Análise Comparativa de Diferentes Algoritmos de Clusterização para Poluição Atmosférica, Condições Climáticas e Internações Hospitalares na cidade de Ponta Grossa, Paraná

Lucas Rover
Programa de Pós-Graduação em Engenharia Mecânica
Universidade Tecnológica Federal do Paraná (UTFPR)
Ponta Grossa – PR, Brasil
lucasrover@alunos.utfpr.edu.br

Profa. Dra. Yara de Souza Tadano
Programa de Pós-Graduação em Engenharia Mecânica
Universidade Tecnológica Federal do Paraná (UTFPR)
Ponta Grossa – PR, Brasil
yaratadano@utfpr.edu.br

Prof. Dr. Hugo Valadares Siqueira
Ministério da Ciência, Tecnologia e Inovação
Universidade Tecnológica Federal do Paraná (UTFPR)
Ponta Grossa – PR, Brasil
hugosiqueira@utfpr.edu.br

***Abstract*—A dinâmica entre poluição atmosférica, clima e saúde é complexa e não-linear, dificultando a identificação de padrões. Este estudo aplicou e comparou cinco algoritmos de clusterização (K-Means, Agglomerative, DBSCAN, GMM e Spectral Clustering) para detectar regimes ambientais associados a internações por doenças respiratórias e circulatórias em 577 dias de observação (2016–2018) em Ponta Grossa (PR). Foram analisadas 10 variáveis ambientais, com exclusão de variáveis de desfecho para garantir agrupamentos baseados apenas em perfis climáticos e de poluição. O K-Means apresentou melhor desempenho (Silhouette = 0,192; Davies-Bouldin = 1,635; Calinski-Harabasz = 125,8), revelando 4 clusters ambientalmente coesos. O cluster "frio e seco com vento forte" apresentou as maiores médias de internações respiratórias, enquanto o cluster "chuvoso e úmido" registrou os menores níveis. A análise por PCA confirmou separabilidade parcial entre os grupos. Os resultados reforçam a eficácia da clusterização na análise exploratória de sistemas ambientais complexos e oferecem subsídios para políticas de saúde ambiental**

***Index Terms*—poluição atmosférica, saúde populacional, clusterização, aprendizado não supervisionado.**

I. INTRODUÇÃO

A qualidade do ar é um fator determinante para a saúde pública global, com forte associação entre a exposição a poluentes atmosféricos e o aumento de doenças respiratórias e cardiovasculares [1], [2]. Compreender a dinâmica complexa entre níveis de poluição, variações climáticas e os impactos resultantes na saúde populacional é crucial para o desenvolvimento de políticas públicas eficazes e estratégias de mitigação [3]. Contudo, a natureza multifatorial e frequentemente não linear dessas interações representa um desafio significativo para métodos analíticos tradicionais [4].

Técnicas de Inteligência Computacional, como algoritmos de clusterização, oferecem abordagens poderosas para a análise exploratória de grandes volumes de dados multidimensionais, permitindo a identificação de padrões e estruturas intrínsecas, que podem não ser aparentes a priori [5], [6]. A clusterização pode agrupar dias ou períodos com características ambientais, climáticas e de saúde semelhantes, revelando "regimes" ou "cenários" típicos.

Este trabalho tem o objetivo de comparar o desempenho de diferentes algoritmos de clusterização para identificação de padrões climáticos em estudos de poluição atmosférica e seus impactos à saúde. Para tanto, um conjunto de dados integrado da cidade de Ponta Grossa, Paraná, abrangendo o período entre 2016 e 2018 será utilizado. O conjunto de dados inclui concentrações diárias de material particulado fino ($MP_{2.5}$), diversas variáveis meteorológicas e contagens diárias de internações hospitalares por causas respiratórias (CID-10 J01-J99) e circulatórias (CID-10 I01-I99). O objetivo principal é identificar grupos (clusters) de dias com perfis distintos de poluição, clima e saúde, e caracterizar esses perfis para entender melhor as condições associadas a diferentes níveis de risco à saúde na população local.

A contribuição deste artigo reside na demonstração metodológica e aplicada de como os algoritmos de clusterização podem ser usados para desvendar padrões complexos e não-lineares que associam condições ambientais e de poluição a desfechos de saúde pública em um contexto específico, fornecendo insights acionáveis para a saúde local. Além disso, ele traz uma comparação sistemática de performances da aplicação de cinco diferentes algoritmos de clusterização (K-Means, Agglomerative Clustering, DBSCAN, GMM, Spectral Clustering) usando métricas de validação interna (Silhouette, Davies-Bouldin, Calinski-Harabasz).

O restante do artigo está organizado da seguinte forma: A Seção II apresenta uma breve revisão de trabalhos relacionados

que utilizam clusterização em contextos similares. A Seção III detalha a metodologia empregada, incluindo a coleta, pré-processamento dos dados e a aplicação. A Seção IV apresenta e discute os resultados da clusterização. Finalmente, a Seção V conclui o trabalho e aponta direções futuras.

II. TRABALHOS RELACIONADOS

A aplicação de algoritmos de clusterização para analisar dados de poluição atmosférica e meteorológicos tem sido muito explorada na literatura. Técnicas como K-Means, análise hierárquica e DBSCAN são frequentemente utilizadas para identificar padrões espaciais ou temporais, agrupar estações de monitoramento com perfis similares ou caracterizar dias com diferentes condições de qualidade do ar.

Paulose et al. [7] utilizaram K-Means para agrupar regiões de Delhi, Índia, com base nas concentrações de múltiplos poluentes ($PM_{2,5}$, PM_{10}, NO_2, SO_2), identificando áreas de maior risco e padrões sazonais claros, correlacionando-os com doenças respiratórias crônicas. Grace et al. [8] desenvolveram uma versão aprimorada do K-Means para analisar dados de sensores de poluição em tempo real, superando limitações da inicialização aleatória e demonstrando melhor acurácia na determinação do Índice de Qualidade do Ar (AQI) em comparação com outros métodos.

