Análise da Qualidade do Ar no Brasil: O Papel do MonitorAr na Saúde Pública

Noam Willyan C[a,*], Thiago Balby C[b], Marcos Antonio A[c]

[a] *Universidade Federal do Maranhão, São Luis, Brazil*
[b] *Universidade Federal do Maranhão, São Luis, Brazil*
[c] *Universidade Federal do Maranhão, São Luis, Brazil*

Abstract

Polluted air affects health and nature conservation, accelerating respiratory diseases and acid rain, causing damage to soil, water and buildings. The Ministry of the Environment has been investing in air quality analysis, including the establishment of a national air surveillance structure and the use of tracking applications. This study aims to use clustering and data assembly methods to examine pollution patterns and their correlations with environmental factors in different regions of Brazil. The analysis will contribute to a deeper understanding of the distribution of pollutants and support more effective strategies for air quality management.

O ar contaminado afeta a saúde e a preservação da natureza, acelerando doenças respiratórias e chuva ácida, causando danos ao solo, água e edifícios. O Ministério do Meio Ambiente vem investindo para análise da qualidade do ar, abrangendo o estabelecimento de uma estrutura nacional de vigilância aérea e empregando aplicações de rastreamento. Este estudo visa empregar métodos de clusterização e montagem de dados para examinar os padrões de poluição e suas correlações com fatores ambientais em diferentes regiões do Brasil. A análise contribuirá para uma compreensão mais aprofundada da distribuição dos poluentes e subsidiará estratégias mais eficazes para a gestão da qualidade do ar.

Keywords: Air quality, Clustering, Environmental monitoring, Pollution patterns

1. Introdução

A poluição do ar é um desafio complexo, resultado da combinação de emissões de veículos, atividades industriais e até mesmo fenômenos naturais [8],[2],[5],[23][31]. A energia, por sua vez, é o motor que move o mundo, sustentando economias e moldando o estilo de vida moderno. No entanto, a maneira como produzimos e consumimos essa energia tem um impacto direto e significativo no planeta. Atualmente, o setor energético é responsável por 75% das emissões de gases de efeito estufa [20], o que o coloca no centro das discussões sobre mudanças climáticas e poluição. Esse cenário nos desafia a buscar soluções mais sustentáveis e inovadoras para um futuro equilibrado.

Além disso, pesquisas recentes mostram que a carga de doenças relacionadas à poluição do ar está mudando, destacando a forte influência da qualidade do ar na saúde da população ao longo do tempo [57][56]. De acordo com o estudo Global Burden of Disease, a poluição do ar externo, incluindo partículas em suspensão, foi responsável por aproximadamente 6,67 milhões de mortes prematuras apenas em 2019 [62]. Esse número alar-

*Corresponding author
Email addresses: noam.costa@discente.ufma.br (Noam Willyan C), tiagobalbyferreiracosta@gmail.com (Thiago Balby C), marcos.alencar@discente.ufma.br (Marcos Antonio A)

mante revela o impacto direto e devastador que a qualidade do ar tem sobre a saúde humana, consolidando a poluição como um dos maiores desafios globais para o bem-estar e a longevidade.

Outro ponto crítico são as partículas inaláveis, classificadas de acordo com seu diâmetro aerodinâmico em $PM_{2.5}$ (partículas com diâmetro $\leq 2.5\,\mu m$) e PM_{10} (partículas com diâmetro $\leq 10\,\mu m$). Essas partículas são especialmente preocupantes devido à sua capacidade de penetrar profundamente no sistema respiratório e de se dispersar por vastas áreas geográficas [41], representando um risco invisível, mas constante, para a saúde pública.

As $PM_{2.5}$, devido ao seu tamanho extremamente reduzido, têm a capacidade de penetrar profundamente nos alvéolos pulmonares e, em alguns casos, até mesmo alcançar a corrente sanguínea. Essa invasão silenciosa pode desencadear processos inflamatórios e estresse oxidativo no organismo [40]. Estudos epidemiológicos reforçam que a exposição crônica a essas partículas está diretamente ligada ao aumento de doenças cardiovasculares e respiratórias [4][64][48][27][22], além de agravar condições preexistentes, como asma, bronquite e rinite alérgica [60].

Diante desse cenário preocupante, o Brasil tem investido em soluções para monitorar e combater a poluição do ar. Diversas estações de monitoramento foram instaladas em diferentes regiões do país, acompanhando a qualidade do ar e rastreando um total de 26 poluentes. Entre os principais estão o Acetaldeído, o Formaldeído, as Partículas Totais em Suspensão (PTS), as Partículas Inaláveis (MP_{10}) e as Partículas Inaláveis Finas ($MP_{2,5}$)[36]. Essas iniciativas são essenciais para entender e mitigar os impactos da poluição na saúde pública.

As partículas finas $PM_{2.5}$ e PM_{10} estão entre os poluentes mais perigosos para a saúde humana. Isso porque, devido ao seu tamanho reduzido, elas podem penetrar profundamente no sistema respiratório e até mesmo na corrente sanguínea, desencadeando problemas graves, como doenças respiratórias e cardiovasculares [66]. Por esse motivo, o monitoramento dessas partículas é uma prioridade, especialmente em grandes centros urbanos, como São Paulo, ou cidades como São Luís.

A qualidade do ar nesses locais tem sido uma preocupação crescente, principalmente após o aumento significativo dos níveis de poluição observado em 2021 [36]. Para ilustrar essa situação, o Gráfico abaixo (Figura 1) mostra a evolução das concentrações médias anuais de PM_{10} na Região Metropolitana de São Paulo (RMSP), além do número de dias meteorologicamente desfavoráveis à dispersão dos poluentes, no período de 2000 a 2022. Já a Figura 2 apresenta a evolução das concentrações médias anuais de $PM_{2.5}$ no mesmo período, reforçando a necessidade de ações contínuas para melhorar a qualidade do ar.

