentou baixa coesão interna dos clusters, refletindo uma separação menos clara entre os grupos (Figura 5). Apesar disso, sua habilidade em identificar anomalias o torna especialmente valioso para análises focadas em detectar padrões atípicos ou regiões de transição geográfica.

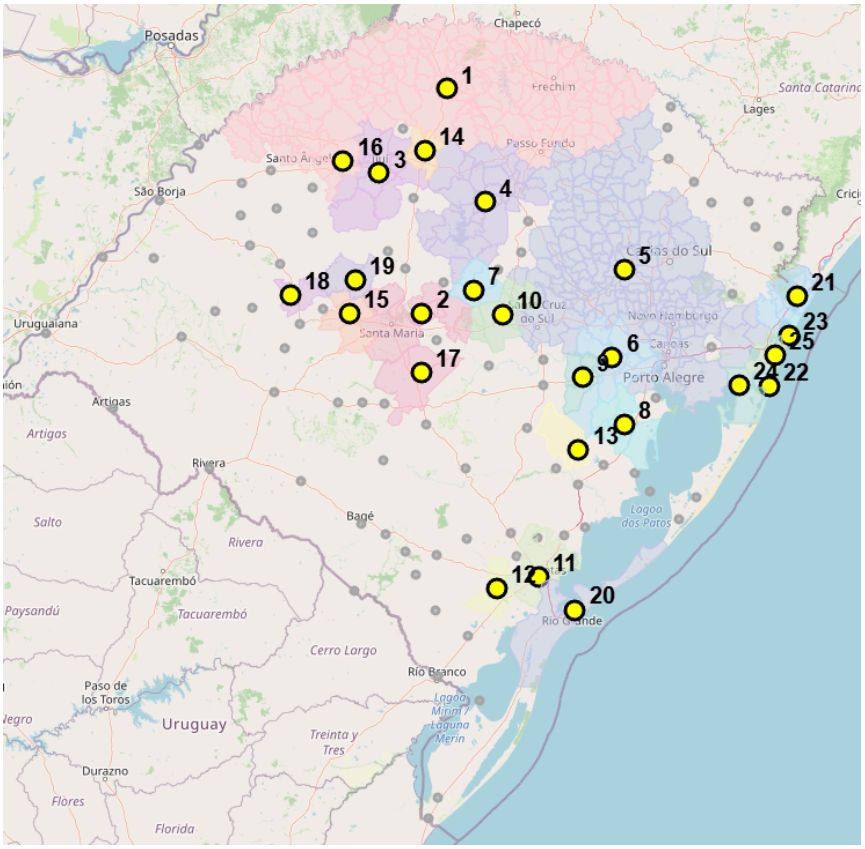

Figura 5. Resultado da clusterização obtida pelo método dbscan

Já a versão adaptada do DBSCAN, configurada para se ajustar a centroides fixos, conseguiu melhorar a coesão dos clusters, garantindo agrupamentos mais consistentes e organizados em torno das localizações predefinidas (Figura 6). No entanto, essa adaptação resultou na perda de sua característica mais distinta: a capacidade de identificar anomalias. Essa versão do DBSCAN é mais adequada para análises onde a consistência e a organização espacial são prioritárias, mas menos eficaz para estudos que necessitam explorar irregularidades nos dados.