Godoy et al. [9] aplicaram K-medoids e regras de associação (Apriori) para analisar concentrações de $MP_{2,5}$ no estado de São Paulo, identificando grupos de estações (metropolitanas vs. afastadas) e padrões sazonais (picos no inverno), além de associar períodos críticos de poluição a condições meteorológicas adversas (baixa umidade, vento fraco, baixas temperaturas). Outros estudos exploraram métodos híbridos, como a combinação de DBSCAN com Redes Neurais Profundas [10] ou CNN-LSTM [11], visando capturar características espaço-temporais mais complexas para fins de previsão, embora a clusterização frequentemente sirva como uma etapa exploratória inicial ou componente desses modelos.

Apesar dos avanços na área, poucos estudos exploram a inclusão de variáveis relacionadas à saúde populacional na clusterização. Desta forma, a análise integrada de dados de poluição, múltiplas variáveis climáticas e dados de saúde (como internações hospitalares) por meio de clusterização ainda oferece oportunidades, especialmente para cidades fora dos grandes centros metropolitanos globais, que também são fortemente impactadas pela poluição atmosférica, visto que nenhum nível de concentração de $MP_{2,5}$ pode ser considerado seguro para a saúde humana [17].

III. METODOLOGIA

Este estudo configura-se como uma análise observacional, retrospectiva e exploratória, utilizando dados para investigar associações entre fatores ambientais (condições climáticas e poluição atmosférica) e desfechos de saúde (internações hospitalares por doenças respiratórias e circulatórias) na cidade de Ponta Grossa, Paraná, Brasil. O objetivo principal foi comparar os diferentes métodos de clusterização aplicados na identificação de padrões ambientais diários distintos e caracterizar esses padrões em relação às médias de internações hospitalares observadas.

O conjunto de dados abrange o período de 1 ° de outubro de 2016 a 30 de abril de 2018 (577 observações diárias) [12]. As variáveis incluídas e suas respectivas fontes foram:

- **Poluição Atmosférica ($MP_{2,5}$)**: Medições diárias de concentração de Material Particulado fino ($\mu g\, m^{-3}$), obtidas com amostrador inercial Harvard [12].
- **Dados Meteorológicos**: Variáveis diárias obtidas do Sistema de Tecnologia e Monitoramento Ambiental do Paraná (SIMEPAR) e do Instituto Nacional de Meteorologia (INMET), incluem: temperatura máxima (C), média (C), mínima (C), umidade relativa (%), precipitação (mm), radiação solar (W/m^2), pressão atmosférica (hPa), direção do vento (graus) e velocidade do vento (m/s).
- **Dados de Saúde**: Contagens diárias de internações hospitalares por Doenças do Aparelho Respiratório (CID-10 J01-J99) e por Doenças do Aparelho Circulatório (CID-10 I01-I99), extraídas do Departamento de Informática do Sistema Único de Saúde (DataSUS) [13] para residentes de Ponta Grossa.

A. Pré-processamento dos Dados

O pré-processamento dos dados foi realizado em etapas sequenciais rigorosas utilizando a linguagem Python e a biblioteca Pandas para assegurar a qualidade analítica dos dados:

1) **Carregamento e Limpeza Inicial:** Os dados foram carregados e seus formatos foram padronizados para facilitar a manipulação. Colunas de metadados e a colunas reduntantes foram removidas. Além disso, garantiu-se a ordenação cronológica das observações e assegurou-se que todas as colunas destinadas à análise numérica fossem convertidas para tipos numéricos apropriados (inteiro ou ponto flutuante).
2) **Tratamento de Dados Ausentes:** foi realizada uma análise quantitativa (`isnull().sum()`) e visual (`missingno.matrix`) dos dados ausentes e, baseado nisso, fez-se a imputação utilizando a mediana de cada variável. Essa abordagem evita a distorção que a média poderia causar em distribuições assimétricas e preserva a integridade temporal do dataset.
3) **Escaolamento de Features para Clusterização:** Para a análise de clusterização, foram selecionadas exclusivamente as seguintes dez variáveis ambientais a partir de uma análise prévia de correlação:`mp25`, `dir_vento` (direção do vento), `prec`, `rad_solar` (radiação solar), `temp_max` (temperatura máxima), `temp_med` (temperatura média), `temp_min` (temperatura mínima), `umid_rel` (umidade relativa), `pressao`, `vel_vento` (velocidade do vento). O objetivo foi agrupar os dias com base em seus perfis ambientais intrínsecos. Todas as variáveis foram nromalizadas por padronização (Standardization), implementada pela classe `StandardScaler` do Scikit-learn.

A padronização transforma os dados para terem média = 0 e desvio padrão = 1. Isso é crucial para algoritmos de clusterização sensíveis à escala das variáveis (como K-Means, DBSCAN, Hierárquico), garantindo que todas as variáveis contribuam de forma equitativa para as métricas de distância ou variância, sem que variáveis com magnitudes maiores dominem o processo. Os dados originais foram mantidos para a EDA e interpretação dos clusters.

B. Análise Exploratória de Dados (EDA)

Uma EDA foi conduzida sobre os dados pré-processados (antes do escalonamento) para compreender suas características e identificar padrões iniciais:

- **Estatísticas Descritivas:** Calculadas (média, mediana, desvio padrão, quartis, mínimo, máximo) para resumir a tendência central, dispersão e amplitude das variáveis numéricas.
- **Análise de Distribuição:** Histogramas e gráficos de densidade de Kernel (KDE) foram gerados (utilizando Seaborn) para visualizar a forma da distribuição de variáveis chave (internações, $MP_{2,5}$, variáveis climáticas selecionadas) e avaliar simetria ou assimetria.
- **Análise de Outliers:** Box plots (Seaborn) foram utilizados para identificar visualmente valores atípicos. Decidiu-se por manter os outliers, justificando que poderiam representar eventos ambientais extremos reais relevantes para a análise.
- **Análise de Correlação:** A matriz de correlação de Pearson foi calculada entre todas as variáveis numéricas relevantes para identificar relações lineares.

