

APLICAÇÃO DE ALGORITMOS DE AGRUPAMENTO E ASSOCIAÇÃO NA IDENTIFICAÇÃO DE PADRÕES DE POLUIÇÃO ATMOSFÉRICA E HÍDRICA EM ESCALA GLOBAL

Application Of Clustering And Association Algorithms In The Identification Of Air And Water
Pollution Patterns On A Global Scale

Gabriel Vinicios Nanneti

Nathan Scremin

RESUMO

A acelerada urbanização global intensificou os desafios de gestão ambiental, especificamente no que tange à qualidade do ar e à poluição dos recursos hídricos. Frequentemente analisados de forma isolada, a correlação entre estes dois vetores de degradação carece de validação estatística em escala global. Este trabalho propõe a aplicação de técnicas de Mineração de Dados para identificar e segmentar perfis de poluição em uma base de dados composta por 3.963 cidades. A metodologia adota uma abordagem descritiva, utilizando o algoritmo não-supervisionado K-Means para o agrupamento de cidades, com o número de clusters validado pelo Método do Cotovelo, e o algoritmo Apriori para a mineração de regras de associação e descoberta de padrões de co-ocorrência. O estudo busca superar a análise unidimensional tradicional, identificando perfis ambientais heterogêneos e cenários intermediários. Os resultados visam contribuir para o cumprimento dos Objetivos de Desenvolvimento Sustentável (ODS) 6 e 11 da ONU, fornecendo subsídios para que gestores públicos desenvolvam políticas de intervenção direcionadas e específicas para cada perfil de risco urbano identificado.

Palavras-chave: Mineração de Dados. K-Means. Poluição Urbana. ODS. Regras de Associação.

ABSTRACT

Accelerated global urbanization has intensified environmental management challenges, specifically regarding air quality and water resources pollution. Often analyzed in isolation, the correlation between these two degradation vectors lacks statistical validation on a global scale. This work proposes the application of Data Mining techniques to identify and segment pollution profiles across a dataset of 3,963 cities. The methodology adopts a descriptive approach, employing the unsupervised K-Means algorithm for city clustering, with the number of clusters validated by the Elbow Method, and the Apriori algorithm for mining association rules and discovering co-occurrence patterns. The study aims to overcome traditional one-dimensional analysis by identifying heterogeneous environmental profiles and intermediate scenarios. The results intend to contribute to the fulfillment of

the UN Sustainable Development Goals (SDGs) 6 and 11, providing a basis for public managers to develop targeted intervention policies specific to each identified urban risk profile.

Keywords: Data Mining. K-Means. Urban Pollution. SDGs. Association Rules.

INTRODUÇÃO

A urbanização acelerada é um dos fenômenos sociodemográficos mais marcantes do século XXI. À medida que as populações globais migram massivamente para os centros urbanos em busca de oportunidades econômicas, a pressão sobre a infraestrutura e os recursos naturais atinge níveis críticos. Neste cenário, a degradação ambiental manifesta-se predominantemente através de dois vetores visíveis e nocivos à saúde pública: a deterioração da qualidade do ar e a poluição dos recursos hídricos. A gestão simultânea destes problemas representa um dos maiores desafios para governos e organizações internacionais na atualidade.

Embora existam vastos repositórios de dados sobre poluição, a análise destes indicadores frequentemente ocorre em "silos", ou seja, de forma isolada. Agências de monitoramento atmosférico e departamentos de saneamento raramente cruzam suas informações para obter uma visão holística da saúde ambiental de uma cidade. Essa fragmentação dificulta a compreensão da complexidade urbana real.

Desafios

Diante deste contexto, surge uma lacuna no entendimento das dinâmicas de poluição global. O senso comum sugere que cidades altamente poluídas sofrem de uma degradação generalizada, afetando tanto o ar quanto a água. No entanto, esta premissa carece de validação estatística global.

Portanto, este trabalho busca responder à seguinte questão norteadora: A poluição do ar e a poluição da água estão necessariamente correlacionadas em ambientes urbanos, ou é possível identificar, através de dados, perfis de cidades com desempenhos ambientais mistos e heterogêneos?

Objetivos

O objetivo geral deste estudo é aplicar técnicas de Mineração de Dados (*Data Mining*), especificamente algoritmos de agrupamento, para identificar, segmentar e analisar padrões globais de poluição urbana, superando a análise unidimensional tradicional.

Para alcançar este propósito, definem-se os seguintes objetivos específicos:

- Implementar o algoritmo de aprendizado de máquina não-supervisionado K-Means para processar uma base de dados contendo indicadores de 3.963 cidades.
- Determinar e validar estatisticamente o número ideal de perfis ambientais (clusters) utilizando o Método do Cotovelo (Elbow Method), visando encontrar agrupamentos que representem fielmente a realidade, incluindo cenários intermediários.

- Analisar comparativamente os perfis encontrados, com ênfase na identificação de discrepâncias entre qualidade do ar e da água.
- Aplicar técnicas de Mineração de Regras de Associação (Algoritmo Apriori) para descobrir padrões de co-ocorrência não triviais e relações probabilísticas entre os níveis de poluição atmosférica e hídrica.
- Contextualizar a posição das cidades brasileiras no cenário global, identificando quais os desafios predominantes no território nacional.

Justificativas

A relevância deste trabalho transcende o exercício acadêmico de análise de dados, alinhando-se diretamente à Agenda 2030 da Organização das Nações Unidas (ONU). A identificação precisa de perfis ambientais é fundamental para o cumprimento de dois Objetivos de Desenvolvimento Sustentável: a ODS 6 (Água Potável e Saneamento) e a ODS 11 (Cidades e Comunidades Sustentáveis).

Ao demonstrar que diferentes cidades possuem diferentes "assinaturas" de poluição, este estudo fornece subsídios para que gestores públicos abandonem soluções genéricas. A segmentação permite a alocação eficiente de recursos: cidades com "Crise Dupla" exigem intervenção emergencial integrada, enquanto cidades com perfis mistos (ex: ar limpo, mas água poluída) necessitam de políticas focadas especificamente em saneamento básico. Assim, a aplicação das técnicas de Mineração de Dados torna-se uma ferramenta estratégica para a promoção da saúde pública e da justiça ambiental.

O restante deste artigo está organizado da seguinte forma: A Seção 2 apresenta a fundamentação teórica. A Seção 3 detalha a metodologia proposta, incluindo o pré-processamento e a configuração dos algoritmos. A Seção 4 discute os resultados obtidos e a análise dos clusters e regras. Por fim, a Seção 5 apresenta as conclusões e trabalhos futuros.

FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos basilares que sustentam a pesquisa, divididos em três eixos: a caracterização da poluição urbana, os fundamentos da Mineração de Dados e o detalhamento matemático do algoritmo K-Means.

Poluição Urbana

A poluição ambiental em centros urbanos é uma consequência direta da industrialização e do crescimento populacional desordenado. Segundo relatórios de agências internacionais como a JICA (*Japan International Cooperation Agency*), a urbanização rápida frequentemente excede a capacidade de infraestrutura das cidades, resultando em saneamento inadequado e emissões atmosféricas descontroladas.

