GERENCIAMENTO DE ÁREAS CONTAMINADAS: ANÁLISE MULTIVARIADA DE RESULTADOS QUÍMICOS E FÍSICO-QUÍMICOS DE ÁGUAS SUBTERRÂNEAS POR APRENDIZAGEM DE MÁQUINA

CONTAMINATED SITE MANAGEMENT: MULTIVARIATE ANALYSIS OF GROUNDWATER CHEMICAL AND PHYSICOCHEMICAL DATA USING MACHINE LEARNING

ALEXANDRE AUGUSTO PINHEIRO SOUZA ARAÚJO
CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

MARCELO DUDUCHI FEITOSA
CENTRO PAULA SOUZA

Comunicação:
O XIV SINGEP foi realizado em conjunto com a 14th Conferência Internacional do CIK (CYRUS Institute of Knowledge) e com a SPARK (Society for Promoting Academic Research and Knowledge), em formato híbrido, com sede presencial na UNINOVE - Universidade Nove de Julho, no Brasil.

Agradecimento à orgão de fomento:
We thank CAPES, FAPESP, and CNPq for their support of the XIV SINGEP-CIK 2026.

GERENCIAMENTO DE ÁREAS CONTAMINADAS: ANÁLISE MULTIVARIADA DE RESULTADOS QUÍMICOS E FÍSICO-QUÍMICOS DE ÁGUAS SUBTERRÂNEAS POR APRENDIZAGEM DE MÁQUINA

Objetivo do estudo
Avaliar a aplicação conjunta PCA, HCA e K-means (técnicas de aprendizado de máquina) na identificação de padrões químicos multivariados em águas subterrâneas e examinar sua utilidade como linha complementar de evidência no gerenciamento de áreas contaminadas (GAC).

Relevância/originalidade
O estudo aplica técnicas de aprendizado de máquina para interpretação de dados ambientais complexos e volumosos no contexto do GAC de forma eficaz e com abordagem multidimensional inovadora, capturando padrões em resultados laboratoriais que muitas vezes não aparecem em abordagens univariadas tradicionais.

Metodologia/abordagem
Estudo de caso quantitativo, exploratório-descritivo e retrospectivo, com três campanhas de monitoramento ambiental consideradas. Aplicou-se pré-processamento de dados, PCA, HCA pelo método de Ward e K-means, métricas de qualidade, análise de estabilidade e sensibilidade de modelo e avaliações de aderência à realidade.

Principais resultados
A estudo demonstrou capturou padrões multivariados importantes e aderentes à realidade, incrementando mais uma linha de evidência ao monitoramento ambiental no contexto do GAC Os modelos de agrupamento HCA e K-means apresentaram concordância integral e revelaram padrões de contaminação multivariados nos dados.

Contribuições teóricas/metodológicas
A abordagem apoia o GAC ao sintetizar bases multivariadas, destacar observações diferenciadas e direcionar investigações adicionais ambientais. Ficou demonstrado que o aprendizado de máquina não supervisionado é capaz tratar grandes volumes de dados ambientais provenientes de áreas contaminadas reconhecendo padrões multivariados.

Contribuições sociais/para a gestão
A estudo demonstrou como técnicas modernas de análises de dados e reconhecimento de padrões podem apoiar a tomada de decisão no GAC contribuindo para a sustentabilidade e preservação dos recursos naturais diante das pressões antrópicas crescentes.

Palavras-chave: Aprendizado de máquina não supervisionado, águas subterrâneas, gerenciamento de áreas contaminadas, análise multivariada, sustentabilidade

CONTAMINATED SITE MANAGEMENT: MULTIVARIATE ANALYSIS OF GROUNDWATER CHEMICAL AND PHYSICOCHEMICAL DATA USING MACHINE LEARNING

Study purpose
Evaluate the joint application of PCA, HCA, and K-means (machine learning techniques) in identifying multivariate chemical patterns in groundwater and examine their usefulness as a complementary line of evidence in contaminated site management (CSM).

Relevance / originality
The study applies machine learning techniques to effectively interpret complex and large-scale environmental data in the context of CSM through an innovative multidimensional approach, capturing patterns in laboratory results that often do not appear in traditional univariate approaches.

Methodology / approach
A quantitative, exploratory-descriptive, and retrospective case study considering three environmental monitoring campaigns. Data preprocessing, PCA, HCA using Ward's method, and K-means were applied, together with quality metrics, model stability and sensitivity analysis, and assessments of consistency with real-world conditions.