C. Análise de Clusterização Comparativa

O núcleo da análise consistiu na aplicação e comparação de cinco algoritmos de clusterização distintos ao conjunto de dados ambientais padronizados. Foram selecionados os seguintes algoritmos, no intuito de contemplar diferentes pressupostos sobre a estrutura dos dados:

- K-Means: Baseado em centróides, particional, adequado para clusters aproximadamente esféricos e amplamente usado em estudos ambientais.
- Agglomerative Clustering (Ward): Hierárquico aglomerativo, utilizando o critério de minimização da variância de Ward, captura hierarquias e pode identificar estruturas mais complexas.
- DBSCAN: Baseado em densidade, útil para dados com ruído e clusters de densidade variável, embora dependa fortemente da calibração dos parâmetros.
- Gaussian Mixture Models (GMM): Baseado em modelos probabilísticos de mistura, é ideal para dados com ruído e clusters de densidade variável, embora dependa fortemente da calibração dos parâmetros.
- Spectral Clustering: Baseado em grafos e análise espectral, indicado para dados não lineares, explorando a conectividade via grafos.

A seleção visou abranger diferentes pressupostos sobre a estrutura dos dados (clusters esféricos, hierárquicos, de densidade variável, formas arbitrárias, etc.), permitindo uma avaliação comparativa robusta da adequação de cada método ao dataset.

D. Determinação de Parâmetros

Para K-Means, Agglomerative, GMM, Spectral: O número ótimo de clusters (k) foi investigado utilizando o Método do Cotovelo (Elbow Method), analisando a inércia intra-cluster (Soma dos Quadrados Intra-clusters - WCSS) para K-Means, e a Análise de Silhueta, calculando o score médio de silhueta para cada k em um intervalo de 2 a 15 clusters. Com base na inspeção visual dos gráficos resultantes e na busca por um ponto de inflexão (cotovelo) ou pico (silhueta), $k = 4$ foi determinado como o número mais apropriado para estes algoritmos neste dataset.

Para DBSCAN, os parâmetros `eps` (raio da vizinhança) e `min_samples` (número mínimo de pontos na vizinhança) foram determinados da seguinte forma: `min_samples` foi definido heuristicamente como 2 vezes o número de features ($2 \times 10 = 20$). O valor de `eps` foi estimado analisando o gráfico de k-distance (distância para o 20º vizinho mais próximo), identificando o "cotovelo" (ponto de máxima curvatura) no gráfico, que ocorreu aproximadamente em $\texttt{eps} = 3.0$.

Cada um dos cinco algoritmos foi ajustado ao conjunto de dados ambientais padronizado (`X_cluster`), e os rótulos de cluster resultantes para cada ponto de dado (dia) foram armazenados.

E. Avaliação e Comparação dos Métodos

A qualidade dos clusters gerados por cada algoritmo foi avaliada utilizando métricas de validação interna, que não requerem conhecimento prévio dos "clusters verdadeiros". As métricas selecionadas foram:

- **Coeficiente de Silhueta:** Mede a coesão intra-cluster e a separação inter-cluster (valores próximos de 1 são melhores).
- **Índice de Davies-Bouldin:** Mede a razão entre a dispersão intra-cluster e a separação inter-cluster (valores próximos de 0 são melhores).
- **Índice de Calinski-Harabasz:** Mede a razão entre a variância inter-cluster e a variância intra-cluster (valores mais altos são melhores).

Os scores dessas métricas foram calculados para os resultados de cada algoritmo (exceto para DBSCAN, que falhou em produzir mais de um cluster significativo com os parâmetros definidos). Os algoritmos foram classificados com base nesses scores para determinar qual produziu a estrutura de cluster mais estatisticamente significativa e robusta para o dataset estudado.

F. Caracterização e Interpretação dos Clusters

Para interpretar o significado dos agrupamentos encontrados pelo algoritmo de melhor desempenho (K-Means com $k = 4$) foram adicionados os rótulos de cluster de volta ao DataFrame original (não escalonado). As médias e medianas de todas as variáveis ambientais e das variáveis de internação (`int_resp`,

int_circ) foram calculadas para cada um dos 4 clusters. A contagem de dias pertencentes a cada cluster foi determinada.

Com base nessas estatísticas descritivas por cluster, foi realizada uma interpretação qualitativa para descrever o perfil ambiental típico de cada cluster e analisar se havia associações aparentes entre esses perfis e os níveis médios de internações respiratórias e circulatórias.

Para visualização, a técnica de Análise de Componente Principal (ACP) foi utilizada para reduzir a dimensionalidade das 10 features ambientais para 2 componentes principais. Um scatter plot foi gerado para visualizar os dias (pontos) coloridos por seus respectivos rótulos de cluster K-Means no espaço ACP, auxiliando na avaliação visual da separação dos clusters.

G. Ferramentas Computacionais

Toda a análise de dados, pré-processamento, implementação dos algoritmos de clusterização, cálculo de métricas e geração de visualizações foram realizados utilizando a linguagem de programação Python e as seguintes bibliotecas científicas principais:

- Pandas: Para manipulação e análise de dados tabulares.
- NumPy: Para operações numéricas eficientes.
- Scikit-learn: Para implementação dos algoritmos de clusterização (KMeans, AgglomerativeClustering, DBSCAN, GaussianMixture, SpectralClustering), pré-processamento (StandardScaler, SimpleImputer), cálculo de métricas de avaliação (silhouette_score, davies_bouldin_score, calinski_harabasz_score) e ACP.
- Matplotlib e Seaborn: Para geração de gráficos estáticos (histogramas, box plots, séries temporais, heatmaps, scatter plots).
- Missingno: Para visualização de padrões de dados ausentes.

IV. RESULTADOS E DISCUSSÃO

O conjunto de dados foi submetido a um fluxo robusto de pré-processamento, essencial para garantir a qualidade analítica e a comparabilidade entre variáveis com diferentes escalas e distribuições conforme descrito na seção metodológica. Para a análise exploratório dos dados, foram aplicadas estatísticas descritivas e gráficos para caracterizar o comportamento das variáveis ambientais, de poluição e de saúde.