A poluição atmosférica refere-se à contaminação do ambiente interno ou externo por qualquer agente químico, físico ou biológico. Os principais poluentes monitorados incluem o material particulado (PM_{2.5} e PM₁₀), dióxido de nitrogênio (NO₂), dióxido de enxofre (SO₂) e ozônio troposférico (O₃). A exposição crônica a estes agentes está associada a doenças respiratórias, cardiovasculares e câncer de pulmão, representando, segundo o NCBI (*National Center for Biotechnology Information*), uma das maiores cargas de doenças em regiões em desenvolvimento.

Paralelamente, a poluição hídrica é caracterizada pela introdução de substâncias em corpos d'água que alteram sua qualidade, tornando-a prejudicial à saúde humana e aos ecossistemas. Em ambientes urbanos, as fontes primárias incluem o descarte de efluentes industriais não tratados e o esgoto doméstico. Indicadores comuns de qualidade da água incluem a Demanda Bioquímica de Oxigênio (DBO), turbidez, pH e a presença de coliformes fecais.

Embora frequentemente geridos por agências distintas, a literatura recente sugere que ar e água são componentes inseparáveis da saúde ambiental urbana, exigindo abordagens analíticas integradas para a formulação de políticas públicas eficazes (ODS 6 e ODS 11).

Mineração de Dados

A Mineração de Dados (*Data Mining*) é uma etapa essencial do processo de Descoberta de Conhecimento em Bancos de Dados ou KDD (*Knowledge Discovery in Databases*). Ela pode ser definida como o processo de exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados com o objetivo de descobrir padrões significativos e regras desconhecidas (*hidden patterns*).

No contexto ambiental, a mineração de dados permite transformar vastos conjuntos de medições brutas de sensores em conhecimento acionável. As tarefas de mineração são geralmente divididas em duas categorias:

- Preditivas: Utilizam dados passados para prever valores futuros (ex: Regressão).
- Descritivas: Buscam encontrar padrões interpretáveis que descrevem os dados (ex: Agrupamento ou *Clustering*).

Este trabalho foca na abordagem descritiva através da análise de agrupamento, que visa organizar objetos em grupos (*clusters*) de tal forma que objetos no mesmo grupo sejam mais semelhantes entre si do que com objetos de outros grupos.

Algoritmo K-Means

O K-Means é um dos algoritmos de aprendizado de máquina não-supervisionado mais utilizados na literatura para problemas de agrupamento, devido à sua eficiência computacional e facilidade de interpretação. Conforme revisado por Sivakumar (2020) em aplicações atmosféricas e aplicado por Hadi et al. (2025) em recursos hídricos, o algoritmo particiona o conjunto de dados em k grupos pré-definidos.

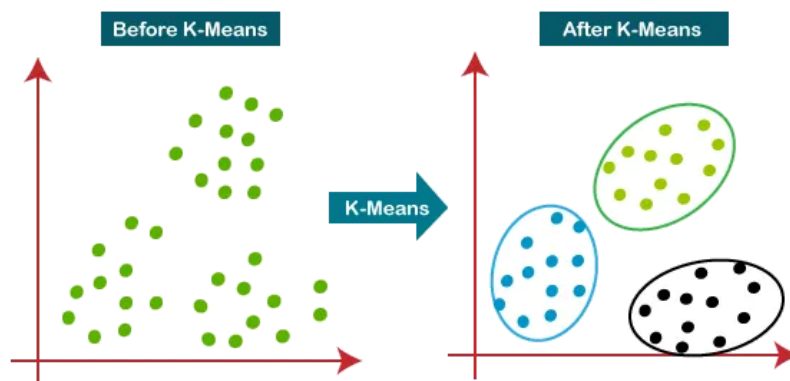


Imagem 1 - K-Means Clustering Algorithm

O objetivo do K-Means é minimizar a "inércia" ou a soma dos erros quadráticos ou SSE (*Sum of Squared Errors*) dentro de cada cluster. O algoritmo opera iterativamente através dos seguintes passos:

1. Inicialização: Seleciona-se aleatoriamente k pontos como centróides iniciais (centros dos grupos).
2. Atribuição: Cada ponto de dado é atribuído ao centróide mais próximo. A proximidade é geralmente calculada pela Distância Euclidiana. Dado um ponto x e um centróide c , a distância é dada por:

$$d(x, c) = \sqrt{\sum_{i=1}^n (x_i - c_i)^2}$$

3. Atualização: Novos centróides são calculados tirando a média aritmética de todos os pontos atribuídos a cada cluster.
4. Repetição: Os passos 2 e 3 se repetem até que os centróides não mudem mais (convergência).

O K-Means é um algoritmo baseado em distância. Portanto, ele é altamente sensível à escala das variáveis. Se uma variável (ex: Poluição da Água) varia de 0 a 1000 e outra (ex: Qualidade do Ar) varia de 0 a 10, a primeira dominará o cálculo da distância, *enviesando* o resultado.

Para corrigir isso, utiliza-se a técnica de padronização Z-Score (*Standard Scaler*), que reescala os dados para que tenham média (μ) igual a 0 e desvio padrão (σ) igual a 1. A fórmula aplicada a cada valor x é:

$$z = \frac{x - \mu}{\sigma}$$

Isso garante que todas as variáveis ambientais contribuam equitativamente para a formação dos clusters.

Método do Cotovelo

Uma limitação do K-Means é a necessidade de definir o número de clusters (k) a priori. Para determinar o k ideal, utiliza-se o Método do Cotovelo.

Este método consiste em rodar o algoritmo para diferentes valores de k (ex: de 1 a 10) e plotar a inércia (soma das distâncias quadráticas intra-cluster) em um gráfico. À medida que k aumenta, a inércia diminui. O ponto ideal é o "cotovelo" da curva, onde o ganho de informação ao adicionar um novo cluster deixa de ser significativo, indicando o equilíbrio entre compressão dos dados e precisão da segmentação.



Imagem 2 - Método Cotovelo

Regras de Associação

A Mineração de Regras de Associação é uma técnica de aprendizado de máquina não-supervisionado utilizada para descobrir relações interessantes, correlações ou estruturas causais entre conjuntos de itens em grandes bancos de dados. Diferente do agrupamento (que organiza dados pela similaridade), a associação busca identificar padrões de co-ocorrência frequente.

Historicamente introduzida por Agrawal et al. (1993) no contexto da "análise de cesta de compras" (*Market Basket Analysis*), a técnica visa encontrar regras do tipo "Se (Antecedente), Então (Consequente)", representadas formalmente por $A \rightarrow B$, onde A e B são conjuntos de itens disjuntos. No contexto ambiental, isso permite inferir probabilidades condicionais, como a probabilidade de a qualidade do ar ser baixa dado que a poluição da água é alta.