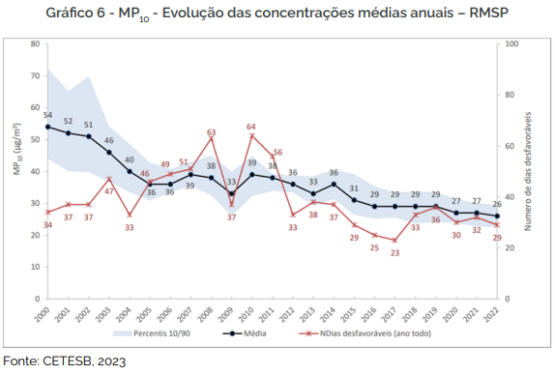

Figure 1: Diagrama de evolução das concentrações médias anuais PM_{10} (RMSP)[36]

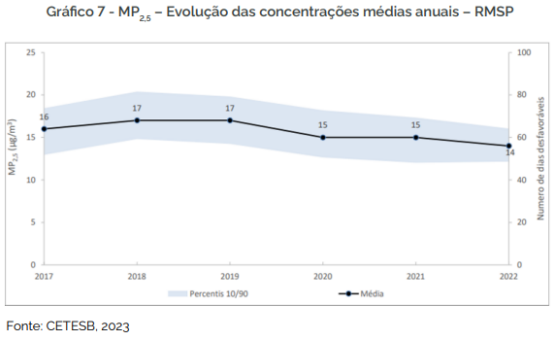

Figure 2: Diagrama de evolução das concentrações médias anuais $PM_{2.5}$ (RMSP)[36]

De acordo com a CETESB (2023), no estado de São Paulo, destacam-se algumas áreas críticas em termos de poluição do ar[36], com o crescimento acelerado da urbanização e das atividades industriais, a poluição do ar virou um problema sério e global. Para enfrentar esses desafios exige tecnologia e inovação. É aí que entra o Ar

Puro - MonitorAr, uma solução que combina Inteligência Artificial Sustentável (IAS), tecnologias avançadas e cluster de dados para oferecer informações em tempo real. Tudo isso disponível, por meio de aplicativos e plataformas online, facilitando o acesso à informação e ajudando a conscientizar as pessoas sobre a importância de respirar um ar de qualidade [60],[36].

O aprendizado de máquina tem se mostrado uma ferramenta poderosa para explorar padrões de poluição do ar em diversas regiões metropolitanas [3][45]e associado a resultados de saúde como aterosclerose e mortalidade [26][24].Austin et al[3]. aplicaram agrupamento k-means em seis anos de dados AP para a cidade de Boston, Massachusetts, A análise revelou a existência de cinco padrões distintos de concentrações diárias de poluentes, demonstrando como a técnica de clusterização pode desvendar padrões complexos e repetitivos que, de outra forma, passariam despercebidos. E não é por acaso: os clusters são famosos por impulsionar a inovação. Essa proximidade favorece a troca de conhecimento, a colaboração e o surgimento de novas ideias, gerando um ambiente dinâmico e competitivo[43],[9],[37]. Mas, claro, nem tudo são flores. Algumas pesquisas levantam dúvidas e apresentam evidências que questionam essa relação tão positiva entre clusters e inovação [51],[50],[11],[39],[46][29]. Por exemplo, estudos indicam que a capacidade de inovação dos clusters pode variar dependendo do grau de desenvolvimento, integração da cadeia produtiva e interação entre os agentes locais.[52]

Este artigo tem como objetivo, analisar os dados coletados pelo MonitorAr, um sistema de monitoramento da qualidade do ar, usando técnicas de clusterização para identificar padrões e agrupamentos que possam nos ajudar a entender melhor os níveis de poluição. A ideia é aplicar métodos tecnológicos avançados para aprimorar a análise desses dados, abrindo caminho para estratégias mais eficientes e inteligentes no controle da qualidade do ar. Em outras palavras, queremos usar a tecnologia a favor de um ar mais limpo e saudável para todos.

2. Aplicações da Análise de Clusters no Estudo da Poluição do Ar

Os clusters têm um papel interessante quando o assunto é inovação: eles podem influenciar de maneiras diferentes antes e depois de um novo paradigma tecnológico surgir. E por que isso acontece? Simplesmente porque o ambiente em que esses clusters estão inseridos faz toda a diferença, moldando como as mudanças se desenrolam [14]. Isso ajuda a identificar padrões – tanto no espaço quanto no tempo – na distribuição de poluentes, o que é essencial para entender como a poluição se comporta. Para isso, algoritmos como K-means, DBSCAN e Hierarchical Clustering são os mais usados para esse tipo de análise [17][13]. As técnicas de clustering são verdadeiras aliadas quando o assunto é reconhecer padrões, entender distribuições e descobrir tecnicas valiosas sobre como os dados se organizam [16][15]. Inclusive, o uso dessas técnicas – especialmente o K-means e a aglomeração hierárquica – na análise de dados sobre poluição do ar não é novidade. Desde a década de 1980, elas vêm sendo exploradas e, ao longo dos anos, só ganharam mais destaque e interesse.

A influência dos clusters na inovação fica clara quando a gente olha como esses agrupamentos facilitam a troca de conhecimento e a colaboração entre empresas e instituições. Segundo [42], os clusters impulsionam a competitividade porque tornam mais fácil o acesso a informações especializadas, tecnologias de ponta e profissionais qualificados. Essa dinâmica é superimportante para a inovação, já que permite que as empresas dividam tanto os riscos quanto os custos ligados ao desenvolvimento de novas tecnologias [35]. Além disso, a proximidade geográfica dentro de um cluster ajuda a criar redes informais de conhecimento, que são um prato cheio para o surgimento de ideias inovadoras [33].

Agora, se a gente for pensar em como aplicar técnicas de clustering na análise de dados ambientais – como no monitoramento da poluição do ar –, fica evidente o quanto identificar padrões pode ser um divisor de águas na hora de tomar decisões [7]. Entre os algoritmos mais usados para isso, o K-means se destaca por organizar os dados em

grupos de forma eficiente.

Quando o assunto é poluição do ar, encontrar padrões pode ser a chave para decisões mais inteligentes. É aí que o clustering entra em cena, ajudando a agrupar informações de um jeito que faz sentido [63]. Um dos algoritmos mais populares é o K-means, que basicamente tenta criar grupos de dados minimizando as diferenças dentro de cada um deles, garantindo que tudo fique bem organizado [30]. Abaixo, a fórmula mostra como o K-means faz esse trabalho [18]:

$$Minimizar : J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \qquad (1)$$

Onde:

- K: Número de clusters.
- C_i: Conjunto de pontos no cluster i.
- μ_i: Centróide do cluster i.
- x: Ponto de dados

Essa fórmula do K-means visa reduzir a soma das distâncias quadradas entre cada ponto x e o centróide μ_i do cluster ao qual ele pertence, buscando uma variação minima dos clusters.[44] Após compreender a função objetivo do K-means e seu papel na minimização da soma das distâncias quadradas dentro dos clusters, surge uma questão fundamental: como determinar o número ideal de clusters para um conjunto de dados?