A. Análise Exploratória de Dados (EDA)

O conjunto de dados compreendeu 577 observações, que foram previamente tratadas, e contém variáveis ambientais, meteorológicas e de saúde pública. A análise descritiva mostrada na Tabela I fornece uma visão inicial sobre a distribuição, centralidade e dispersão dessas variáveis.

As estatísticas descritivas revelam alta variabilidade intra-anual em variáveis como precipitação, radiação, $MP_{2,5}$ e temperatura mínima. A dispersão nos dados e a presença de assimetrias justificam o uso de técnicas robustas de normalização (como o StandardScaler) e a aplicação de métodos de clusterização

TABLE I: Estatísticas Descritivas das Variáveis Analisadas (N=577)

Variável	count	mean	std	min	25%	50%	75%	max
int_resp	577,00	5,43	3,06	0,00	3,00	5,00	7,00	16,00
int_circ	577,00	9,29	4,69	0,00	6,00	9,00	12,00	27,00
mp25	577,00	10,38	8,48	0,93	5,82	8,27	11,98	79,16
dir_vento	577,00	144,12	101,27	0,00	90,00	90,00	270,00	315,00
prec	577,00	4,63	11,07	0,00	0,00	0,00	2,80	84,80
rad_solar	577,00	881,51	239,59	176,70	713,80	921,00	1073,00	1357,00
temp_max	577,00	25,24	4,15	11,20	22,10	25,70	28,40	33,70
temp_med	577,00	18,97	3,25	4,46	16,77	19,57	21,24	25,41
temp_min	577,00	14,53	3,44	0,00	12,70	15,00	17,20	20,30
umid_rel	577,00	78,05	8,33	48,39	73,15	78,65	84,35	96,45
pressao	577,00	915,71	3,74	906,58	913,17	915,29	918,12	927,55
vel_vento	577,00	3,13	1,16	0,84	2,30	3,00	3,83	7,87
diasem	577,00	4,00	2,01	1,00	2,00	4,00	6,00	7,00
feriado	577,00	0,10	0,31	0,00	0,00	0,00	0,00	1,00

multivariada, capazes de identificar padrões não lineares e segmentar regimes ambientais com maior precisão.

Essa análise oferece uma base sólida para interpretar os agrupamentos obtidos e suas implicações em saúde pública, considerando tanto os eventos extremos (outliers) quanto os padrões centrais do sistema clima-poluição-saúde.

B. Análise de Distribuição de Variáveis Chave

A análise exploratória das distribuições empíricas das variáveis-chave revela características assimétricas e não-normais que influenciam diretamente a escolha dos métodos estatísticos e a interpretação dos resultados, conforme demonstrado na Figura 1.

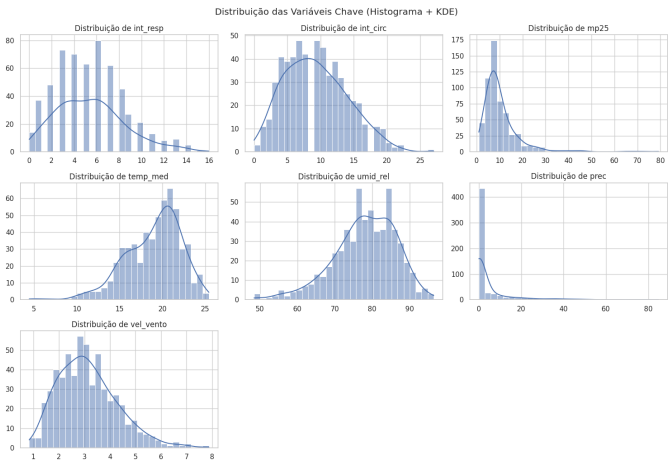

Fig. 1: Distribuição das Variáveis Chave (Histograma + KDE).

As variáveis de saúde (int_resp e int_circ), apresentam distribuições fortemente assimétricas à direita. Essa estrutura sugere a presença de dias atípicos ou surtos localizados, os quais, embora raros, exercem impacto relevante sobre a média. Este tipo de assimetria reforça a necessidade de análises baseadas em medidas robustas de tendência central, como a mediana, além de cuidados na modelagem preditiva para não superestimar padrões baseados em outliers.

A distribuição da variável mp25 (concentração de $MP_{2,5}$) também é assimétrica à direita, com grande concentração de valores altos (acima de $10\,\mu g\,m^{-3}$) e uma cauda longa representando episódios de poluição atmosférica acentuados. A variável

`temp_med` (temperatura média) apresenta uma distribuição relativamente simétrica, possivelmente com bimodalidade leve, refletindo a sazonalidade do clima local (ex: verão vs inverno). Essa distribuição sugere que o uso de medidas como média e desvio padrão é apropriado neste caso, e pode indicar a existência de dois regimes térmicos relativamente distintos ao longo do ano.

A variável `umid_rel` (umidade relativa) mostra uma assimetria à esquerda, com forte concentração de valores elevados (acima de 70 %). Essa distribuição é típica de climas úmidos, onde há pouca variação em relação à umidade, exceto em dias de secura atípica. Tais observações com umidade anormalmente baixa podem, no entanto, estar associadas a picos de internações respiratórias, sendo importantes para análises de associação. A `prec` (precipitação) é uma das variáveis mais fortemente assimétricas à direita, com mediana igual a zero, indicando que a maioria dos dias não registra chuva significativa. A cauda longa representa poucos dias de chuvas intensas, o que é comum em regiões com regimes pluviométricos concentrados. Essa estrutura justifica o uso de métodos não-paramétricos ou transformações logarítmicas em análises de regressão. A `vel_vento` (velocidade do vento) também exibe assimetria à direita, com maioria dos registros abaixo de 4 m/s e uma cauda com ventos mais intensos. Essa distribuição impacta a dispersão de poluentes e pode modular os efeitos da qualidade do ar sobre a saúde.

C. Análise de Outliers com Box Plots

Durante a análise exploratória de dados, foi possível identificar a presença de outliers potenciais em diversas variáveis ambientais e de saúde. Esses valores extremos foram evidenciados, principalmente, por pontos fora dos "whiskers" observado na Figura 2, indicando observações que se distanciam significativamente do intervalo interquartil (IQR).