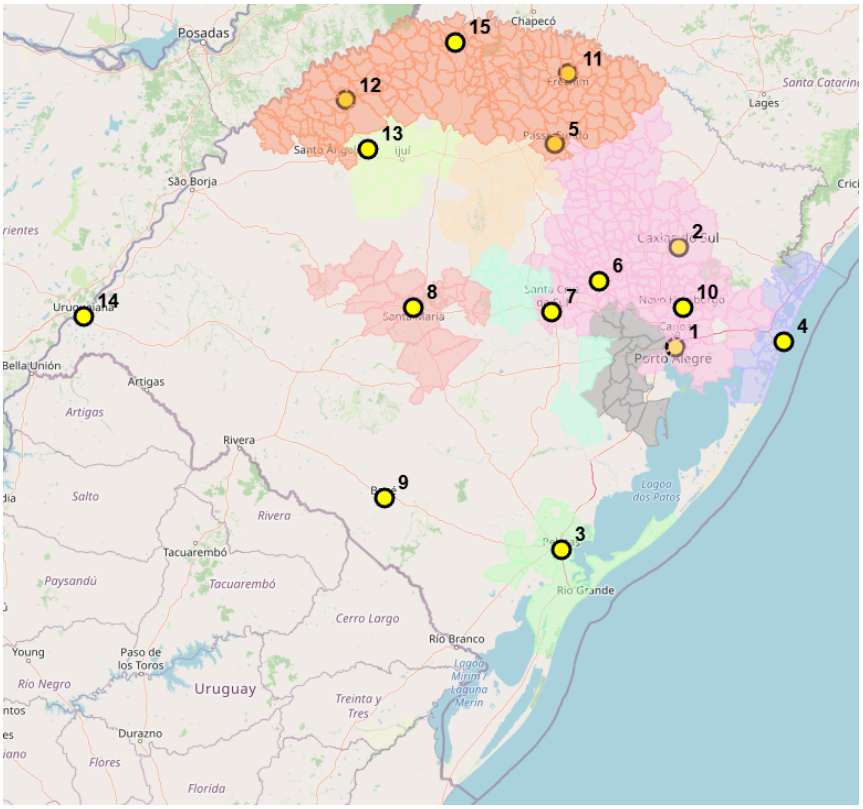

Figura 6. Resultado da clusterização obtida pelo método dbscan adaptado

Esses resultados destacam a importância de selecionar o método de clusterização mais adequado ao objetivo da análise e às características dos dados. Enquanto métodos como o K-means e o Weighted K-means são ideais para análises estruturadas com centroides fixos, o DBSCAN se sobressai em cenários onde a detecção de outliers e a flexibilidade dos agrupamentos são essenciais. O Agglomerative Clustering, por sua vez, oferece uma perspectiva complementar, útil para análises hierárquicas ou exploratórias. A escolha do método, portanto, deve ser guiada por uma avaliação criteriosa dos requisitos da aplicação e das limitações de cada técnica.

4.2. Análise Comparativa dos Resultados

Na Tabela 1 é apresentada uma visão comparativa dos métodos de clusterização aplicados no estudo, destacando os valores do Silhouette Score, o tempo de execução e o número de outliers detectados para cada método. Essa análise fornece informações fundamentais sobre a eficiência, a qualidade dos agrupamentos e a capacidade de lidar com outliers em diferentes cenários.

Tabela 1. Comparação de eficiência

method	silhouette_score	execution_time	outliers_detected
K-means	0.371	0.0168	0
Weighted K-means	0.3347	5.588	0
Agglomerative Clustering	0.3281	0.0323	0
DBSCAN	-0.0029	0.028	80

Os resultados de cada método foram representados em mapas interativos, com cores vivas para destacar os clusters formados. Os centróides de cada cluster foram marcados como pontos distintos, facilitando a identificação visual dos agrupamentos. No caso específico do DBSCAN, os outliers foram representados como pontos isolados, ressaltando regiões que não se encaixam em nenhum cluster identificado.

Os resultados foram visualizados em mapas interativos. Métodos como K-means e Weighted K-means respeitaram os centroides fixos, enquanto Agglomerative Clustering apresentou subgrupos hierárquicos mais complexos. O DBSCAN sem centroides fixos detectou outliers significativos, enquanto sua versão com centroides ajustados ofereceu maior consistência visual.

Diversos gráficos foram elaborados para analisar e comparar os métodos:

- Silhouette Score: Indicou a qualidade dos agrupamentos para cada método, refletindo o quão bem os dados foram agrupados.

- Tempo de Execução: Avaliou a eficiência computacional de cada técnica, evidenciando diferenças no tempo necessário para processar os dados.
- Número de Outliers: Destaque para o DBSCAN, que, devido à sua abordagem baseada em densidade, identificou um número significativamente maior de outliers em comparação aos outros métodos.

5. DISCUSSÃO

O presente estudo apresentou uma análise detalhada e comparativa entre métodos de clusterização automáticos e a abordagem manual adotada por Gomes (2021) para regionalizar os municípios do Rio Grande do Sul. Enquanto os métodos automáticos, como K-means, Weighted K-means, Agglomerative Clustering e DBSCAN, foram aplicados para criar agrupamentos consistentes com base em critérios matemáticos, a abordagem manual trouxe um diferencial qualitativo importante ao integrar fatores sociais, econômicos e culturais na definição dos clusters.

5.1. Comparação entre os Métodos Automáticos e a Abordagem Manual

A análise dos resultados apresentados na tabela de divisão dos clusters (Tabela 2) evidencia uma certa proximidade entre os agrupamentos gerados automaticamente e os clusters manuais em áreas específicas. Por exemplo, os clusters de "Santa Maria" e "Frederico Westphalen" foram consistentes entre os métodos automáticos e a abordagem manual devido à relativa homogeneidade demográfica e geográfica dessas regiões. Isso reflete a capacidade dos métodos automáticos de capturar padrões objetivos em contextos onde as características regionais estão bem delimitadas.