O algoritmo mais clássico e amplamente utilizado para esta tarefa é o Apriori. Ele opera sob o princípio da "propriedade Apriori", que afirma que qualquer subconjunto de um conjunto de itens frequente deve ser também frequente.

O funcionamento do algoritmo ocorre em duas etapas principais:

1. Geração de Conjuntos de Itens Frequentes: O algoritmo identifica todos os itens (ou conjuntos de itens) que aparecem na base de dados com uma frequência superior a um limite mínimo pré-definido (suporte mínimo).

2. Geração de Regras: A partir destes conjuntos frequentes, o algoritmo constrói as regras de associação que satisfazem uma confiança mínima.

Para determinar a relevância e a força de uma regra de associação, utilizam-se métricas estatísticas fundamentais. As três principais são:

Support: O suporte indica a frequência com que a regra aparece na base de dados total. É a probabilidade da interseção entre o antecedente (A) e o conseqüente (B).

$$Suporte(A \rightarrow B) = P(A \cup B) = \frac{n^{\circ} \text{ de transações contendo } A \text{ e } B}{n^{\circ} \text{ total de transações}}$$

Confidence: A confiança mede a confiabilidade da inferência. Ela expressa a probabilidade condicional de o conseqüente (B) ocorrer, dado que o antecedente (A) já ocorreu. Uma confiança de 1.0 (ou 100%) indica uma relação lógica absoluta nos dados observados.

$$Confiança(A \rightarrow B) = P(B|A) = \frac{Suporte(A \cup B)}{Suporte(A)}$$

Lift: O Lift avalia a força da regra comparando a confiança observada com a confiança esperada se A e B fossem estatisticamente independentes.

- Lift = 1: Indica independência (a regra é mera coincidência).
- Lift > 1: Indica uma associação positiva (a presença de A aumenta a probabilidade de B).

$$Lift(A \rightarrow B) = \frac{Confiança(A \rightarrow B)}{Suporte(B)}$$

METODOLOGIA

Este capítulo descreve os procedimentos computacionais, as técnicas de tratamento de dados e os critérios estatísticos adotados para a segmentação das cidades e a descoberta de padrões de associação. O fluxo de trabalho foi estruturado sequencialmente em: coleta, pré-processamento, modelagem de agrupamento e mineração de regras.

Base de Dados

O estudo utilizou a base de dados pública denominada *cities_air_quality_water_pollution.18-10-2021.csv*. O conjunto de dados original contém 3.963 registros referentes a cidades de diversas regiões globais. É importante afirmar que a base de dados continha uma coluna *Region*, que foi desconsiderada durante a limpeza dos dados.

Nome da Variável	Tipo de Dado	Descrição	Unidade/Escala	Fonte Original
City	Categórico (Nominal)	Nome da cidade onde a medição foi feita.	Texto	Kaggle
AirQuality	Numérico (Contínuo)	Índice composto da qualidade do ar.	0 - 100	-
WaterPollution	Numérico (Contínuo)	Índice composto da poluição hídrica.	0 - 100	-

Tabela 1 – Categorias Principais da Tabela

Para a análise, foram selecionadas as duas variáveis numéricas centrais, ambas representadas em uma escala indexada de 0 a 100:

- *AirQuality*: Índice de qualidade do ar (quanto maior o valor, melhor a qualidade do ar).
- *WaterPollution*: Índice de poluição da água (quanto maior o valor, mais poluída é a água).

Ferramentas Utilizadas

O processamento e análise dos dados foram realizados utilizando a linguagem de programação Python, executada em ambiente de nuvem (Google Colab). O ecossistema de bibliotecas utilizado incluiu:

- Pandas: Para manipulação, limpeza e estruturação dos dados tabulares.

- Scikit-Learn (*sklearn*): Para a aplicação dos algoritmos de agrupamento (K-Means) e padronização dos dados (*StandardScaler*).
- Mlxtend: Para a aplicação do algoritmo Apriori (geração de regras de associação).
- Seaborn e Matplotlib: Para a visualização gráfica e análise estatística dos resultados.

Pré-processamento dos Dados

Para garantir a integridade da análise, os dados brutos foram submetidos a três etapas de tratamento:

- Limpeza e Agregação: Inicialmente, realizou-se a higienização dos nomes das colunas e o tratamento de duplicatas. Para cidades com múltiplas medições, optou-se pela utilização da Mediana para agregar os valores, visando mitigar o impacto de outliers extremos e obter um valor representativo por localidade.
- Aplicação da Padronização: Considerando que o algoritmo K-Means é sensível à escala das variáveis, aplicou-se a técnica de Padronização Z-Score (*StandardScaler*). Esta técnica transforma os dados para que tenham média (μ) igual a 0 e desvio padrão (σ) igual a 1, garantindo que as variáveis de Ar e Água contribuam equitativamente para o cálculo das distâncias.
- Agrupamento de Dados: O algoritmo escolhido para a segmentação das cidades foi o K-Means, configurado com a inicialização *k-means++* para otimizar a convergência.

Para definir o número ideal de grupos (k), utilizou-se o Método do Cotovelo (*Elbow Method*). A análise qualitativa indicou que $k = 3$ seria a escolha mais robusta, pois permitiu capturar não apenas os extremos (bom/ruim), mas também um perfil intermediário de cidades com discrepância entre ar e água.

Regras de Associação

Como etapa complementar ao agrupamento, utilizou-se a técnica de Regras de Associação para descobrir relações probabilísticas entre as variáveis (ex: "Se a poluição da água é alta, qual a probabilidade do ar ser ruim?").

Visto que algoritmos de associação requerem dados categóricos, as variáveis numéricas contínuas foram discretizadas em 5 categorias ordinais (*bins* de amplitude fixa): Muito Baixo (0-20), Baixo (20-40), Médio (40-60), Alto (60-80) e Muito Alto (80-100). Essa abordagem foi escolhida em detrimento da binarização simples para permitir a detecção de nuances e estágios intermediários de poluição, enriquecendo as regras geradas.

Então aplicou-se o algoritmo Apriori (via biblioteca *mlxtend*) para identificar conjuntos de itens frequentes e gerar regras condicionais. A relevância das regras foi avaliada através das métricas de Suporte (frequência) e Confiança (probabilidade condicional).

RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos a partir da execução do fluxo de trabalho computacional (*pipeline*) descrito na metodologia. A discussão está estruturada de forma a acompanhar as etapas de transformação dos dados, desde a caracterização inicial até a identificação dos perfis ambientais e regras de associação.

Caracterização Inicial da Base de Dados

A etapa inicial consistiu na carga e higienização estrutural do conjunto de dados. Após a correção sintática dos cabeçalhos (remoção de caracteres especiais e espaçamentos), obteve-se uma matriz de dados contendo identificadores geográficos (*City*, *Region*, *Country*) e os dois indicadores numéricos centrais (*AirQuality* e *WaterPollution*).

A Tabela 2 apresenta uma amostra dos cinco primeiros registros da base bruta, ilustrando a estrutura das variáveis que serão submetidas ao processamento.