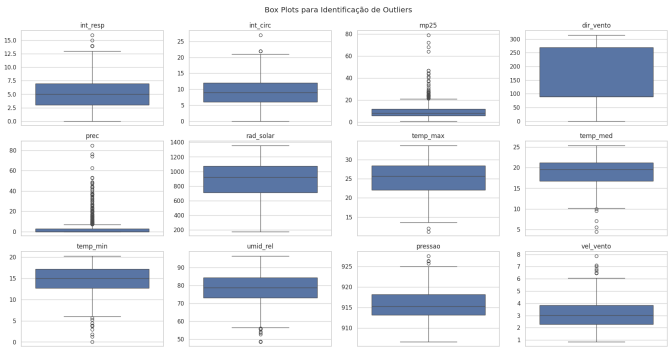

Fig. 2: Box Plots para Identificação de Outliers.

A presença desses outliers em variáveis como $MP_{2,5}$, precipitação e temperaturas, com destaque para os valores extremos de poluição e chuva intensa. Esses outliers, longe de representarem erros, são plausíveis do ponto de vista ambiental e refletem eventos reais e críticos — como dias altamente poluídos ou com precipitação extrema. Espera-se que as técnicas de clusterização adotadas sejam robustas e capazes de agrupar esses dias atípicos em padrões específicos, preservando a integridade e a riqueza dos dados.

D. Análise de Correlação

A análise de correlação linear entre as variáveis ambientais e os desfechos de saúde (internações por causas respiratórias e circulatórias) demonstrou associações fracas ou nulas na maioria dos casos. As relações mais expressivas, foram entre internações por doenças circulatórias e respitarórias (0,36) e internações por doenças respiratórias e temperatura mínima (-0,34) indicando que relações lineares simples não capturam bem a complexidade subjacente entre ambiente e saúde. Tal evidência justifica a adoção de métodos não supervisionados, como a clusterização, para identificação de padrões multivariados latentes que escapam às análises tradicionais.

E. Determinação do Número de Clusters (k)

A definição do número de clusters para os algoritmos K-Means, Agglomerative Clustering (Ward), Gaussian Mixture Model (GMM) e Spectral Clustering foi baseada em uma abordagem exploratória combinada, utilizando o método do cotovelo sobre a inércia (soma das distâncias intra-cluster) e a análise do coeficiente de Silhueta para diferentes valores de k. Os gráficos apresentados nas Figuras 3 e 4, mostram que a configuração com 4 clusters representou o melhor equilíbrio entre parcimônia e separabilidade, sendo o ponto de inflexão mais evidente no gráfico de inércia, ao mesmo tempo em que maximizava a separação e coesão entre os grupos no índice de Silhueta.

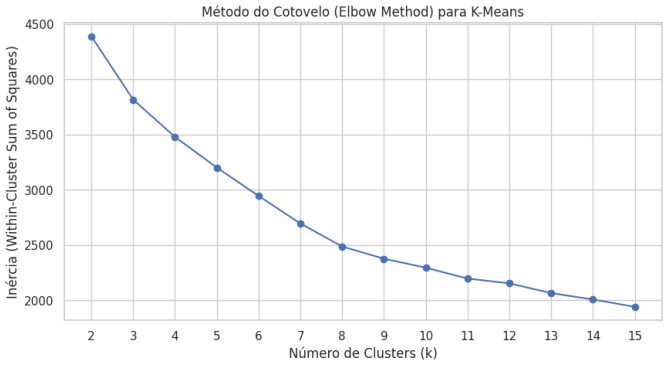

Fig. 3: Método do Cotovelo (Elbow Method) para K-Means.

A clusterização foi realizada com base em 10 variáveis ambientais e de poluição (`mp25`, `dir_vento`, `prec`, `rad_solar`, `temp_max`, `temp_med`, `temp_min`, `umid_rel`, `pressao`, `vel_vento`), a fim de garantir que os clusters refletissem estruturas ambientais puras, sem interferência direta de resultados ou datas.

Para o DBSCAN, que não requer a definição prévia de k, os parâmetros foram ajustados com base no gráfico k-distance (Figura 5), definindo $\texttt{eps} = 3.0$ como o ponto de inflexão mais representativo e $\texttt{min\_samples} = 20$, seguindo a heurística de número de dimensões. No entanto, o DBSCAN identificou apenas um cluster principal e 13 pontos de ruído, sugerindo

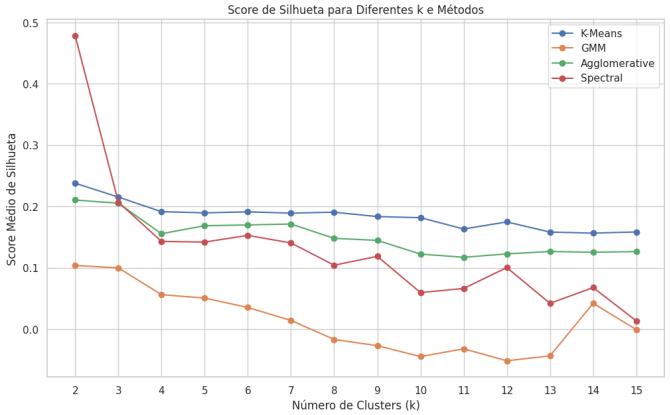

Fig. 4: Score de Silhueta para Diferentes k e Métodos.

que os dados não possuem densidades locais bem definidas sob esses parâmetros.

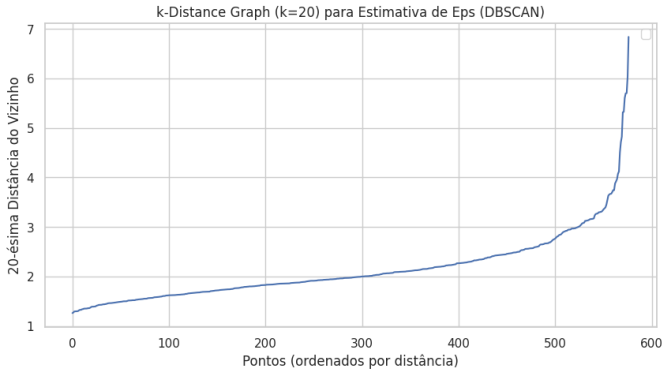

Fig. 5: k-Distance Graph (k=20) para Estimativa de Eps (DBSCAN).