Uma das abordagens mais utilizadas para essa escolha é o **Método do Cotovelo (Elbow Method)**. Esse método analisa a variação da inércia, que representa a soma das distâncias quadradas entre os pontos e seus centróides, em função do número de clusters, abaixo a fórmula de como funciona o método do Cotovelo[25][44]:

$$\text{Inércia} : J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \qquad (2)$$

A inércia é uma métrica usada para avaliar a qualidade da clusterização em algoritmos como o K-Means, medindo a compactação dos Clusters,

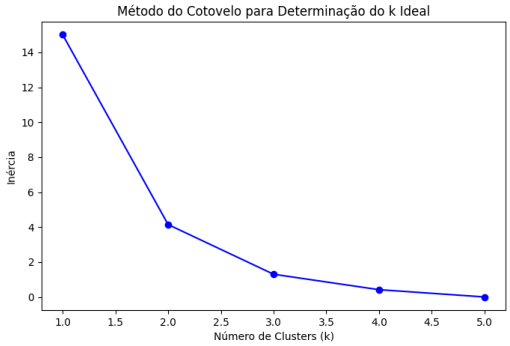

Figure 3: Diagrama ilustrativo do Método do Cotovelo

abaixo podemos analisar um gráfico (figure 3) de como funciona o Método do Cotovelo no Cluster:

Percebe-se portanto que a medida que k aumenta: A inércia cai rapidamente, pois os pontos são agrupados em clusters menores e mais compactos, logo K = 2 e k = 3 são bons candidatos, pois não estão em valores de inercia tão alto e também, não tão baixos.

Estudos recentes têm apostado em algoritmos de clustering para identificar de onde vêm as emissões de poluentes e avaliar se as políticas públicas de controle da qualidade do ar estão funcionando [59]. Essas análises são úteis porque ajudam a apontar áreas críticas que precisam de atenção especial, além de mostrar se as medidas regulatórias estão surtindo efeito ao longo do tempo [58]. Para entender na prática como um cluster funciona, este trabalho vai analisar dados reais do Ministério do Meio Ambiente e Mudança do Clima. Mas, para facilitar a visualização, a (Figura 4) vai usar dados fictícios sobre a concentração de partículas finas em algumas regiões:

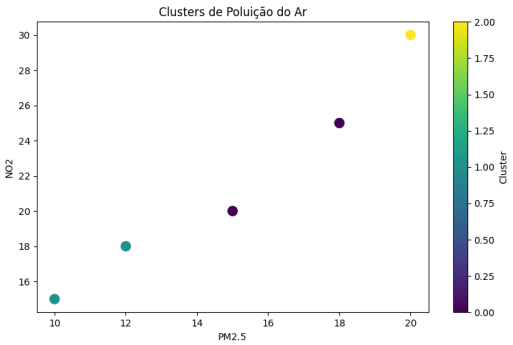

Figure 4: Diagrama ilustrativo do Método do Cotovelo

Cidades	PM2.5	NO2	SO2	Cluster
São Paulo	20	30	10	C1
Rio de Janeiro	18	25	8	C2
Belo Horizonte	15	20	7	C3
Curitiba	12	18	5	C2
Salvador	10	15	4	C3

Table 1: Agrupamento de cidades baseado nos níveis de poluentes atmosféricos.

Os dados apresentados são meramente ilustrativos, mas demonstram que os níveis de poluição em São Paulo são mais elevados em comparação com Salvador. No agrupamento por clusters, São Paulo pertence ao Cluster 1 (amarelo), enquanto Rio de Janeiro e Curitiba estão no Cluster 2 (roxo). Já Belo Horizonte e Salvador fazem parte do Cluster 3 (azul). Observa-se que os níveis de poluição são menores no Cluster 3, indicando uma melhor qualidade do ar nessas regiões.

Além disso, a integração de técnicas de clustering com outras abordagens analíticas, como a modelagem preditiva e a análise de séries temporais, tem se mostrado promissora para a compreensão de fenômenos complexos, como a dispersão de poluentes atmosféricos. Por exemplo, [65] combinaram técnicas de clustering com modelos de aprendizado de máquina para prever a concentração de partículas finas (PM2.5) em diferentes regiões urbanas[6]. Essa abordagem permitiu não apenas a identificação de padrões espaciais e temporais, mas também a previsão de eventos de alta poluição com antecedência, facilitando a implementação de medidas preventivas[10].

3. Outros métodos de Clustering

Neste artigo, vamos explorar o algoritmo K-means, seus usos práticos, exemplos de aplicação e a matemática que o sustenta [61]. No entanto, é fundamental também abordar outro tema relevante: o Agrupamento Hierárquico, que consiste na formação de clusters de maneira hierárquica. Essa abordagem é amplamente utilizada na análise de dados para revelar padrões ocultos em conjuntos de informações [21]. Uma característica marcante do Hierarchical Clustering é a criação de uma cadeia de grupos, frequentemente representada por um diagrama em forma de árvore, conhecido como dendrograma [55]. Esse processo pode ser realizado por meio de duas técnicas principais: a junção (que começa de baixo para cima) e a separação (que começa de cima para baixo). No método hierárquico, cada ponto de dados é tratado individualmente no início, e os grupos são formados gradualmente à medida que os pontos mais próximos são unidos, até que todos os dados estejam agrupados em um único cluster [32].

Abaixo (Figura 5), podemos observar um dendrograma que ilustra graficamente como funciona o Agrupamento Hierárquico simples [49]. O dendrograma do eixo X mostra os pontos de dados, e o eixo Y indica a distância entre clusters. Linhas horizontais conectam grupos conforme eles se unem, e quanto mais alta a fusão, mais distintos são os clusters antes da junção.[38] Esse gráfico ajuda a visualizar quantos clusters podem ser formados, bastando cortar em um nível desejado. A explicação da separação entre agrupamentos é um ponto-chave do Agrupamento Hierárquico. Há diversas abordagens para medir essa separação, tais como:

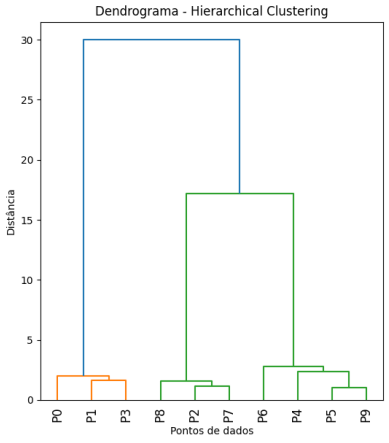

Figure 5: Diagrama para identificar padrões e estruturas subjacentes em conjuntos de dados.

- Single linkage: considera a menor distância entre quaisquer dois pontos pertencentes a clusters diferentes.
- Complete linkage: utiliza a maior distância

entre quaisquer dois pontos de clusters distintos.

- Average linkage: calcula a média das distâncias entre todos os pares de pontos entre clusters (Sibson, 1973).

4. metodologia

4.1. Coleta de dados

Para analisar a qualidade do ar, utilizamos dados coletados por estações de monitoramento espalhadas por todo o Brasil. Essas estações têm a importante tarefa de medir a concentração de vários poluentes atmosféricos, que são essenciais para entender como está a qualidade do ar que respiramos. Neste estudo, os poluentes que estamos considerando incluem:

- $PM_{2.5}$: Partículas inaláveis com diâmetro menor que 2.5 µm
- PM_{20}: Partículas inaláveis com diâmetro menor que 10 µm

Na coleta de dados para a clusterização, serão analisadas as seguintes variáveis: estado, item monitorado (poluente), concentração, IQAR, data e situação. A análise abrangerá todos os estados do nordeste como Bahia, Maranhão, Rio Grande do Norte e Pernambuco: As demais regiões não foram incluídas devido à ausência de estações de monitoramento da rede **MonitorAr**. A (Table2) exibe as estações, cidades, estados, concentrações de PM10 e PM2.5 (Figura 6), além do número do cluster. Abaixo, a (Figura 7) traz a análise 3D.