Entretanto, discrepâncias significativas foram observadas em clusters como os de "Bagé" e "Porto Alegre". No caso de "Bagé", a abordagem manual de Gomes (2021) agrupou um número significativo de municípios com base em fatores socioeconômicos e históricos. Por outro lado, os métodos automáticos, como o K-means e o Agglomerative Clustering, tendem a distribuir os municípios entre clusters vizinhos devido à proximidade geográfica, ignorando nuances qualitativas. O cluster de "Porto Alegre", um polo econômico crucial,

apresentou melhor agrupamento na abordagem manual, refletindo maior interação econômica, enquanto os métodos automáticos priorizaram distâncias geométricas.

Tabela 2. Comparação quantidade de cidades

Cluster	K-means	Weighted K-means	Agglomerative Clustering	DBSCAN	Tese Gomes (2021)
1 (Porto Alegre)	24	27	24	7	21
2 (Caxias do Sul)	45	37	65	0	38
3 (Pelotas)	15	22	54	0	18
4 (Tramandaí)	28	27	20	10	19
5 (Passo Fundo)	29	45	18	20	57
6 (Lajeado)	53	50	24	140	32
7 (Santa Cruz do Sul)	33	28	63	10	28
8 (Santa Maria)	19	29	44	8	33
9 (Bagé)	11	8	23	0	8
10 (Novo Hamburgo)	64	35	23	0	57
11 (Erechim)	47	54	33	0	56
12 (Santa Rosa)	32	45	22	21	55
13 (Santo Ângelo)	45	20	3	8	20
14 (Uruguaiana)	9	8	56	0	8
15 (Frederico Westphalen)	45	64	27	117	57
Total	**499**	**499**	**499**	**341**	**507**

A Tabela 2 apresenta uma análise comparativa da quantidade de cidades atribuídas a cada cluster pelos métodos automáticos de clusterização e a abordagem manual utilizada por Gomes (2021). Observa-se que, em clusters como os de "Santa Maria" e "Frederico Westphalen", os métodos automáticos, especialmente o Weighted K-means, se aproximaram consideravelmente da divisão proposta por Gomes (2021), refletindo características geográficas e demográficas consistentes. Contudo, discrepâncias notáveis surgem em clusters como o de "Bagé" e "Lajeado", onde a abordagem manual, por levar em conta fatores qualitativos como relevância econômica e conectividade regional, apresentou agrupamentos

mais coerentes com a realidade prática do estado. Além disso, enquanto métodos como o K-means e o Agglomerative Clustering mantiveram consistência em clusters mais homogêneos, o DBSCAN revelou uma maior dispersão e um número significativo de outliers, indicando limitações no alinhamento com centroides fixos. Essas diferenças ressaltam a importância de combinar análises automáticas e qualitativas para uma regionalização mais fiel às particularidades locais.

5.2. Avaliação de Eficiência e Precisão

Com base na tabela de métricas de eficiência e precisão Tabela 1, os métodos apresentaram comportamentos distintos, para uma melhor visualização foi desenvolvido um gráfico Figura 7.

K-means demonstrou-se eficiente em termos de tempo de execução (0,0168 segundos) e produziu um bom Silhouette Score (0,371). No entanto, sua limitação em lidar com clusters de formas irregulares ficou evidente em áreas de maior complexidade geográfica, como "Porto Alegre".

Weighted K-means incorporou a dimensão populacional, resultando em agrupamentos mais próximos à abordagem manual em regiões densamente povoadas, como "Santa Cruz do Sul". Apesar disso, o tempo de execução mais alto (5,588 segundos) reflete a sobrecarga computacional causada pela ponderação.

Agglomerative Clustering destacou-se por fornecer insights hierárquicos e subgrupos internos, mas teve maior dificuldade em se alinhar aos centroides fixos devido à sua natureza iterativa. Seu tempo de execução (0,0323 segundos) foi competitivo, mas o Silhouette Score (0,3281) revelou desafios em regiões mais complexas.

DBSCAN apresentou as maiores discrepâncias, especialmente sem o uso de centroides fixos. Sua versão padrão detectou 80 outliers, mas com um Silhouette Score negativo (-0,0029), indicando baixa coesão dos clusters. A adaptação para centroides fixos eliminou os outliers, mas sacrificou parte de sua capacidade de refletir a densidade real dos dados.