Assim, a escolha de $k = 4$ se mostrou mais adequada para os algoritmos baseados em partição, hierarquia, distribuição e grafos, permitindo a identificação de padrões climáticos distintos e regimes ambientais coerentes, com suporte quantitativo robusto.

F. Avaliação e Comparação de Métodos

Com base na avaliação quantitativa das métricas de validação interna (Tabela II), o algoritmo K-Means apresentou o melhor desempenho geral entre os métodos de clusterização testados. A métrica Silhouette, que mede a coesão interna dos clusters em relação à separação entre grupos, alcançou um valor de 0,192 — o mais alto entre todos os modelos. Além disso, o índice Davies-Bouldin (1,635), que penaliza a sobreposição entre clusters, reforça essa qualidade ao indicar uma separação razoável entre os grupos formados. Complementando essa performance, o valor de Calinski-Harabasz (125,8) sugere que os clusters formados pelo K-Means são bem definidos, com alta variância intergrupo em relação à variância intragrupo. Esses resultados, em conjunto, sinalizam que o K-Means gerou agrupamentos com boa distinção estrutural e coesão interna.

O segundo melhor desempenho foi observado com o algoritmo Agglomerative (Ward), cujo valor de Silhouette foi de 0,156 e o índice Calinski-Harabasz atingiu 100,5. Apesar de ligeiramente inferiores ao K-Means, esses resultados ainda indicam uma estrutura de agrupamento com razoável coesão e separação. A abordagem hierárquica utilizada no Agglomerative pode ter capturado nuances estruturais nos dados, especialmente em contextos em que a forma dos clusters não é perfeitamente esférica.

Já o método Spectral Clustering apresentou uma avaliação mista. Embora tenha alcançado um bom valor de Davies-Bouldin (1,035) — o mais baixo entre os algoritmos testados, o que teoricamente indicaria clusters bem separados —, as demais métricas foram menos favoráveis. A Silhouette (0,143) e o índice Calinski-Harabasz (12,995) foram substancialmente mais baixos, sugerindo que os clusters gerados são menos compactos e apresentam maior sobreposição interna. Isso pode indicar que, embora os centróides estejam relativamente afastados, os dados dentro de cada cluster são mais dispersos ou mal definidos.

Enquanto o algoritmo Gaussian Mixture Model (GMM) teve o desempenho mais fraco entre os métodos baseados em k. A Silhouette foi a menor de todas (0,056), indicando fraca coesão intracluster e proximidade entre clusters distintos. Além disso, o índice Davies-Bouldin atingiu 2,823, sinalizando sobreposição elevada entre os grupos. O índice Calinski-Harabasz (37,561) também reforça esse diagnóstico de má separação e estrutura pouco definida nos agrupamentos. Esses resultados sugerem que a suposição de distribuição normal implícita no GMM pode não ser adequada para os dados analisados, resultando em clusters com baixa qualidade estrutural.

Por fim, o resultado do método DBSCAN encontrou 1 cluster e 13 pontos de ruído e as métricas indicam que, com `eps` $= 3.0$ e `min_samples` $= 20$, o algoritmo considerou quase todos os pontos como pertencentes a um único grande cluster de alta densidade (ou interconectados), classificando apenas alguns como ruído. Isso sugere que os dados não têm uma estrutura de densidade clara que o DBSCAN possa capturar com esses parâmetros, ou os parâmetros precisam de um ajuste mais fino (talvez um eps menor ou `min_samples` diferente).

TABLE II: Métricas de Validação Interna por Algoritmo (k=4, exceto DBSCAN)

Algoritmo	Silhouette	Davies-Bouldin	Calinski-Harabasz	Num_Clusters
K-Means	0,192	1,635	125,847	4,000
Agglomerative	0,156	1,884	100,505	4,000
Spectral	0,143	1,035	12,995	4,000
GMM	0,056	2,823	37,561	4,000
DBSCAN	−1,000	inf	0,000	1,000

G. Caracterização dos Clusters K-means

Com base na análise das médias dos clusters gerados via K-Means, é possível caracterizar distintos regimes ambientais e seus respectivos impactos sobre indicadores de saúde. A partir das Tabelas III e IV, são descritos os perfis de cada cluster, utilizando as médias, e tendo as medianas como apoio confirmatório da consistência desses padrões.

TABLE III: Medianas das Variáveis por Cluster (K-Means, k=4)

Cluster	mp25	dir_vento	prec	rad_solar	temp_max	temp_med	temp_min	umid_rel	pressao	vel_vento	int_resp	int_circ
Cluster 0	7,18	90,00	13,90	631,00	23,45	19,38	16,60	87,32	913,73	2,81	4,00	8,00
Cluster 1	7,96	90,00	0,00	720,00	21,30	15,35	11,20	77,64	919,14	3,83	7,00	9,00
Cluster 2	8,27	315,00	0,00	1055,50	27,15	20,62	16,65	80,65	914,55	2,86	5,00	9,00
Cluster 3	8,62	90,00	0,00	1044,50	28,50	21,18	15,80	73,88	914,11	2,81	5,00	9,00

TABLE IV: Médias das Variáveis por Cluster (K-Means, k=4)

Cluster	mp25	dir_vento	prec	rad_solar	temp_max	temp_med	temp_min	umid_rel	pressao	vel_vento	int_resp	int_circ
Cluster 0	7,96	125,24	19,16	645,90	23,36	18,99	16,15	87,20	913,52	2,93	4,72	8,28
Cluster 1	10,15	127,64	0,45	759,64	20,92	15,00	10,74	77,50	919,53	3,65	6,72	9,70
Cluster 2	10,09	302,69	2,70	1013,16	27,21	20,71	16,28	79,10	914,57	2,88	4,78	9,40
Cluster 3	11,94	84,98	1,36	1029,07	28,52	21,12	15,72	73,20	914,47	2,95	5,13	9,43