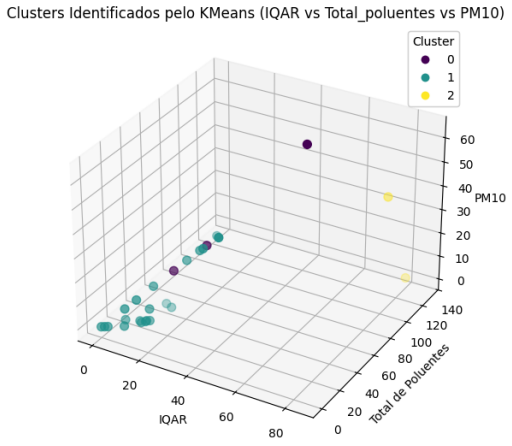

Figure 6: Diagrama 3D para identificar padrões de poluição IQAR e TotalPoluição

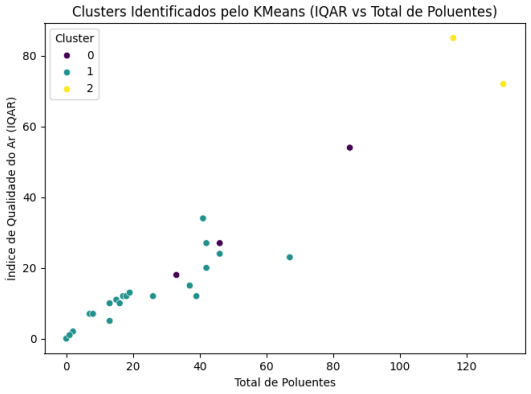

Figure 7: Diagrama 2D para identificar padrões de poluição IQAR e TotalPoluição

Estaçao	**Cidade**	**Estado**	**PM10**	**PM2.5**	**O3**	**NO2**	**SO2**	**CO**	**IQAR**	**Total**	**Cluster**
Capinzal.N	São Luís	MA	33	0	72	0	26	0	72	131	2
Ybacanga	São Luís	MA	18	15	0	0	0	0	18	33	0
PostoSaúde	São Luís	MA	64	21	0	0	0	0	64	85	0
Gapara	São Luís	MA	24	0	7	1	12	2	24	70	1
AERCA	Açailândia	MA	27	19	0	0	0	0	27	46	0
UTE	São Luís	MA	20	0	12	1	8	1	20	42	1
Santo Ant.	Santo Ant.	MA	33	0	72	0	26	0	72	131	2
Areias II	Camaçari	BA	0	0	0	0	7	0	7	7	1
Areias	Camaçari	BA	0	0	0	0	0	0	0	0	1
Machadinho	Camaçari	BA	0	15	13	3	5	1	15	37	1
Caboto	Candeias	BA	23	0	18	4	21	1	23	67	1
Botelho	Salvador	BA	0	0	13	3	2	1	13	19	1
Camâra	D. d'Ávila	BA	0	0	12	3	2	0	12	17	1
Cobre	D. d'Ávila	BA	0	0	0	0	2	0	2	2	1
Leandrinho	D. d'Ávila	BA	12	10	0	2	2	0	12	26	1
Gravatá	Camaçari	BA	10	0	0	3	0	0	10	13	1
Lamarão	S.Sebastião	BA	0	0	12	2	4	0	12	18	1
FuturamaI	D. d'Ávila	BA	0	0	0	0	1	0	1	1	1
Malemba	Camaçari	BA	7	10	13	1	85	0	85	116	2
Gamboa	Candeias	BA	5	0	13	1	0	0	13	19	1
Concórdia	D. d'Ávila	BA	7	0	0	1	0	0	7	8	1
Escola	D. d'Ávila	BA	0	0	11	2	2	0	11	15	1
EDCUPE	Recife	PE	0	0	10	1	3	2	10	16	1
IFPE	Recife	PE	0	0	5	0	8	0	5	13	1
CPRH	SantoAgos	PE	34	0	0	1	4	2	34	41	1
Ipojuca	Ipojuca	PE	27	0	7	3	3	2	27	42	1
PortoAlegre	CETE/RN	RN	0	11	5	4	12	7	12	39	1

Table 2: Agrupamento de cidades baseado nos níveis de poluentes atmosféricos.

Nesta etapa final da análise do MonitorAR, concluímos o processo de clusterização, avaliando os padrões identificados e a distribuição geográfica dos dados coletados. Para facilitar a visualização, utilizamos um mapa do Brasil, que destaca as regiões analisadas e mostra onde as concentrações de informações mais relevantes estão localizadas. Além disso, o mapa ajuda a identificar possíveis relações entre os pontos de monitoramento, tornando os resultados mais fáceis de encontrar. A extração dos dados foi feita com bastante cuidado, empregando técnicas avançadas de processamento e modelagem. Fatores como a densidade de ocorrências, tendências sazonais e possíveis anomalias nos registros foram levados em consideração. A visualização geoespacial no mapa é uma ferramenta importante, pois simplifica a compreensão desses padrões e permite uma interpretação mais clara da abrangência do estudo e do impacto dos resultados.

Com essa abordagem, o MonitorAR se mostra importante na análise ambiental e monitoramente de qualidade do ar. Abaixo o mapa que ilustra como funciona o site do MonitorAR: A análise

Figure 8: Diagrama geografico de monitoramento do Ar

completa do cluster será expandida para trabalhar com bases de dados mais robustas, permitindo uma investigação detalhada da dispersão e concentração de poluentes atmosféricos. O estudo não se limitará apenas às partículas finas da cadeia de PM2.5 e PM10, mas também abrangerá outros contaminantes relevantes, como óxidos de nitrogênio (NO_x), dióxido de enxofre (SO), ozônio (O_3) e compostos orgânicos (COVs), que desempenham um papel significativo na qualidade do ar e na saúde pública.

Com essa abordagem, o estudo vai oferecer informações mais precisas para o monitoramento ambiental, ajudando no desenvolvimento de políticas públicas e estratégias que possam melhorar a qualidade do ar em diversas regiões.