Main results
The study demonstrated that it captured important multivariate patterns consistent with real-world conditions, adding another line of evidence to environmental monitoring in the context of CSM The HCA and K-means clustering models showed complete agreement and revealed multivariate contamination patterns in the

Theoretical / methodological contributions
The approach supports CSM by synthesizing multivariate datasets, highlighting differentiated observations, and guiding additional environmental investigations. It was demonstrated that unsupervised machine learning is capable of handling large volumes of environmental data from contaminated sites while identifying multivariate patterns.

Social / management contributions
The study demonstrated how modern data analysis and pattern recognition techniques can support decision-making in CSM, contributing to sustainability and the preservation of natural resources in the face of increasing anthropogenic pressures.

Keywords: Unsupervised machine learning, groundwater, contaminated site management, multivariate analysis, sustainability

GERENCIAMENTO DE ÁREAS CONTAMINADAS: ANÁLISE MULTIVARIADA DE RESULTADOS QUÍMICOS E FÍSICO-QUÍMICOS DE ÁGUAS SUBTERRÂNEAS POR APRENDIZAGEM DE MÁQUINA

1 Introdução

As águas subterrâneas constituem recurso estratégico para o abastecimento humano, as atividades econômicas e a manutenção dos ecossistemas (Hirata et al., 2019). Sua proteção está diretamente relacionada à conservação da qualidade do meio físico, uma vez que substâncias liberadas no solo podem alcançar o meio subterrâneo e comprometer os aquíferos. Nesse contexto, o gerenciamento de áreas contaminadas insere-se como instrumento de proteção integrada do solo e das águas subterrâneas, contribuindo para a prevenção e o controle de impactos ambientais e para o uso sustentável dos recursos naturais.

O gerenciamento de áreas contaminadas (GAC) envolve a integração progressiva de informações sobre fontes de contaminação, substâncias químicas de interesse, condições do meio físico e comportamento dos contaminantes, de modo a subsidiar as etapas de investigação, avaliação e intervenção. Nesse processo, o monitoramento das águas subterrâneas constitui importante fonte de informação, devendo seus resultados ser interpretados em conjunto com o modelo conceitual e as demais evidências disponíveis para a área (CETESB, 2024).

Redes de monitoramento de águas subterrâneas compostas por diferentes poços, campanhas, substâncias químicas e parâmetros físico-químicos produzem matrizes multivariadas cuja complexidade pode dificultar a identificação de relações simultâneas entre variáveis e de similaridades entre os perfis dos pontos monitorados. A avaliação individual dos parâmetros permanece necessária, porém pode não representar integralmente a estrutura conjunta dos dados. Zhang et al. (2023) destacam que conjuntos hidrogeoquímicos de elevada dimensionalidade apresentam relações internas complexas que podem demandar abordagens analíticas complementares.

Costanzo et al. (2026) demonstraram que a interpretação de agrupamentos hidrogeoquímicos pode ser fortalecida quando esses padrões são confrontados com informações ambientais obtidas por métodos independentes, como dados geofísicos. Nguyen et al. (2025) demonstraram a aplicação de análise de componentes principais (PCA) e análise hierárquica de agrupamentos (HCA) para simplificar bases ambientais complexas e reconhecer padrões composicionais. De forma complementar, Panneerselvam et al. (2026) utilizaram PCA, HCA e K-means na avaliação de águas subterrâneas, evidenciando a aplicabilidade conjunta da redução dimensional e das técnicas de agrupamento.

A utilização desses métodos, entretanto, exige cautela interpretativa. Um cluster representa inicialmente uma estrutura estatística de similaridade definida pelas variáveis e pelo processamento empregado. Assim, a inclusão de dois poços no mesmo agrupamento não demonstra, isoladamente, pertencimento à mesma pluma, origem comum da contaminação ou conectividade hidráulica. Costanzo et al. (2026) demonstraram que agrupamentos hidrogeoquímicos adquirem maior significado ambiental quando confrontados com informações independentes. Da mesma forma, fatores temporais, espaciais, hidrogeológicos e geoquímicos podem influenciar os padrões observados e precisam ser considerados posteriormente na interpretação (Peng et al., 2025).

Dessa forma, o problema investigado neste trabalho consiste em determinar em que medida estruturas multivariadas identificadas exclusivamente a partir dos resultados químicos e físico-químicos podem acrescentar informação útil ao GAC sem que similaridade estatística seja confundida com evidência direta de fonte, pluma ou conexão hidrogeológica. Para isso, propõe-se separar a etapa de descoberta dos padrões de sua interpretação ambiental: os agrupamentos são inicialmente obtidos sem utilização do histórico da área como rótulo e, posteriormente, são avaliados quanto à consistência estatística, às variáveis responsáveis por

sua diferenciação e à correspondência com informações independentes do gerenciamento ambiental.