1) Cluster 0 – "Chuvoso e Úmido" (106 dias): Este grupo é caracterizado por um ambiente predominantemente chuvoso, com precipitação média de 19,16 mm, alta umidade relativa (87,2 %) e radiação solar reduzida (645,9 W/m^2) — condições típicas de dias encobertos ou fortemente nublados. As temperaturas se mantêm em níveis moderados (temperatura média de 19 C) e a pressão atmosférica é relativamente baixa (913,52 hPa), o que reforça a presença de sistemas meteorológicos de instabilidade. No que diz respeito à poluição, os níveis de material particulado fino ($MP_{2,5}$) são os mais baixos entre os clusters (7,96 µg m^{-3}), sugerindo que a alta umidade e as chuvas contribuem significativamente para a dispersão e deposição dos poluentes atmosféricos. Do ponto de vista da saúde pública, este cluster está associado às menores médias de internações por doenças circulatórias (8,28) e respiratórias (4,72), indicando um ambiente menos agressivo tanto em termos térmicos quanto de qualidade do ar.

2) Cluster 1 - "Frio e Vento Forte" (159 dias): Este cluster representa um ambiente marcadamente frio, com a temperatura média mais baixa (15,0 C) e mínima de 10,7 C. A pressão atmosférica é a mais alta entre os grupos (919,53 hPa), e a velocidade média do vento (3,65 m/s) também é superior, refletindo a atuação de massas de ar mais densas e secas. A precipitação é baixa (0,45 mm) e a umidade relativa reduzida (77,5 %), compondo um cenário frio e seco. A radiação solar é intermediária-baixa, e os níveis de $MP_{2.5}$ são intermediários (10,15 µg m^{-3}). Este regime climático se relaciona com os piores desfechos em saúde respiratória, apresentando a maior média de internações respiratórias (6,72), além de internações circulatórias também elevadas (9,70). Esse perfil indica um potencial agravamento de doenças cardiovasculares e respiratórias associado ao frio, ar seco e maior transporte de poluentes por ventos mais intensos.

3) Cluster 2 - "Quente, Vento Oeste/Noroeste" (106 dias): Definido por temperaturas elevadas (média de 20,7 C) e radiação solar intensa (1013,2 W/m^2), este cluster reflete dias quentes e relativamente estáveis, com baixo volume de precipitação (2,7 mm). A direção do vento predominante é de Oeste/Noroeste (302°), fator importante para o transporte de poluentes regionais. A umidade é intermediária (79,1 %) e a velocidade do vento é mais baixa (2,88 m/s). O nível de $MP_{2.5}$ permanece intermediário (10,09 µg m^{-3}). Sob essa configuração, observa-se um nível reduzido de internações respiratórias (4,78), o que pode estar relacionado ao maior conforto térmico e menor presença de frentes frias. Entretanto, as internações circulatórias são intermediárias a altas (9,40), sugerindo que o calor e a exposição solar prolongada ainda podem representar um risco cardiovascular moderado.

4) Cluster 3 - "Quente, Seco, Alta Radiação e Poluído" (206 dias): Este é o cluster mais frequente, refletindo o padrão climático dominante ao longo do período analisado. Apresenta as temperaturas médias mais altas (21,1 C), acompanhadas da maior radiação solar (1029,1 W/m^2) e da menor precipitação (1,36 mm). A umidade relativa é a mais baixa entre todos os grupos (73,2 %), o que acentua o caráter seco. A direção do vento predominante é de Leste/Nordeste (85°), e os níveis médios de $MP_{2.5}$ são os mais elevados (11,94 µg m^{-3}), indicando potencial acúmulo de poluentes. Esse cenário combina calor intenso, ar seco e poluição elevada, o que se traduz em níveis intermediários de internações respiratórias (5,13) e altas internações circulatórias (9,43). Tal perfil é compatível com o impacto fisiológico do estresse térmico combinado à má qualidade do ar, especialmente entre populações vulneráveis.

H. Interpretação Técnica da Variância Explicada

A decomposição dos dados ambientais em componentes principais revelou que os dois primeiros componentes (PCA1 e PCA2) explicam 53,10 % da variância total dos dados. Esse valor, embora razoável, também evidencia que quase metade da estrutura informacional ainda está dispersa em dimensões superiores.

De forma geral, o espaço projetado em 2D não é capaz de capturar completamente a complexidade multivariada do sistema. Isso é esperado em contextos com alta correlação entre variáveis ambientais (como temperatura, umidade e pressão), que exigem mais dimensões para uma representação fidedigna.

Dessa forma, a análise visual do gráfico PCA (Figura 6) permite verificar a coerência dos agrupamentos obtidos. Quando os clusters aparecem bem separados no gráfico bidimensional, isso sugere que há diferenciação substancial nas direções de maior variância — o que reforça a validade dos agrupamentos, alinhando-se com as métricas supracitadas.

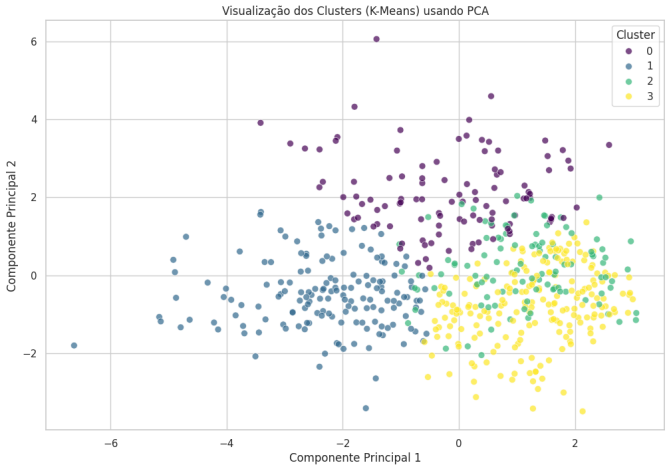

Fig. 6: Visualização dos Clusters (K-Means) usando PCA.