4.2. Contextualização da problematica

Os dados fornecidos pelo MonitorAr ofereciam informações sobre as regiões e os níveis de concentração de cada poluente. Para facilitar a análise, realizamos uma filtragem desses dados, já que a base contava com mais de 2.000 linhas, abrangendo todas as cidades e estados. Durante a análise, observamos que as estações de Malemba, Santo Antônio e Capinzal do Norte apresentam os maiores níveis de poluição, com destaque para Malemba, que registra o pior índice. Seu IQAR chega a 85, um valor preocupante, já que está acima do nível considerado seguro pelo MonitorAr. Essa situação reforça a necessidade de atenção especial a essas áreas, abaixo a tabela mostrando o IQAR e seus níveis: Além do cluster de total de poluentes vs IQAR,

Classificação	**IQAR**
BOA	0 - 40
MODERADA	41 - 80
RUIM	81 - 120
MUITO RUIM	121 - 200
PÉSSIMA	+200

Table 3: Classificação dos níveis de qualidade do ar conforme o IQAR.

criamos outros gráficos de cluster para análise da quatidade de poluente por regiões. segue abaixo um desses gráficos:

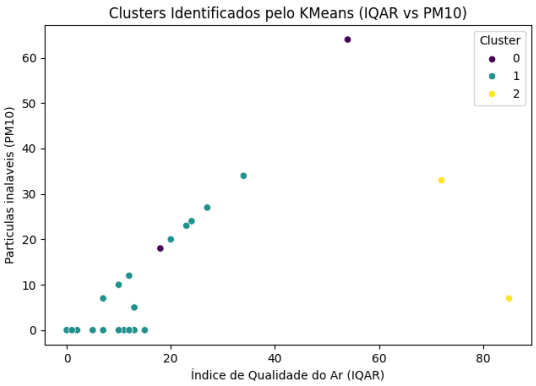

Figure 9: Diagrama 2D para identificar padrões de poluição IQAR vs PM10

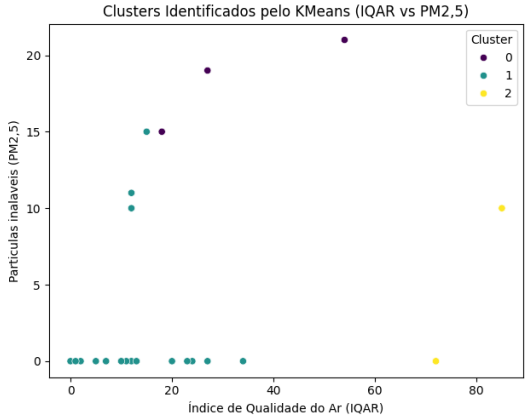

Figure 10: Diagrama 2D para identificar padrões de poluição IQAR vs PM2.5

Na Figura 9, é possível observar que um dos clusters (Cluster 2) se mantém abaixo do limite de PM10, porém apresenta um nível elevado de IQAR. Isso indica que o principal poluente responsável pelo alto índice de qualidade do ar não é o PM10. De acordo com a tabela de referência, esse ponto corresponde à estação Malemba, onde o maior fator de poluição é a presença de dióxido de enxofre (SO_2), enquanto um dos clusters (cluster 0) apresenta níveis elevados de PM10 indicando o Posto de saúde Ybacanga. As análises escolhidas foram referentes aos poluentes PM10 e PM2.5, por serem os mais comuns e considerados os mais perigosos à saúde. No entanto, é importante destacar que todos os poluentes apresentam riscos. Além disso, a medida analisada foi o Índice de Qualidade do Ar (IQAR), que mede a concentração e o impacto desses poluentes no ambiente, abaixo um gráfico de como funciona o cluster, por estação, total de poluentes e IQAR:

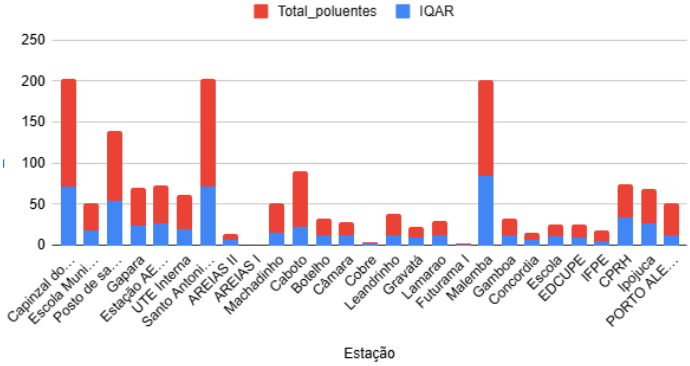

Figure 11: Grafico de colunas IQAR vs TOTAL POLUENTES

As estações com IQAR mais elevados são aquelas classificadas como ruins, pois os níveis de poluentes presentes no ar ultrapassam os limites considerados seguros. Isso indica que a qualidade do ar nessas áreas está comprometida, exigindo medidas para reduzir os impactos negativos, abaixo será apresentado mapas de calor de cada cluster, onde será explicado como eles se comportam na análise de poluentes:

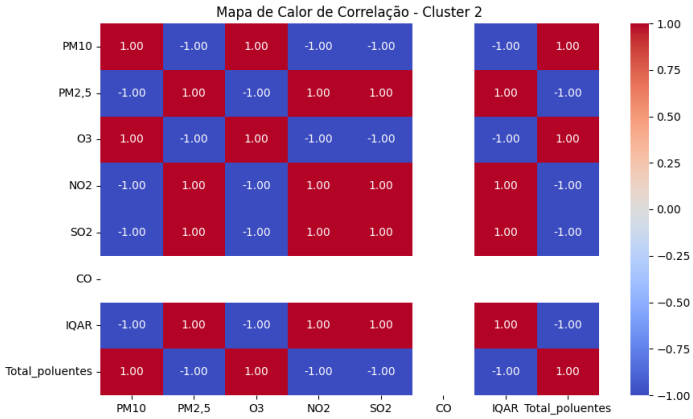

Figure 12: Mapa de Calor: Cluster 2

No mapa de calor do Cluster 2 (Figura 12), mostra que, à medida que os níveis de PM10 aumentam, os de PM2.5 diminuem, o que sugere que não há uma relação direta entre esses dois

poluentes. Por outro lado, o O3 tende a aumentar junto com o PM10, indicando que esses são os poluentes mais relevantes nesse cluster. Já os níveis de NO2 e SO2 permanecem baixos, sem apresentar um aumento significativo em relação ao PM10. Em relação à primeira análise, o total de poluentes nesse cluster é consideravelmente mais alto e altamente influenciado pelo PM10. O mapa de calor pode ser analisado olhando a relação das linhas com os poluentes das colunas No cluster 2, Os materiais poluentes que apresentam maiores indices são o ozônio e enxofre.