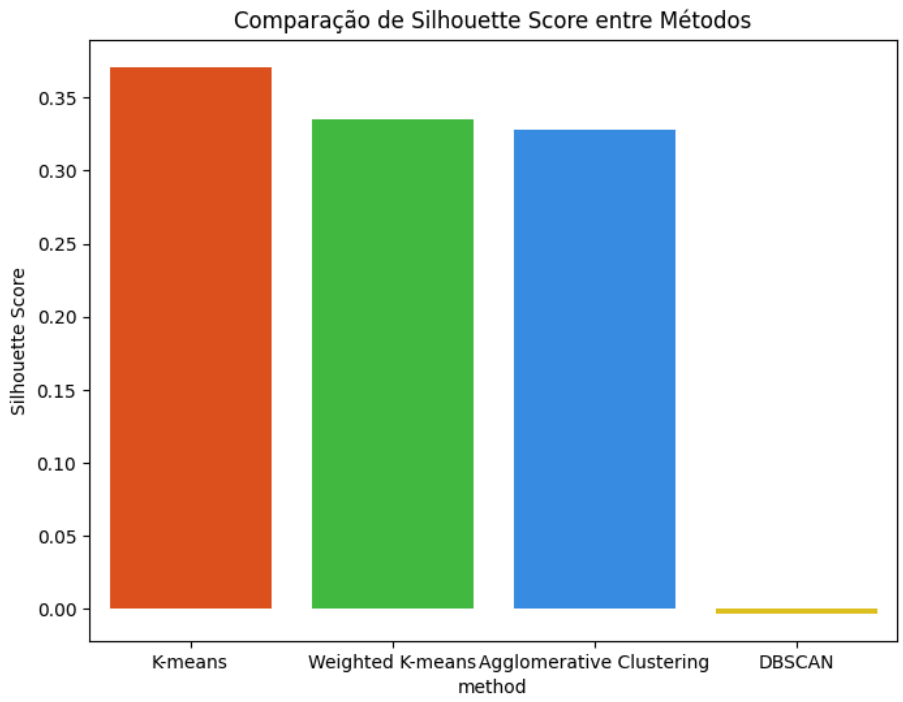

Figura 7. Comparação Silhouette Score

5.3. A Influência dos Centroides Fixos

A abordagem manual de Gomes (2021) destacou-se pelo uso criterioso de centroides fixos, definidos com base em critérios qualitativos, como relevância econômica, localização estratégica e características culturais das regiões do Rio Grande do Sul. Esses centroides serviram como polos para a regionalização, permitindo agrupamentos que refletissem não apenas proximidade geográfica, mas também conexões econômicas e sociais significativas. Inspirada por essa abordagem, este estudo introduziu centroides fixos nos métodos automáticos de clusterização, buscando aproximar os resultados quantitativos das divisões qualitativas propostas manualmente.

No caso de métodos como K-means e Weighted K-means, que dependem explicitamente da inicialização dos centroides, essa estratégia mostrou-se eficaz. Os agrupamentos formados a partir dos centroides fixos tiveram maior alinhamento com as divisões regionais estabelecidas por Gomes, especialmente em regiões com características demográficas e econômicas bem definidas. No entanto, para métodos baseados em densidade, como o DBSCAN, cuja lógica não considera centroides pré-definidos, os resultados foram mais desafiadores. A tentativa de ajustar o DBSCAN a padrões fixos comprometeu sua

capacidade de detectar clusters de forma arbitrária e de identificar outliers, suas principais vantagens.

A contribuição de Gomes ao enfatizar a importância de centroides cuidadosamente escolhidos destaca uma limitação crucial nos métodos automáticos: a ausência de contexto qualitativo. Este estudo demonstra que os centroides fixos podem ser uma ferramenta poderosa para alinhar métodos quantitativos a realidades regionais. Contudo, sua eficácia varia entre os métodos utilizados, reforçando a necessidade de adaptar a metodologia às características dos dados e ao objetivo da análise. A integração de práticas qualitativas, como a seleção manual de centroides, e métodos quantitativos pode ser o caminho para análises geoespaciais mais representativas e robustas.

5.4. Reflexões e Sugestões

Os resultados sugerem que a integração de métodos automáticos e abordagens manuais pode oferecer uma solução híbrida eficiente para análises regionais. A aplicação inicial de métodos como Weighted K-means pode servir como uma base quantitativa, seguida de ajustes manuais para incorporar fatores locais. Essa combinação seria particularmente útil em políticas públicas e planejamento urbano, onde aspectos como conectividade, infraestrutura e relevância econômica desempenham papeis críticos.