City	Region	Country	AirQuality	WaterPollution
New York City	New York	United States of America	46.81	49.50
Washington, D.C.	District of Columbia	United States of America	66.12	49.10
San Francisco	California	United States of America	60.51	43.00
Berlin		Germany	62.36	28.61
Los Angeles	California	United States of America	36.62	61.29

Tabela 2 – Amostra inicial dos Dados

Pré-processamento e Padronização dos Dados

Para a etapa de modelagem, foram selecionadas exclusivamente as variáveis numéricas AirQuality e WaterPollution. Como o algoritmo K-Means é sensível à magnitude dos vetores de dados, a utilização dos valores brutos poderia enviesar o agrupamento.

Para mitigar este problema, aplicou-se a técnica de Padronização Z-Score (*StandardScaler*). Esta transformação reescalou os dados para que apresentassem média igual a 0 e desvio padrão igual a 1. A Tabela 3 demonstra o efeito desta transformação nas primeiras amostras do conjunto de dados, onde se observa que valores próximos à média global tornam-se próximos de zero, enquanto valores extremos (*outliers*) tornam-se positivos ou negativos de maior magnitude.

Cidade (Amostra)	Ar Original	Água Original	Ar Padronizado (Z)	Água Padronizada (Z)
New York City	46.81	49.50	-0.49	0.18
Washington, D.C.	66.12	49.10	0.12	0.17
San Francisco	60.51	43.00	-0.05	-0.06
Berlin	62.36	28.61	0.01	-0.62
Los Angeles	36.62	61.29	-0.82	0.65

Tabela 3 – Amostra Padronizada dos Dados

A aplicação da padronização permitiu uma comparação direta entre as métricas, independentemente de suas escalas originais. Por exemplo, cidades como Los Angeles apresentam um índice padronizado de qualidade do ar negativo (-0.82), indicando uma performance inferior à média global, simultaneamente a um índice de poluição da água positivo (0.65), indicando níveis de poluição

acima da média. Esta transformação é fundamental para que o algoritmo K-Means interprete corretamente a "distância" de cada cidade em relação aos centros dos grupos.

Determinação do Número de Clusters

Para definir a quantidade ideal de perfis ambientais (k), o algoritmo K-Means foi executado iterativamente para valores de k variando de 1 a 10. A métrica de avaliação utilizada foi a Inércia (Soma dos Erros Quadráticos - SSE), que quantifica a compactação dos clusters.

A Figura 3 apresenta a curva de decaimento da inércia obtida pelo Método do Cotovelo. Observa-se que a inércia decresce rapidamente até $k = 2$ e $k = 3$, estabilizando-se progressivamente a partir deste ponto.

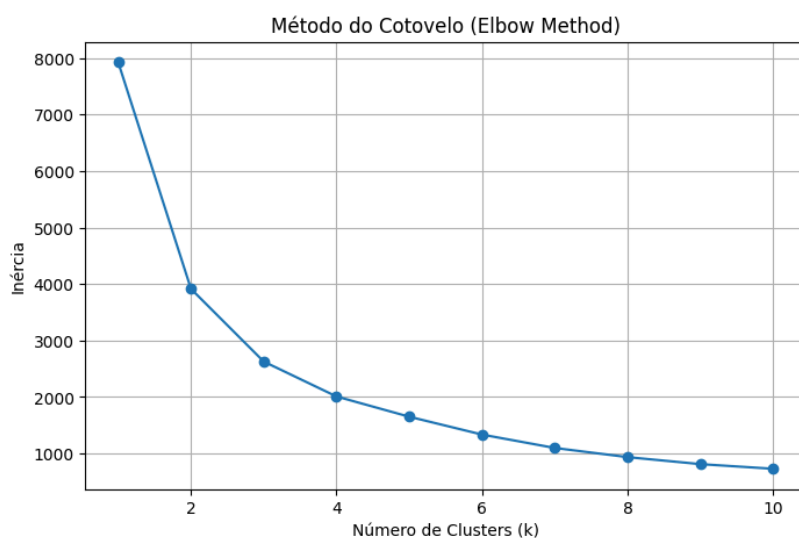


Imagem 3 - Método Cotovelo Aplicado

Embora a inflexão matemática mais acentuada ocorra em $k = 2$ (sugerindo uma divisão binária entre "bom" e "ruim"), optou-se pela definição de $k = 3$. A análise qualitativa demonstrou que a inclusão de um terceiro cluster permitiu isolar um grupo de cidades com características heterogêneas (discrepância entre ar e água), enriquecendo a interpretação fenomenológica dos dados sem incorrer em *overfitting* (superajuste), o que ocorreria com valores de k superiores a 4.

Segmentação e Definição dos Perfis Ambientais

Com o parâmetro $k = 3$ definido na etapa anterior, o algoritmo K-Means foi executado sobre a base de dados completa e padronizada. A segmentação resultou na classificação de cada uma das 3.963 cidades em um grupo específico, baseado na similaridade vetorial de seus indicadores.

A Figura 4 ilustra a distribuição espacial das cidades após o agrupamento. O eixo horizontal representa a Qualidade do Ar e o eixo vertical a Poluição da Água. A distinção cromática evidencia a

formação de três regiões bem delimitadas, validando a capacidade do modelo em separar comportamentos distintos e não-lineares.

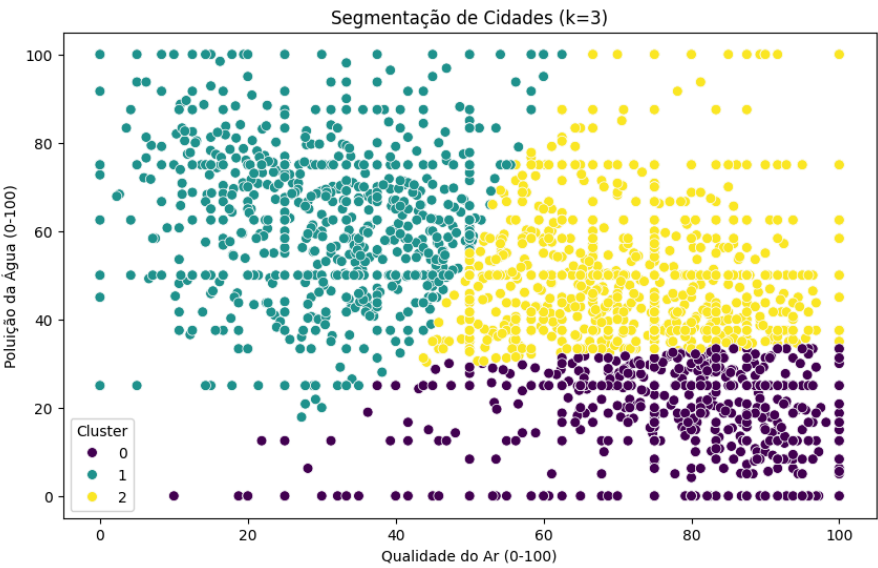


Imagem 4 - Agrupamento em $k = 3$

Para interpretar o significado fenomenológico de cada grupo, analisaram-se os centróides (valores médios) de cada cluster. A Tabela 4 apresenta os perfis médios calculados, revertidos para a escala original de 0 a 100 para fins de interpretação.