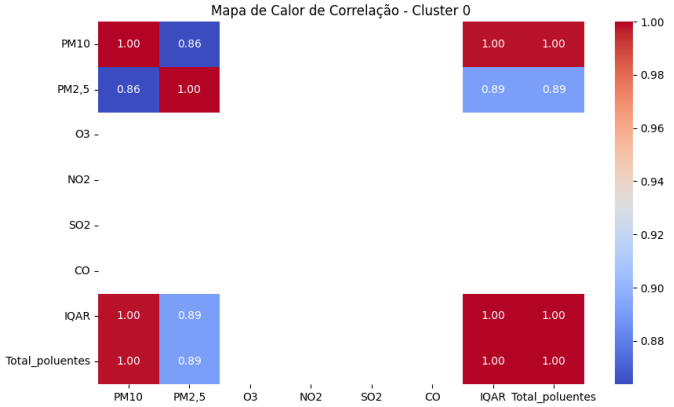

Figure 13: Mapa de Calor: Cluster 0

No Cluster 0, não há presença de [O_3, NO_2, CO]. No entanto, ao analisarmos o PM10, percebemos uma forte correlação com o PM2.5, pois à medida que o PM10 aumenta, o PM2.5 também aumenta, o IQAR e total de poluentes também estão correlacionados, podendo indicar que o PM10 é o maior poluente em relação ao IQAR no cluster 0.

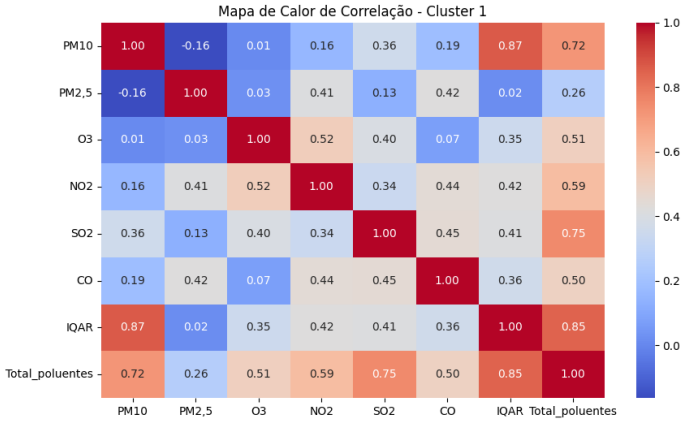

Figure 14: Mapa de Calor: Cluster 1

No cluster 1 os níveis de poluentes são bem baixos em relação ao IQAR do MonitorAR, por isso não iremos nos aprofundar no cluster 1. O cluster 0 embora que também não tem poluentes altos, no posto de saúde do bacanga chega aos aproximidas 54 IQAR, o que é moderado, porém seu total de poluentes não é tão grave em comparação com estações: Malemba, Santo Antonio, Capinzal do Norte.

5. Dados e formulas de outros metodos de Clustering

Neste trabalho, exploramos dois métodos essenciais para análise de agrupamentos: o K-means e o Clustering Hierárquico. No entanto, é importante lembrar que existem vários outros métodos de clustering, cada um com suas características e aplicações específicas. Conhecer essas alternativas é importante para escolher a abordagem mais adequada a cada situação e para realizar análises mais detalhadas e precisas.

Relembrando a função principal do clustering, conforme destacado por [53], trata-se do processo de agrupar um conjunto de objetos de modo que aqueles pertencentes ao mesmo grupo (cluster) sejam mais semelhantes entre si do que com os objetos de outros grupos. Essa definição reforça a importância de selecionar o método de clustering mais apropriado, considerando as características dos dados e os objetivos da análise, eles podem se

classificados como: K-means, Hierarchical Clustering, DBSC [28][13].

5.1. DBSC

O DBSCAN é robusto a outliers e não requer o número de clusters como entrada.[13] ele agrupa pontos que estão próximos em regiões de alta densidade e identifica pontos de ruído, sua formula pode ser idenficada como:

Densidade(x) = Número de pontos dentro do raio ϵ de x

5.2. GMM

O GMM é uma abordagem probabilística que permite clusters com formas elípticas[34], onde cada cluster é modelado como uma distribuição normal multivariada, sua formula pode ser identificada como:

$$P(x) = \sum_{i=1}^{k} \pi_i . N(x|\mu_i, \sum_i) \tag{3}$$

5.3. Avaliação de qualidade de um clustering

As métricas de um cluster são ferramentas essenciais para avaliar a qualidade e a eficácia da segmentação de dados realizada por um algoritmo de agrupamento. Elas ajudam a quantificar o quão bem os dados foram divididos em clusters, fornecendo informações sobre a homogeneidade e a separação dos grupos formados.[12] Abaixo, são apresentados algumas das principais métricas de avaliação de clusters e como o teste de qualidade pode ser realizado:

- **Coeficiente de silhueta:** O coeficiente de silhueta é uma métrica útil para avaliar a separação e a coesão dos clusters[47][12], pode ser medido seguindo a formula abaixo:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

  Ela Mede quão bem um ponto está agrupado em relação a outros clusters. Varia de -1 a 1, onde valores próximos de 1 indicam clusters bem definidos.[47]

- **Índice de Davies-Bouldin:** Mede a razão entre a dispersão intra-cluster e a separação inter-cluster. Valores menores indicam clusters melhores.

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{5}$$

  Onde:

  - k: Número de clusters.
  - σ_i: Dispersão média dos pontos no cluster i.
  - $d(c_i, c_j)$: Distância entre os centroides dos clusters i e j.

6. Resolução de dados complexos com o UMAP

O UMAP (Uniform Manifold Approximation and Projection) é um algoritmo que ajuda a simplificar dados complexos, reduzindo sua dimensionalidade.[54] Ele não é exatamente um método de clustering, ou seja, não agrupa os dados diretamente. No entanto, ele é muito útil como uma etapa preparatória antes de aplicar técnicas de agrupamento. Ao usar o UMAP, os dados ficam mais organizados e separados, o que facilita bastante a identificação de padrões e a formação de clusters mais claros e significativos.[54] É como se ele preparasse o terreno para que outras técnicas possam trabalhar de forma mais eficiente! Nesse trabalho não teremos necessidade de trabalhar com o UMAP, porém vale a pena ressaltar a sua importância e explicar como funciona.

O UMAP foi apresentado em 2018 por (Leland McInnes, John Healy e James Melville) no artigo [19] essa técnica ganhou popularidade para redução de dimensionalidade, competindo com métodos como PCA e t-SNE. Como dito a técnica T-SNE, também reduz a dimensionalidade, porém o UMAP é mais rápido e eficiênte tendo um desempenho maior nas relações estruturais entre os pontos. Enquanto o TSNE utiliza as distâncias euclidianas, o UMAP usa distribuição de probabilidade exponencial de uma forma diferente do TSNE, abaixo a formula de como isso funciona:

$$p_{i|j} = e - \frac{d(x_i, x_j) - p_i}{\sigma_i} \tag{6}$$

UMAP usa o número de vizinhos mais próximos em vez de perplexidade sendo próximo k sem a função log2, ou seja, da seguinte forma[1]:

$$k = 2, \sum_i p_{ij} \tag{7}$$

Outro fato sobre o UMAP é que foi inspirado em fundamentos da teoria de grafos e geometria diferencial, o algoritmo constrói um grafo de proximidade no espaço de alta dimensão e o mapeia para um espaço de menor dimensão de forma a minimizar a distorção da relação entre os pontos.Abaixo tem a formula de família de curva que calculo a probabilidade de distâncias em dimensões baixas[1]:

$$q_{ij} = (1 + a(y_i - y_j)^{2b})^{-1} \tag{8}$$

O UMAP é uma ferramenta poderosa para lidar com dados complexos e de alta dimensionalidade. No entanto, como o sistema apresentado aqui é mais simples, ele será mencionado apenas como referência, permitindo que compreendamos sua finalidade e papel na organização de clusters.