Por fim, a análise reafirma que não há um método universal para clusterização geoespacial. Enquanto os métodos automáticos, como K-means e Weighted K-means, são ideais para análises rápidas e replicáveis, a abordagem manual de Gomes (2021) ressalta a importância de capturar as nuances regionais e contextuais. Este estudo contribui ao demonstrar que a integração de abordagens pode produzir análises mais completas e alinhadas às demandas regionais e práticas.

6. CONCLUSÃO

Este estudo analisou diferentes métodos de clusterização aplicados a dados geoespaciais dos municípios do Rio Grande do Sul, comparando abordagens automáticas – como K-means, Weighted K-means, Agglomerative Clustering e DBSCAN – com a metodologia manual desenvolvida por Gomes (2021). Os resultados destacaram as forças e

limitações de cada técnica em termos de representatividade dos agrupamentos, eficiência computacional e adequação ao contexto regional.

Os métodos baseados em centroides fixos, especialmente K-means e Weighted K-means, mostraram maior alinhamento com os padrões regionais, sendo eficazes em cenários com clusters homogêneos e populações densas. O Weighted K-means, ao ponderar os dados pela população, aproximou-se mais das divisões manuais em áreas densamente povoadas. Já o Agglomerative Clustering forneceu insights hierárquicos, mas enfrentou dificuldades em respeitar os centroides fixos devido à sua natureza iterativa. Por sua vez, o DBSCAN, embora eficaz na detecção de outliers, perdeu parte de sua capacidade ao ser ajustado para centroides fixos, destacando suas limitações em contextos de regionalização predefinida.

A comparação com a abordagem manual evidenciou que, apesar da eficiência e replicabilidade dos métodos automáticos, eles não conseguem capturar fatores qualitativos, como características culturais, históricas e econômicas regionais. O trabalho manual de Gomes (2021) demonstrou superioridade em representar as especificidades locais, evidenciando a importância do conhecimento contextual em análises geoespaciais. Dessa forma, este estudo reforça a necessidade de integrar abordagens automáticas e qualitativas. Métodos automáticos podem fornecer uma base quantitativa eficiente, enquanto ajustes manuais são essenciais para refinar os resultados conforme as necessidades práticas. Essa combinação híbrida é especialmente relevante para aplicações como planejamento urbano e políticas públicas, onde aspectos regionais desempenham um papel crucial.

Por fim, este trabalho contribui para o avanço das análises geoespaciais ao destacar os benefícios de uma abordagem híbrida, abrindo espaço para estudos futuros que explorem novos parâmetros e combinações de técnicas. A integração entre métodos quantitativos e qualitativos se apresenta como um caminho promissor para análises mais completas e alinhadas às demandas regionais e globais.

7. REFERÊNCIAS

METZ, Ryan. *The Power of Machine Learning in Geospatial Analysis: Enhancing Decision-Making and Insights*. Medium, [s.d.]. 2023. Disponível em: https://ryankmetz.medium.com/the-power-of-machine-learning-in-geospatial-analysis-enhancing-decision-making-and-insights-287cf1bebc51. Acesso em: 13 nov. 2024.

AYA DATA. *The Ultimate Guide to Geospatial Data Science*. Aya Data, [s.d.]. Disponível em: https://www.ayadata.ai/the-ultimate-guide-to-geospatial-data-science/. Acesso em: 13 nov. 2024.

URBAN SPATIAL. *Introduction to Geospatial Machine Learning: Part 1*. Public Policy Analytics, [s.d.]. Disponível em: https://urbanspatial.github.io/PublicPolicyAnalytics/intro-to-geospatial-machine-learning-part-1.html#intro-to-geospatial-machine-learning-part-1. Acesso em: 13 nov. 2024.

REY, Sergio; ANSELIN, Luc; KNAAP, Elijah; WOLF, Levi. *Clustering and Regionalization*. Geographic Data Science with Python, [s.d.]. Disponível em: https://geographicdata.science/book/notebooks/10_clustering_and_regionalization.html. Acesso em: 13 nov. 2024.

ESRI. *Spatially Constrained Multivariate Clustering*. ArcGIS Pro, [s.d.]. Disponível em: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/spatially-constrained-multivariate-clustering.htm. Acesso em: 13 nov. 2024.

GOMES, Aline Pimentel. *Proposta de regionalização para constituição de consórcios públicos para gestão integrada de resíduos sólidos urbanos*. 2021. 368 f. Tese (Doutorado em Engenharia Civil e Ambiental) - Universidade de Passo Fundo, Passo Fundo, RS, 2021.