Cluster ID	Qualidade do Ar (Méd)	Poluição da Água (Méd)	Definição do Perfil
0	82.66 (Alta)	13.48 (Baixa)	Sustentabilidade Alta
1	24.06 (Baixa)	63.41 (Alta)	Crise Ambiental Dupla
2	77.49 (Alta)	52.74 (Média/Alta)	Desafio Hídrico

Tabela 4 – Tabela com os valores e suas respectivas definições

Análise dos Perfis Identificados

A análise quantitativa dos centróides permite a seguinte caracterização dos perfis urbanos:

Cluster 0 - Sustentabilidade Alta (Ar Bom / Água Boa): Representa o cenário ideal (*Benchmark*), onde altos índices de qualidade do ar coexistem com baixíssima poluição hídrica. Cidades neste grupo atingiram, simultaneamente, metas cruciais das ODS 6 e 11.

Cluster 1 - Crise Ambiental Dupla (Ar Ruim / Água Ruim): O grupo mais crítico. Cidades aqui localizadas sofrem de degradação sistêmica, falhando em ambas as métricas. A correlação positiva entre ar ruim e água ruim neste grupo sugere problemas estruturais graves de gestão urbana e industrialização.

Cluster 2 - Desafio Hídrico (Ar Bom / Água Poluída): Este perfil é o achado mais significativo do estudo, pois refuta a hipótese de correlação linear absoluta. Ele agrupa cidades que conseguem manter uma boa qualidade atmosférica (média 77.49), mas falham drasticamente no controle da poluição da água (média 52.74). Este grupo evidencia que o "Ar Limpo" não é garantia de "Água Limpa", exigindo políticas públicas focadas especificamente em saneamento básico.

Distribuição Quantitativa Global

Após a classificação das 3.963 cidades, realizou-se a contagem absoluta e relativa dos registros em cada cluster para compreender a prevalência de cada perfil no cenário mundial.

Perfil (Cluster)	Quantidade de Cidades	Percentual (%)
0. Sustentabilidade	1.156	29,2%
1. Crise Dupla	1.242	31,3%
2. Desafio Hídrico	1.565	39,5%

Tabela 5 – Valores Globais

Os dados revelam que o perfil predominante globalmente é o Cluster 2 (Desafio Hídrico), abrangendo quase 40% das cidades analisadas. Este dado é revelador: indica que uma parcela significativa dos centros urbanos globais obteve êxito na mitigação da poluição atmosférica, mas ainda falha na gestão dos recursos hídricos. O perfil de Sustentabilidade (Cluster 0), embora representativo

(29,2%), permanece como o grupo minoritário, evidenciando o longo caminho a ser percorrido para o cumprimento integral das ODS.

O Cenário Brasileiro

A análise focada no território brasileiro revelou uma heterogeneidade ambiental maior do que a observada na amostragem inicial. Ao total, foram identificadas cidades em todos os três espectros de classificação, com uma predominância clara do desafio hídrico.

Cidade	Perfil Identificado	Ar (0-100)	Água (0-100)	Interpretação
São Paulo	Crise Dupla	24,1	73,7	Índices críticos refletem a alta densidade industrial e gargalos no saneamento da bacia do Tietê.
Rio de Janeiro	Crise Dupla	46,8	77,2	A topografia e o trânsito impactam o ar, enquanto a poluição da Baía de Guanabara eleva o índice hídrico.
Manaus	Crise Dupla	31,2	46,4	A presença do Polo Industrial impacta a qualidade do ar, inserindo a cidade amazônica no grupo de crise.
São Luís	Crise Dupla	50,0	100,0	Alerta Crítico: Apresentou o índice máximo de poluição da água na amostra, indicando situação de emergência sanitária.

Brasília	Desafio Hídrico	76,5	38,0	O planejamento urbano favorece a dispersão de poluentes (Ar bom), mas a gestão hídrica ainda requer atenção.
Salvador	Desafio Hídrico	83,0	51,3	A forte ventilação costeira garante excelente qualidade do ar, mascarando os problemas de balneabilidade e esgoto.
Curitiba	Desafio Hídrico	73,1	48,9	Referência em planejamento, mantém o ar sob controle, mas enfrenta desafios intermediários na qualidade da água.

Tabela 6 – Cidades Brasileiras e suas Situações.

A estatística descritiva do Brasil aponta para uma realidade complexa:

- Desafio Hídrico (54 cidades): É a realidade da maioria das cidades brasileiras analisadas. O país tende a ter uma qualidade do ar aceitável (muitas vezes por fatores geográficos favoráveis), mas sofre sistemicamente com a poluição da água.
- Crise Dupla (29 cidades): Concentra as grandes metrópoles e pólos industriais, onde a capacidade de suporte do meio ambiente foi excedida em ambas as frentes.
- Sustentabilidade (20 cidades): Diferente da hipótese inicial, o Brasil possui representantes neste grupo, indicando a existência de "ilhas de excelência" ou cidades de menor porte com menor pressão antrópica.

Esta distribuição reforça a urgência de priorizar a ODS 6 (Água e Saneamento) nas políticas públicas nacionais, visto que é o vetor de poluição mais abrangente no território, afetando até mesmo cidades com boa qualidade atmosférica.

Análise da Distribuição Espacial

Para visualizar a fronteira de decisão entre os perfis identificados, plotou-se a distribuição de todas as cidades no plano bidimensional, conforme a Figura 3. A coloração representa os clusters atribuídos pelo algoritmo.