7. Conclusão

Conclui-se que o monitoramento da qualidade do ar é de extrema importância para a compreensão e redução dos impactos ambientais e à saúde pública. Usando a base de dados MonitorAr, usamos técnicas de clustering (K-means) para analisar dados ambientais, e os resultados foram bastante interessantes. Aplicando métodos de agrupamento, conseguimos identificar padrões de diversos estados e regiões do nordeste, que mostram como os poluentes se distribuem. Isso ajudou a entender melhor de onde vem a contaminação e quais fatores influenciam a qualidade do ar em cada região.

Essa abordagem baseada em dados mostrou o quanto ela é importante para a gestão ambiental. Os dados gerados podem ser utilizados para tomar decisões mais embasadas e criar políticas públicas mais eficientes. Outro fato a ser identificado foi a qualidade dos dados e a escolha dos algoritmos de clustering, pois, são pontos que exigem atenção, fazendo toda a diferença no resultado final. Para avançar ainda mais, futuros estudos podem incluir outras variáveis, como outras regiões do Brasil, além de técnicas mais avançadas de machine learning. Isso poderia aumentar a precisão e a utilidade do monitoramento. Por fim, o MonitorAr com clustering se mostrou uma ferramenta promissora para enfrentar os desafios da poluição do ar. Com ela, podemos contribuir não só para a saúde pública, mas também para um futuro mais sustentável.

References

[1] Como funciona exatamente umap, 2020. Como funciona exatamente UMAP,24 de dezembro de 2020.

[2] Al-Hasnawi, S., H. H. A.-A. N., and nutsson, S. The effect of the industrial activities on air pollution at baiji and its surrounding areas. *Iraq. Engineering 8* (2016), 34–44.

[3] Austin, E., Coull, B., Thomas, D., and Koutrakis, P. A framework for identifying distinct multipollutant profiles in air pollution data. *Environment International 45* (2012), 112–121.

[4] Brook RD, Rajagopalan S, P. C. r. B. J. B. A. D.-R. A. H. F. H. Y. L. R. M. M. P. A. S. D. S. S. J. W. L. K. J. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association. *American Heart Association Council on Epidemiology and Prevention* (2010).

[5] Can, A. Time series analysis of air pollutants for karabük province. *ITM Web Conf. 9* (2017), 02002.

[6] Choi, D., Ham, J., Heo, G., Lee, S.-H., Yoo, J.-W., Yang, G.-H., Jeon, S., and Kim, C.-H. Mesoscale wind field patterns conducive to the high-pm2.5 episodes over south korea: Cluster analysis. *Atmospheric Environment 333* (2024), 120653.

[7] Chunhui Li, Lian Sun, J. J. Y. C. X. W. Risk assessment of water pollution sources based on an integrated k-means clustering and set pair analysis method in the region of shiyan, china. *Science of The Total Environment 557–558* (2016), 307–316.

[8] de Lima Brum, R., Penteado, J. O., Ramires, P. F., Tavella, R. A., Honscha, L. C., da Silva Freitas, L., de Moura, F. R., da Silva Bonifácio, A., da Silva, V. M., dos Santos da Silva, L., Santos, J. E. K., and da Silva Júnior, F. M. R. Southern air project - scientific efforts to monitor and measure the impacts of air pollution in southern brazil. *Societal Impacts 4* (2024), 100074.

[9] Delgado, M. The co-location of innovation and production in clusters. *Industry and Innovation 27*, 8 (2020), 842–870.

[10] Dirk Fornahl, R. H., and Menzel, M.-P. Broadening our knowledge on cluster evolution. *European Planning Studies 23*, 10 (2015), 1921–1931.

[11] Elisa Giuliani, Pierre-Alexandre Balland, A. M. Straining but not thriving: understanding network dynamics in underperforming industrial clusters. *Journal of Economic Geography 19* (2019), 147–172.

[12] .Eréndira Rendón, Itzel Abundez, A. A., and Quiroz., E. M. Internal versus external cluster validation indexes. *INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS 5* (2011).

[13] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), KDD'96, AAAI Press, p. 226–231.

[14] Gilbert, B. A., and Campbell, J. T. The geographic origins of radical technological paradigms: A configurational study. *Research Policy 44*, 2 (2015), 311–327.

[15] Govender, P., and Sivakumar, V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research 11*, 1 (2020), 40–56.

[16] Halkidi, M., B. Y., and Vazirgiannis, M. On clustering validation techniques. *Journal of Intelligent Information Systems 17* (2001), 107–145.

[17] Han, J., K. M. . P. J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann,2012, 2012.

[18] Hartigan, J. A., and Wong, M. A. Algorithm AS 136: A K-Means clustering algorithm. *Applied Statistics 28*, 1 (1979), 100–108.

[19] Healy, John, McInnes, Leland. Uniform manifold approximation and projection. *Nature Reviews Methods Primers 4* (2024). Pagination: 27.

[20] IEA. International energy agency, "iea". Acessado em: 15 fev. 2025.

[21] Jain, A.K., M. M., and Flynn, P. Data clustering: A review. *ACM Computing Surveys 31* (1999), 264–323.

[22] James Bergstra JAMES, Y. B. Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13* (2012), 281–305.

[23] Jos Lelieveld, Andrea Pozzer, U. P. M. F. A. H. T. M. Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. *Cardiovascular Research 116* (2020), 1910–1917.

[24] Keller, H. Universality claim of attachment theory: Children's socioemotional development across cultures. *Proceedings of the National Academy of Sciences 115*, 45 (2018), 11414–11419.

[25] Ketchen, D. J., . S. C. L. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal 17* (1996), 441–458.

[26] Kioumourtzoglou MA, Schwartz J, J. P. D. F. Z. A. Pm2.5 and mortality in 207 us cities: Modification by temperature and city characteristics. *Epidemiology 27* (2016), 221–7.

[27] Laumbach, R., and Kipen, H. Respiratory health effects of air pollution: Update on biomass smoke and traffic pollution. *The Journal of allergy and clinical immunology 129* (01 2012), 3–11; quiz 12.

[28] Leonard Kaufman, P. J. R. Finding groups in data: An introduction to cluster analysis. *John Wiley Sons* (1990).

[29] Li, P. Cluster-based routines and paradigm-bound innovation. *Research Policy 54*, 3 (2025), 105192.