V. CONCLUSÃO

Este estudo demonstrou a aplicabilidade e a eficácia de técnicas de clusterização para a identificação de regimes ambientais distintos em dados multivariados que integram poluição atmosférica, variáveis meteorológicas e desfechos em saúde. A abordagem exploratória, aplicada ao município de Ponta Grossa (PR) entre 2016 e 2018, evidenciou a complexidade das relações entre clima, qualidade do ar e internações hospitalares, reforçando a inadequação de métodos puramente lineares para representar tais dinâmicas.

Dentre os algoritmos avaliados, o K-Means apresentou o melhor desempenho com base em métricas de validação interna (Silhouette = 0,192; Davies-Bouldin = 1,635; Calinski-Harabasz = 125,8), revelando quatro clusters ambientalmente coesos e estatisticamente robustos. A caracterização dos grupos identificou padrões climáticos específicos, como o cluster “chuvoso e úmido” associado a menores níveis de internações, e o cluster "frio e seco com vento forte”, que apresentou as maiores médias de internações respiratórias. Essa associação sugere que certas configurações ambientais — especialmente frio intenso, ar seco e baixa dispersão de poluentes — estão mais fortemente ligadas a riscos à saúde.

A análise exploratória também destacou a presença de outliers e assimetrias marcantes em variáveis como precipitação, $MP_{2,5}$ e internações, que, longe de serem removidos, foram interpretados como eventos críticos e incorporados na modelagem como parte essencial da variabilidade do sistema. O uso da redução de dimensionalidade via PCA permitiu a visualização dos agrupamentos e confirmou visualmente a separabilidade relativa entre os clusters, embora os dois primeiros componentes explicassem apenas 53,1 % da variância total — um indicativo de que o sistema analisado possui uma estrutura informacional de alta complexidade.

A integração de métodos baseados em centroides, hierarquia, densidade e grafos demonstrou que, embora algoritmos como DBSCAN ou GMM tenham utilidade teórica em outros contextos, sua aplicação prática neste caso foi limitada por características específicas d os d ados, c omo b aixa densidade local e ausência de estrutura elipsoidal.

Portanto, este trabalho contribui metodologicamente ao demonstrar que a clusterização é uma ferramenta poderosa para explorar padrões latentes em sistemas ambientais complexos e, ao mesmo tempo, fornece subsídios aplicáveis para gestão hospitalar e vigilância em saúde ambiental. Estudos futuros poderão aprofundar essa abordagem com a integração de variáveis socioeconômicas, séries temporais multivariadas e modelos preditivos híbridos, ampliando a capacidade de antecipação e resposta a cenários críticos de risco à saúde coletiva.

AGRADECIMENTOS

Agradecimento à Fundação Araucária pela concessão de bolsa de mestrado via NAPI Emergência Climática e pela bolsa produtividade à Profa. *Yara de Souza Tadano*.

REFERENCES

[1] World Health Organization (WHO), *WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Geneva: WHO, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240034228

[2] M. G. Arévalo León et al., "Air pollution and its impact on human health: A systematic review," *Contribution to Environmental and Health Sciences*, vol. 5, no. 2, pp. 85–102, 2023.

[3] L. G. Ardiles et al., "Negative binomial regression model to analyze the relationship between hospitalization and air pollution," *Atmospheric Pollution Research*, vol. 9, no. 2, pp. 333-341, 2018.

[4] Y. S. Tadano et al., "Impacto da Poluição Atmosférica e das Alterações Climáticas na Saúde Populacional utilizando Redes Neurais Artificiais," *Revista da Associação Portuguesa de Análise Experimental de Tensões*, vol. 29, pp. 35-42, 2017.

[5] S. Chander et al., "Unsupervised learning methods for data clustering," in *Data Clustering: Algorithms and Applications*, C. C. Aggarwal and C. K. Reddy, Eds. Chapman and Hall/CRC, 2021, pp. 41-64.

[6] R. Bonner, "On Some Clustering Techniques," *IBM J. Res. Dev.*, vol. 8, no. 1, pp. 22-32, Jan. 1964.

[7] B. Paulose, S. Sabitha, R. Punhani, and I. Sahani, "Identification of Regions and Probable Health Risks Due to Air Pollution Using K-Mean Clustering Techniques," in *Proc. Int. Conf. Computational Intelligence and Communication Technology (CICT)*, 2018, pp. 1-5.

[8] K. Grace et al., "Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data," in *Proc. IEEE Region 10 Conf. (TENCON)*, 2016, pp. 2793-2797.

[9] A. R. L. Godoy et al., "Application of machine learning algorithms to PM2.5 concentration analysis in the state of São Paulo, Brazil," *Revista Brasileira de Ciências Ambientais*, vol. 56, pp. 152–165, Mar. 2021.

[10] X. Lu et al., "Estimating hourly PM2.5 concentrations using Himawari-8 AOD and a DBSCAN-modified deep learning model over the YRDUA, China," *Atmospheric Pollution Research*, vol. 12, no. 1, pp. 183–192, Jan. 2021.

[11] H. Dai et al., "Prediction of air pollutant concentration based on one-dimensional multi-scale CNN-LSTM considering spatial-temporal characteristics: A case study of Xi'an, China," *Atmosphere*, vol. 12, no. 12, p. 1626, Dec. 2021.

[12] R. B. Nishida, "Análise do Material Particulado Emitido na Cidade de Ponta Grossa," B.S. thesis, Universidade Tecnológica Federal do Paraná, Ponta Grossa, Brazil, 2017.

[13] Brasil, Ministério da Saúde, Departamento de Informática do SUS (DataSUS). [Online]. Available: http://www.datasus.gov.br

[14] H. Liu et al., "Transforming Complex Problems Into K-Means Solutions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9149-9168, Jul. 2023.

[15] E. Schubert et al., "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, article 19, Jul. 2017.

[16] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *Proc. IEEE 7th Int. Conf. Data Science and Advanced Analytics (DSAA)*, 2020, pp. 747-748.

[17] G. Polezer et al., "The new WHO air quality guidelines for PM2.5: predicament for small/medium cities" in *Environmental Geochemistry and Health*, vol. 45, no. 5, pp. 1841-1860, 2023.