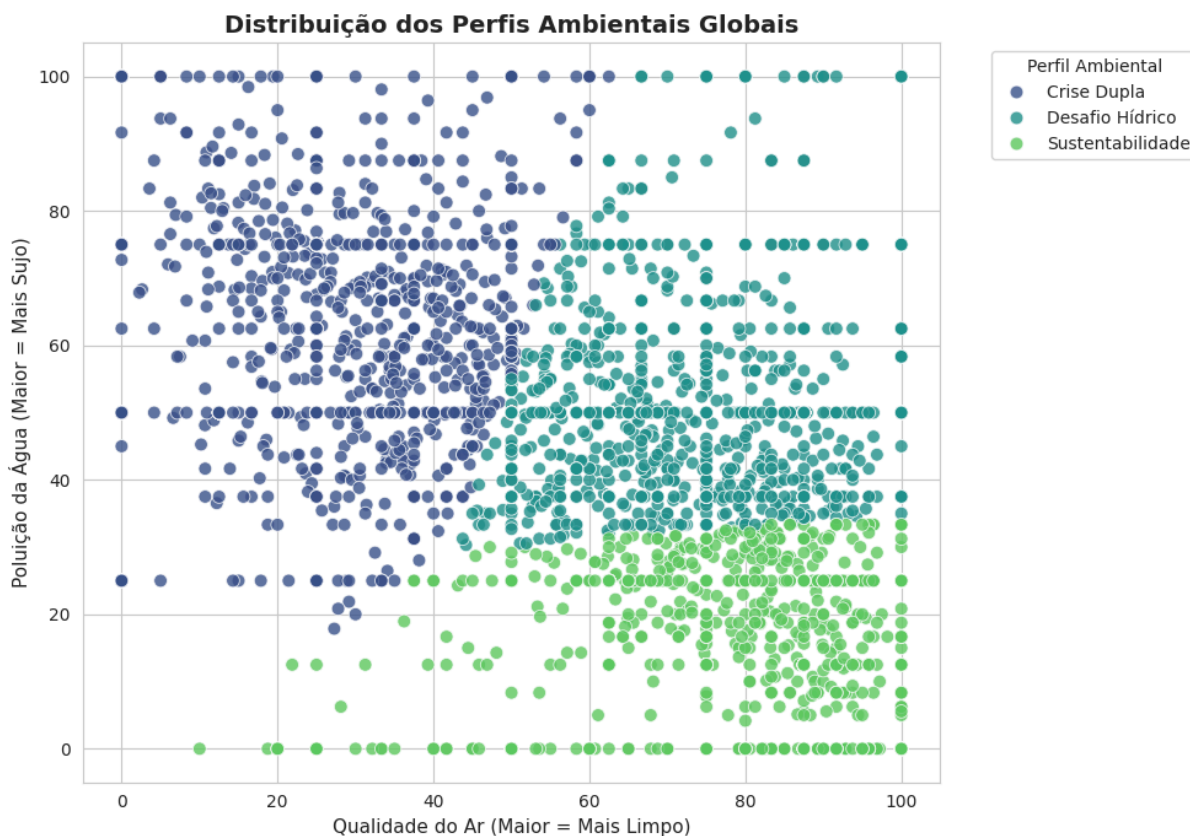


Imagem 5 - Distribuição dos Perfis Ambientais

A análise visual confirma a separação não linear dos dados:

- O Cluster "Sustentabilidade" concentra-se no quadrante inferior direito (Alta Qualidade do Ar, Baixa Poluição da Água).
- O Cluster "Crise Dupla" ocupa a região superior esquerda, evidenciando a correlação negativa entre qualidade do ar e pureza da água.
- O Cluster "Desafio Hídrico" ocupa uma região híbrida. Nota-se que, no eixo X (Ar), ele se sobrepõe ao cluster de sustentabilidade, mas no eixo Y (Água), ele ascende aos níveis de poluição do cluster de crise.

Validação Estatística

Para validar a distinção estatística entre os grupos, utilizou-se a visualização por *Boxplots* (Diagrama de Caixa), permitindo comparar as medianas e a dispersão dos dados para cada variável.

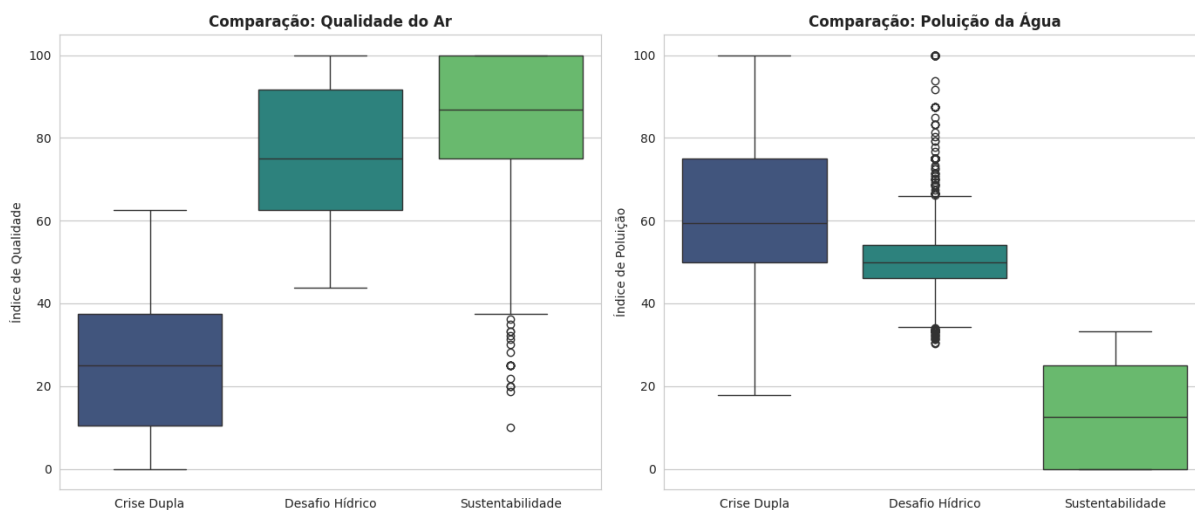


Imagem 6 - Comparação entre a Qualidade do Ar e a Poluição da Água

No gráfico da esquerda (Qualidade do Ar): A caixa do "Desafio Hídrico" está alinhada com a de "Sustentabilidade". Isso prova que, estatisticamente, a qualidade do ar desses dois grupos é similarmente boa.

No gráfico da direita (Poluição da Água): A caixa do "Desafio Hídrico" eleva-se e alinha-se com a de "Crise Dupla". Isso prova que a água dessas cidades é tão poluída quanto a das cidades em crise.

Esta análise de variância confirma que o algoritmo K-Means foi capaz de isolar um grupo de cidades que possui "o melhor do ar" mas "o pior da água", justificando a necessidade de políticas públicas diferenciadas.

Preparação dos Dados para Regras de Associação

Para a etapa de mineração de padrões frequentes, foi necessário transformar as variáveis numéricas contínuas em dados categóricos ordinais. Diferente da abordagem binária simples, optou-se por uma discretização em 5 níveis (bins) de amplitude fixa (20 pontos cada), permitindo uma granularidade maior na detecção de regras.

As classes definidas foram:

- 0-20: Muito Baixo
- 20-40: Baixo
- 40-60: Médio
- 60-80: Alto
- 80-100: Muito Alto

A Tabela 7 exemplifica a conversão dos dados numéricos originais para estas categorias semânticas.

Amostra	Categoria Ar	Categoria Água
0	Médio	Médio
1	Alto	Médio
2	Alto	Médio
3	Alto	Baixo
4	Baixo	Alto

Tabela 7 – 5 Classes definidas

Como o algoritmo Apriori requer uma matriz lógica (Verdadeiro/Falso) para calcular o suporte dos itens, aplicou-se a técnica de *One-Hot Encoding*. As duas colunas categóricas foram expandidas em colunas binárias para cada nível possível.

A Tabela 8 apresenta o formato final da matriz "cesta de compras" (*basket*) utilizada como entrada para o algoritmo, onde o valor "True" indica a presença daquela característica na cidade.