[30] MacQueen, J. Some methods for classification and analysis of multivariate observations. *Statistics, University of California Press 1* (1967), 281–297.

[31] Mallik, C. Anthropogenic sources of air pollution. *CABI* (2019), 6–25.

[32] Manning, C.D., S. H., and Raghavan, P. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge., 2008.

[33] Maskell, P. *Towards a knowledge-based theory of the geographical cluster*, vol. 10. 2001.

[34] McLachlan, G., . P. D. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[35] Michaud, P. Clustering techniques. *Future Generation Computer Systems 13*, 2 (1997), 135–147. Data Mining.

[36] Ministério do Meio Ambiente e Mudança do Clima, B. *RELATÓRIO ANUAL DE ACOMPANHAMENTO DA QUALIDADE DO AR 2023*. 2023.

[37] Moretti, E. The effect of high-tech clusters on the productivity of top inventors. *American Economic Review 111*, 10 (October 2021), 3328–75.

[38] Murtagh, F., . L. P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification 31* (2014), 274–295.

[39] Noni, I. D., and Belussi, F. Breakthrough invention performance of multispecialized clustered regions in europe. *Economic Geography 97*, 2 (2021), 164–186.

[40] Pope, C., and Dockery, D. Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air Waste Management Association 56* (2006), 709–742.

[41] Pope, C. A., Coleman, N., Pond, Z. A., and Burnett, R. T. Fine particulate air pollution and human mortality: 25+ years of cohort studies. *Environmental Research 183* (2020), 108924.

[42] Porter, M. E. Clusters and the new economics of competition. *Harvard Business Review 76* (1998), 77–90.

[43] Porter, M. E. Location, competition, and economic development: Local clusters in a global economy. *Economic Development Quarterly 14*, 1 (2000), 15–34.

[44] Qu, F., Shi, Y., Yang, Y., Hu, Y., and Liu, Y. Coordinate descent k-means algorithm based on split-merge. *Computers, Materials and Continua 81*, 3 (2024), 4875–4893.

[45] Riches, N. O., Gouripeddi, R., Payan-Medina, A., and Facelli, J. C. K-means cluster analysis of cooperative effects of co, no2, o3, pm2.5, pm10, and so2 on incidence of type 2 diabetes mellitus in the us. *Environmental Research 212* (2022), 113259.

[46] Ron Boschma, Ernest Miguelez, R. M., and Ocampo-Corrales, D. B. The role of relatedness and unrelatedness for the geography of technological breakthroughs in europe. *Economic Geography 99*, 2 (2023), 117–139.

[47] Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20* (1987), 53–65.

[48] S. Harinath, U. M. Effect of air pollution on human health in industrial areas - a case study. *Journal of Industrial Pollution Control 28* (2012), 9–11.

[49] (ScikitLearnPlot)Hierarchical Clustering Dendrogram,2025.

[50] Sergio Mariotti, Rocco Mosconi, L. P. Location and survival of mnes' subsidiaries: Agglomeration and heterogeneity of firms. *Strategic Managent 40* (2019), 2242–2270.

[51] Shaver, J.M, F. F. Agglomeration economies, firm heterogeneity. *foreign direct investment in the United States 21* (2000).

[52] SUZIGAN, W., FURTADO, J., GARCIA, R., and SAMPAIO, S. Clusters or local production systems: Mapping, classification and suggestions for polices. *Brazilian Journal of Political Economy*, 4 (2004), 548–570.

[53] Tan, P.-N., S. M. . K. V. *Introduction to Data Mining*. Pearson Education Limited, 2019.

[54] Telkmann, K., Gudi-Mindermann, H., Bogers, R., Ahrens, J., Tönnies, J., van Kamp, I., Vrijkotte, T., and Bolte, G. Identification of exposome clusters based on societal, social, built and natural environment – results of the abcd cohort study. *Environment International 197* (2025), 109335.

[55] Trevor Hastie, Jerome H. Friedman, R. T. *The Elements of Statistical Learning*. Data Mining, Inference, and Prediction, Second Edition, 2009.

[56] Unep2021. *Actions on Air Quality: A Global Summary of Policies and Programmes to Reduce Air Pollution*. 07 September 2021, 2021.

[57] Wan Hu, Lanlan Fang, H. Z. R. N. . G. P. Changing trends in the air pollution–related disease burden from 1990 to 2019 and its predicted level in 25 years. *Environmental Science and Pollution Research 30* (2023), 1761–1773.

[58] Wang, S., and Hao, J. Air quality management in china: Issues, challenges, and options. *Journal of Environmental Sciences 24*, 1 (2012), 2–13.

[59] Wei, Y., Jing, X., Chen, Y., Sun, W., Zhang, Y., and Zhu, R. Spatial–temporal characteristics, source apportionment, and health risks of atmospheric volatile organic compounds in china: A comprehensive review. *Toxics 12*, 11 (2024).

[60] (WHO), W. H. O. *WHO global air quality guidelines: Particulate matter (PM, and PM), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. 22 September 2021, 2021.

[61] Wu, R. Behavioral analysis of electricity consumption characteristics for customer groups using the k-means algorithm. *Systems and Soft Computing 6* (2024), 200143.

[62] Yu, W., Ye, T., Zhang, Y., Xu, R., Lei, Y., Chen, Z., Yang, Z., Zhang, Y., Song, J., Yue, X., Li, S., and Guo, Y. Global estimates of daily ambient fine particulate matter concentrations and unequal spatiotemporal distribution of population exposure: a machine learning modelling study. *The Lancet Planetary Health 7*, 3 (2023), e209–e218.

[63] Zhang, L., and Yang, G. Cluster analysis of pm2.5 pollution in china using the frequent itemset clustering approach. *Environmental Research 204* (2022), 112009.

[64] Zhang, Q., Smith, G., and Wu, Y. Catalytic hydrolysis of sodium borohydride in an integrated reactor for hydrogen generation. *International Journal of Hydrogen Energy 32*, 18 (2007), 4731–4735.

[65] Zhou, S., Wang, W., Zhu, L., Qiao, Q., and Kang, Y. Deep-learning architecture for pm2.5 concentration prediction: A review. *Environmental Science and Ecotechnology 21* (2024), 100400.

[66] Ziqiang Zhang, Ning Song, J. W. J. L. L. S. J. D. Effect of pm2.5 air pollution on the global burden of neonatal diarrhea from 1990 to 2019. *Environmental Pollution 367* (2025), 125–604.

8. Apêndice

Fórmulas Adicionais[34]

- **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - Y_i| \tag{9}$$

- **Mean Absolute Percentage Error (MAPE):**

$$MAPE = \frac{1oo\%}{n}\sum_{i=1}^{n}\frac{|y_i - Y_i|}{y_i} \tag{10}$$

- **Coefficient of Determination (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

- **Symmetric Mean Absolute Percentage Error (SMAPE):**

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{2|F_i - A_i|}{|F_i| + |A_i|} \tag{12}$$