A contribuição do estudo não reside na proposição de novos algoritmos, mas na aplicação estruturada de técnicas consolidadas de PCA, HCA e K-means a uma rede real de monitoramento, associando redução de dimensionalidade, agrupamento, avaliação de robustez e posterior confronto com evidências ambientais. Nesse enquadramento, a questão de pesquisa é a seguinte: em que medida a combinação de análise de componentes principais e técnicas de agrupamento permite identificar padrões químicos multivariados em uma rede de monitoramento de águas subterrâneas e produzir uma linha complementar de evidência para o gerenciamento de áreas contaminadas?

O objetivo do presente trabalho é, portanto, avaliar a aplicação de técnicas de aprendizado de máquina não supervisionado, combinando PCA e análise de agrupamentos, para identificar padrões químicos multivariados em águas subterrâneas e examinar sua utilização como linha complementar de evidência no GAC. Para tanto, busca-se caracterizar a estrutura multivariada dos resultados químicos e físico-químicos, identificar grupos e observações diferenciadas por HCA e K-means, avaliar sua consistência por métricas internas, concordância entre métodos, estabilidade e análises de sensibilidade, identificar as variáveis que mais contribuem para essa diferenciação e, posteriormente, confrontar os padrões obtidos com informações independentes, incluindo histórico de contaminação, ocorrência de LNAPL, comportamento temporal, localização dos poços e contexto hidrogeológico.

É ainda importante não perder de vista que, no processo de GAC, uma composição de diversas linhas de evidência sustenta a tomada de decisão e, portanto, não se pode confundir a adição de uma linha de evidência por meio de algoritmos de aprendizado de máquina com uma ferramenta determinística de tomada de decisão autossuficiente e excludente de outras linhas de evidência.

2 Referencial Teórico

O gerenciamento de áreas contaminadas (GAC) constitui um processo progressivo de aquisição e interpretação de informações sobre fontes, substâncias químicas de interesse, características do meio físico e comportamento dos contaminantes. Essas informações subsidiam o modelo conceitual da área e as decisões relacionadas à investigação, avaliação e intervenção. No caso das águas subterrâneas, os resultados obtidos nos poços de monitoramento constituem uma linha quantitativa de evidência que deve ser interpretada em conjunto com o contexto espacial, hidrogeológico e histórico da área (CETESB, 2024).

Redes de monitoramento podem produzir matrizes de elevada dimensionalidade ao combinarem diferentes poços de monitoramento de água subterrânea, campanhas, substâncias químicas e parâmetros físico-químicos. Além do número de variáveis, diferenças de escala, unidades, assimetria, não detecções e correlações entre parâmetros aumentam a complexidade analítica. Nesse contexto, técnicas multivariadas permitem avaliar simultaneamente relações entre variáveis e reconhecer estruturas que podem não ser evidentes em análises parâmetro a parâmetro. Nguyen et al. (2025), por exemplo, utilizaram PCA e HCA para reduzir a complexidade de grandes conjuntos ambientais e reconhecer padrões composicionais.

O aprendizado de máquina não supervisionado busca identificar estruturas nos dados sem utilização prévia de classes conhecidas. Em uma rede de monitoramento, isso permite investigar similaridades entre os perfis químicos sem informar previamente quais poços são considerados contaminados, hotspots, integrantes de determinada pluma ou associados a fontes específicas. A literatura apresenta aplicações dessas técnicas em águas subterrâneas combinadas posteriormente com informações hidrogeoquímicas, geológicas e hidrogeológicas,

reforçando que os agrupamentos constituem inicialmente estruturas estatísticas cuja interpretação ambiental requer evidências adicionais (Costanzo et al., 2026; Zhang et al., 2023).

2.1 Análise de componentes principais

A análise de componentes principais (PCA) é uma técnica multivariada de redução de dimensionalidade que transforma variáveis originais correlacionadas em componentes obtidos por combinações lineares. Os componentes concentram progressivamente a variabilidade do conjunto de dados, permitindo representar sua estrutura em um número menor de dimensões (Fávero & Belfiore, 2025).