ID	Ar_MuitoBaixo	...	Ar_Alto	...	Agua_Médio	Agua_Alto
0	False	...	False	...	True	False
1	False	...	True	...	True	False

2	False	...	True	...	True	False
3	False	...	True	...	False	False
4	False	...	False	...	False	True

Tabela 8 – Formato final da Matriz

Análise de Regras de Associação

Na etapa final do experimento, aplicou-se o algoritmo Apriori para descobrir padrões de co-ocorrência entre as categorias de poluição. Devido à alta granularidade da discretização (matriz 5x5 níveis), o parâmetro de suporte mínimo foi ajustado para 0.03 (3%) e a confiança mínima para 0.2 (20%), visando capturar regras relevantes mesmo em subgrupos menores.

O algoritmo identificou 20 conjuntos frequentes, resultando na extração das regras de associação listadas na Tabela 9, ordenadas pela métrica de confiança.

ID	Antecedente (Se...)	Consequente (Então...)	Suporte	Confiança	Interpretação
#9	<p>Água Muito Limpa</p> <p><i>(Agua_Cat_Muito Baixo)</i></p>	<p>Ar Muito Limpo</p> <p><i>(Ar_Cat_Muito Alto)</i></p>	3,8%	60,8%	<p>Forte Sinergia Positiva: Se a água é excelente, há 60% de chance do ar também ser.</p>

#3	Ar Médio	Água Média	6,7%	43,5%	Tendência à Mediocridade: Cidades medianas tendem a sê-lo em ambos os aspectos.
#0	Ar Ruim (Ar_Cat_Baixo)	Água Média	5,8%	41,6%	Incerteza: Ar ruim não garante água ruim, mas aponta para água média.
#11	Água Limpa (Agua_Cat_Baixo)	Ar Muito Limpo	7,9%	38,5%	Reforça a regra #9, indicando correlação positiva no espectro sustentável.
#1	Ar Ruim (Ar_Cat_Baixo)	Água Poluída (Agua_Cat_Alto)	4,6%	32,9%	Sinal de Crise: Em 33% dos casos, ar ruim está associado a água poluída.

Tabela 9 – Regras de Associação

A Prova da Sustentabilidade (Regra #9): Esta é a regra mais forte do conjunto. Com 60,8% de confiança, ela indica que cidades que conseguem resolver a poluição da água a níveis excelentes

("Muito Baixo") quase invariavelmente possuem qualidade do ar excelente ("Muito Alto"). Isso valida estatisticamente o Cluster 0 (Sustentabilidade), sugerindo que a excelência na gestão hídrica é um forte preditor de excelência atmosférica.

A Assimetria da Crise (Regras #0 e #1): Ao analisarmos o cenário oposto, a relação não é tão linear. Quando o antecedente é "Ar Ruim" (*Ar_Cat_Baixo*), o consequente varia entre "Água Média" (41%) e "Água Poluída" (33%). Isso corrobora a descoberta do Cluster 2 (Desafio Hídrico): ter ar ruim é um sinal de alerta, mas não é uma sentença determinística de água poluída, demonstrando que a degradação hídrica e atmosférica podem ocorrer em ritmos diferentes.

Regras de Associação e Análise de Lift

Para refinar a descoberta de padrões, aplicou-se uma etapa de limpeza textual nas regras geradas pelo algoritmo Apriori, facilitando a interpretação semântica. A Tabela 10 apresenta as regras de maior destaque, ordenadas pela confiança, mas com ênfase na métrica de Lift (Elevação), que indica a força da correlação.

Antecedente (Causa)	Consequente (Efeito)	Confiança	Lift	Análise de Impacto
Água: Muito Baixo	Ar: Muito Alto	60,9%	1,77	Sinergia Sustentável: A regra mais forte. Confirma que excelência em água quase sempre arrasta a excelência do ar junto. O Lift de 1,77 indica uma associação positiva robusta.

Ar: Baixo	Água: Alto	33,0%	2,17	O Maior Risco Identificado: Embora a confiança seja moderada (33%), o Lift de 2,17 é o maior do estudo. Isso significa que, em cidades com ar ruim, a probabilidade de encontrar água poluída é duas vezes maior do que a aleatoriedade. É a prova estatística do "Cluster de Crise".
Ar: Médio	Água: Médio	43,5%	1,11	Inércia: Cidades com desempenho mediano tendem a estagnar nesse patamar em ambas as frentes (Lift próximo de 1 indica comportamento esperado).

Tabela 10 – Regras de maior destaque

A utilização do algoritmo Apriori para identificar padrões ocultos em dados ambientais alinha-se à metodologia proposta por Bagan et al. (2025) e Billah et al. (2025), que demonstraram a eficácia desta técnica na previsão de comportamentos de poluentes urbanos que escapam às análises de correlação linear simples.

A análise combinada de *Confiança* e *Lift* permite concluir que os extremos se atraem: a excelência ambiental é um ciclo virtuoso (Ar Limpo → Água Limpa), enquanto a degradação apresenta um risco agravado de cumulatividade (Ar Ruim aumenta drasticamente o risco de Água Ruim).

Visualização Gráfica dos Padrões de Associação

Para sintetizar a complexidade das regras geradas e facilitar a identificação visual de padrões, elaboraram-se duas representações gráficas complementares.

A Figura 7 apresenta a dispersão das regras baseada nas métricas de qualidade. O tamanho e a cor dos pontos representam o *Lift* (Força da regra).

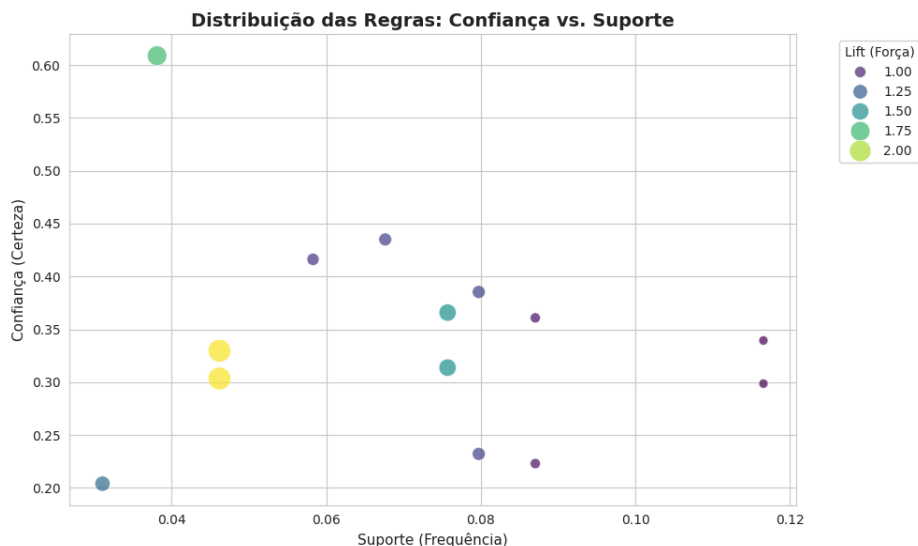


Imagem 7 - Mapa da distribuição das Regras

A análise do gráfico de dispersão revela um *trade-off* (troca) interessante: as regras com maior Lift (pontos maiores e mais escuros) tendem a ter um suporte menor. Isso indica que as associações mais fortes e críticas (como a relação de risco extremo entre ar ruim e água poluída) são fenômenos específicos, e não comportamentos médios da base inteira.

Por fim, a Figura 8 apresenta a Matriz de Causalidade (Heatmap), cruzando todos os antecedentes e consequentes encontrados. A intensidade da cor indica a Confiança da regra.

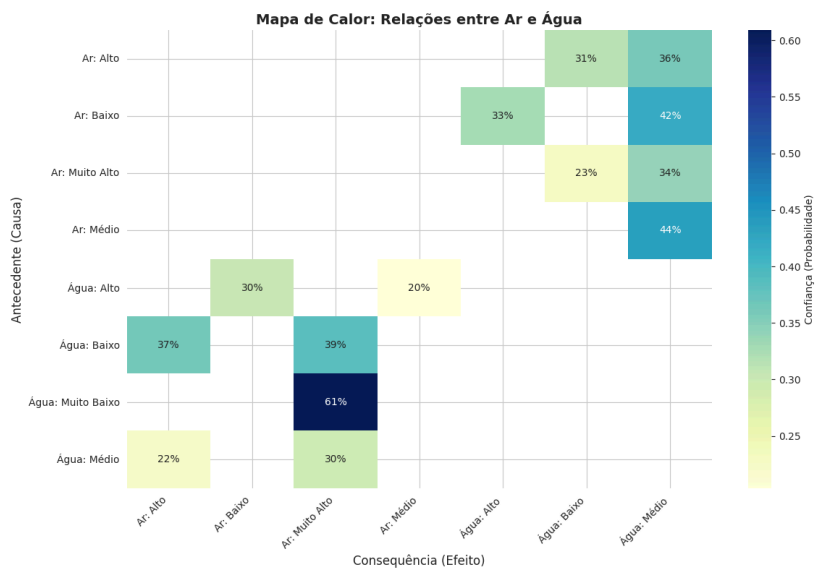


Imagem 8 - Matriz de Causalidade

O mapa de calor evidencia visualmente a assimetria das relações:

- O "Ponto Quente" da Sustentabilidade: A interseção entre *Água: Muito Baixo* e *Ar: Muito Alto* apresenta a coloração mais intensa (azul escuro), confirmando visualmente a robustez desta conexão positiva (>60%).
- A Dispersão da Crise: As categorias de poluição "Média" e "Alta" apresentam cores mais difusas, indicando que a transição entre esses estados é mais fluida e menos determinística do que o estado de excelência ambiental.

CONCLUSÃO

O presente estudo investigou, através de técnicas de Mineração de Dados, a relação entre a poluição atmosférica e hídrica em 3.963 cidades globais, questionando a premissa de correlação linear entre estes vetores de degradação.

A aplicação do algoritmo K-Means ($k = 3$) revelou um cenário global heterogêneo. Identificou-se que, enquanto 31,3% das cidades sofrem de uma "Crise Dupla" (falha simultânea nas ODS 6 e 11) e 29,2% atingiram a "Sustentabilidade Alta", a maior parcela das cidades (39,5%) enquadra-se no perfil de "Desafio Hídrico". Este agrupamento intermediário é o principal achado da pesquisa, pois comprova que a qualidade do ar não é um preditor perfeito para a qualidade da água: é plenamente possível, e comum, que cidades com boa qualidade atmosférica enfrentem graves problemas de saneamento.

A análise focada no Brasil corroborou esta tese com nuances importantes. Diferente da hipótese inicial de homogeneidade, o país apresentou cidades em todos os três perfis. Contudo, a predominância do "Desafio Hídrico" (ex: Salvador, Brasília) sugere que, embora fatores geográficos favoreçam a dispersão de poluentes aéreos em muitas capitais, a infraestrutura de saneamento permanece como o gargalo crítico para o desenvolvimento sustentável nacional.

Por fim, a mineração de Regras de Associação (Apriori) adicionou uma camada probabilística à análise. A descoberta de um *Lift* elevado (2.17) para a regra (Ar: Baixo \rightarrow Água: Alta) alerta que, embora existam exceções, a degradação ambiental tende a ser sistêmica nos casos extremos: negligenciar a qualidade do ar aumenta estatisticamente o risco de colapso hídrico.

Conclui-se que a gestão ambiental urbana não pode ser monolítica. Recomenda-se que gestores públicos utilizem essa segmentação baseada em dados para direcionar investimentos: cidades do "Desafio Hídrico" exigem foco exclusivo em saneamento (ODS 6), enquanto as da "Crise Dupla" necessitam de intervenção emergencial integrada.

Recomenda-se, portanto, que gestores públicos abandonem políticas ambientais genéricas em favor de abordagens baseadas em dados (*data-driven*). Esta conclusão corrobora os achados de Mohammed e Khalid (2021), que alertam que, embora a industrialização atue como um vetor comum de degradação para ar e água (validando nosso cluster de "Crise Dupla"), a gestão destes recursos frequentemente ocorre de forma assíncrona. Assim, cidades do "Desafio Hídrico" exigem foco

exclusivo em saneamento (ODS 6), enquanto as da "Crise Dupla" necessitam de intervenção emergencial integrada.

REFERÊNCIAS BIBLIOGRÁFICAS

SIVAKUMAR, B. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, v. 11, n. 6, p. 40-56, 2020.

HADI, S. P.; RIZKIANA, S.; SULISTYA, J. Cluster Based Classification of River Water Pollution Using K-Means for Policy Intervention. *Journal of Law, Environmental and Justice*, v. 3, n. 2, p. 102-118, 2025. (Nota: Se o ano for 2024 ou outro, ajuste, mas mantive o que estava no texto).

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: *PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES*. São Francisco: Morgan Kaufmann Publishers Inc., 1994. p. 487-499.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). Objetivos de Desenvolvimento Sustentável. Disponível em: <https://brasil.un.org/pt-br/sdgs>. Acesso em: dez. 2024.

JAPAN INTERNATIONAL COOPERATION AGENCY (JICA). Cluster Strategy for Promotion of Healthy Environment through Appropriate Environmental Regulations. JICA Report, 2022.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI). Air and Water Pollution: Burden and Strategies for Control. In: *Disease Control Priorities in Developing Countries*. 2. ed. Washington (DC): World Bank, 2006.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3. ed. Waltham: Morgan Kaufmann, 2011.

FACELI, K. et al. *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. Rio de Janeiro: LTC, 2011.

MOHAMMED, K. K.; KHALID, K. Correlation between Air Quality and Wastewater Pollution. *Asian Journal of Environment & Ecology*, v. 14, n. 4, p. 12-21, 2021.

BILLAH, M. et al. Discovering Patterns in Environmental Data for Air Quality Analysis Using Association Rule Mining. *IEEE Access*, v. 13, p. 1-15, 2025.