Na interpretação da PCA, os escores representam a posição das observações no espaço dos componentes, enquanto os loadings indicam a contribuição das variáveis originais para essas dimensões. Em aplicações hidroquímicas, a técnica pode sintetizar variáveis correlacionadas e auxiliar na identificação dos principais fatores associados à variabilidade da qualidade das águas subterrâneas (Mohammed et al., 2025). No presente estudo, a PCA é empregada tanto para reduzir a dimensionalidade quanto para apoiar a interpretação dos padrões químicos e físico-químicos posteriormente submetidos às técnicas de agrupamento.

2.2 Análise hierárquica de agrupamentos e método de Ward

A análise hierárquica de agrupamentos (HCA) organiza observações segundo medidas de similaridade ou dissimilaridade, construindo sucessivamente grupos a partir das relações de proximidade existentes nos dados. O resultado pode ser representado por um dendrograma, no qual é possível examinar a sequência das fusões, subestruturas, observações isoladas e diferentes possibilidades de particionamento do conjunto (Fávero & Belfiore, 2025).

Em dados métricos, a distância euclidiana constitui uma das medidas de proximidade mais utilizadas. Como essa distância é influenciada pela escala das variáveis, a padronização prévia é importante quando os parâmetros apresentam diferentes unidades e ordens de grandeza. No método de Ward, empregado neste estudo, as fusões são realizadas buscando limitar o aumento da variabilidade interna decorrente da combinação dos grupos. O procedimento parte das observações individuais e realiza sucessivas agregações, cuja estrutura é representada pelo dendrograma (Costanzo et al., 2026; Fávero & Belfiore, 2025).

Para redes de monitoramento, o dendrograma permite avaliar não apenas uma solução final de agrupamento, mas também relações hierárquicas entre os poços. Essas relações, entretanto, representam similaridade no espaço multivariado analisado e não podem ser interpretadas diretamente como evidência de conectividade hidráulica, propagação de contaminantes ou origem comum.

2.3 Agrupamento K-means

O K-means é um método de agrupamento não hierárquico que particiona as observações em um número previamente definido de grupos. O algoritmo utiliza centroides como centros de aglomeração e atribui cada observação ao grupo correspondente ao centro mais próximo, buscando obter maior similaridade interna e maior diferenciação entre os grupos (Fávero & Belfiore, 2025).

Sua natureza particional é complementar à HCA. Enquanto o dendrograma permite observar diferentes níveis de agregação e subestruturas existentes no conjunto, o K-means fornece uma classificação explícita para determinado número de grupos. A utilização conjunta das técnicas possibilita verificar se estruturas semelhantes são recuperadas por procedimentos

distintos. Essa combinação tem sido empregada em estudos hidrogeoquímicos, incluindo aplicações de PCA, HCA e K-means para diferenciação de grupos com características composicionais distintas (Panneerselvam et al., 2026).

2.4 Validação e interpretação dos agrupamentos

A obtenção de uma partição matemática não é suficiente para estabelecer sua adequação. Em problemas não supervisionados, a avaliação deve combinar diferentes critérios. Métricas como Silhouette, Calinski-Harabasz e Davies-Bouldin permitem avaliar propriedades internas dos agrupamentos, enquanto a estrutura do dendrograma, a distribuição das observações e a concordância entre métodos oferecem evidências adicionais para escolha da solução. Guan et al. (2025) utilizaram conjuntamente os índices Silhouette, Calinski-Harabasz e Davies-Bouldin para avaliar o desempenho de métodos de agrupamento empregados na classificação do risco de poluição de águas subterrâneas em uma instalação industrial.

A robustez pode ser adicionalmente examinada pela concordância entre HCA e K-means, por procedimentos de reamostragem e por análises de sensibilidade às decisões de pré-processamento. Entretanto, consistência estatística e significado ambiental representam níveis distintos de avaliação. Uma solução bem separada matematicamente não demonstra, isoladamente, a existência de uma fonte, pluma ou unidade hidrogeológica. No contexto do GAC, os agrupamentos são mais adequadamente tratados como uma linha complementar de evidência, cuja interpretação depende da caracterização das variáveis responsáveis pela diferenciação e do confronto posterior com informações ambientais independentes.

3 Metodologia

A metodologia adotada no desenvolvimento do presente trabalho foi compartimentada em etapas, conforme apresentado a seguir.

3.1 Delineamento da pesquisa

Este estudo apresenta abordagem quantitativa, natureza aplicada e caráter exploratório-descritivo, sendo desenvolvido segundo delineamento observacional e retrospectivo. Quanto ao procedimento, caracteriza-se como estudo de caso único, no qual dados secundários provenientes do monitoramento de águas subterrâneas de uma área monitorada são submetidos a técnicas de análise multivariada e aprendizado de máquina não supervisionado. O caráter exploratório decorre da busca por estruturas presentes nos próprios dados químicos e físico-químicos obtidos a partir ensaios em amostras destas águas subterrâneas realizadas por laboratórios acreditados de acordo com norma ISO/IEC 17025 (2017), sem utilização prévia de classificações ambientais dos poços como variável-resposta.

O delineamento não envolve manipulação experimental das condições ambientais nem treinamento dos algoritmos a partir de classes previamente conhecidas. Os padrões multivariados são determinados exclusivamente a partir da matriz de dados utilizada na modelagem. Somente após essa etapa os resultados são confrontados com informações ambientais independentes, preservando a distinção entre identificação estatística de padrões e sua posterior interpretação no contexto do gerenciamento de áreas contaminadas.

3.2 Área de estudo

A aplicação empírica foi realizada na área denominada PDD, integrante de um empreendimento industrial localizado no Estado de São Paulo. A área integra uma rede de

monitoramento de águas subterrâneas inserida em programa de acompanhamento ambiental do empreendimento. Os documentos técnicos disponíveis para a área incluem histórico ambiental, dados de potenciometria e sentido do fluxo subterrâneo, parâmetros físico-químicos medidos in situ e resultados laboratoriais de diferentes grupos de substâncias, incluindo BTEX, hidrocarbonetos policíclicos aromáticos e hidrocarbonetos de petróleo.

3.3 Base de dados e campanhas

Foram analisadas três campanhas de monitoramento, identificadas no banco de dados como 2025_1, 2025_2 e 2026_1, correspondentes, respectivamente, à primeira campanha de 2025 considerada no estudo, à segunda campanha de 2025 e à primeira campanha de 2026. A matriz final continha 28 poços na campanha 2025_1, 21 na campanha 2025_2 e 25 na campanha 2026_1.

Cada registro foi estruturado por poço e campanha, contendo identificador da amostra, coordenadas espaciais e as variáveis de interesse para este estudo: parâmetros físico-químicos de campo e resultados químicos laboratoriais associados aos poços de monitoramento. Entre os grupos analíticos disponíveis encontram-se BTEX, PAHs, frações de hidrocarbonetos de petróleo, compostos orgânicos voláteis, n-alcanos e metais dissolvidos.

3.4 Seleção das variáveis

A seleção das variáveis foi realizada antes da PCA e independentemente dos agrupamentos obtidos, evitando que o conhecimento posterior dos clusters interferisse na definição da matriz de entrada. Foram excluídos identificadores, campos administrativos, controles de qualidade laboratorial, padrões internos, surrogates, cabeçalhos ambíguos e metais totais. As coordenadas dos poços de monitoramento também foram mantidas fora da modelagem e utilizadas apenas posteriormente na avaliação espacial.

O cenário combinou resultados químicos laboratoriais com seis parâmetros físico-químicos medidos em campo: condutividade elétrica, potencial de oxirredução, oxigênio dissolvido, pH, temperatura e turbidez. Todos os metais totais foram excluídos e somente o subconjunto previamente curado de metais dissolvidos foi mantido como candidato à modelagem, decisão adotada para reduzir a influência potencial de variações de background metálico sobre a estrutura multivariada. Essa escolha constitui uma decisão específica do delineamento deste estudo, e não uma regra geral para análise de áreas contaminadas.

3.5 Tratamento dos dados

As variáveis químicas foram submetidas a critérios de qualidade antes da modelagem. Foram excluídas variáveis com proporção de dados ausentes superior a 40%, proporção de zeros superior a 95% ou menos de três resultados não nulos. Posteriormente, foi aplicado filtro de redundância baseado em correlação de Spearman absoluta igual ou superior a 0,95 entre variáveis químicas.

No cenário principal, resultados químicos registrados como não detectados ou censurados foram representados por zero, preservando a convenção utilizada na análise principal. Como avaliação de sensibilidade, foi adotado alternativamente o valor correspondente à metade do limite analítico quando esse limite estava explicitamente disponível no banco de dados.

3.6 Transformação e padronização dos dados

As variáveis químicas elegíveis e não negativas foram submetidas à transformação log1p para reduzir a assimetria das distribuições, enquanto os parâmetros físico-químicos foram mantidos sem transformação logarítmica. Em seguida, as variáveis foram padronizadas pelo StandardScaler antes da PCA. Valores ausentes dos parâmetros físico-químicos não foram convertidos em zero, sendo exigida, no cenário principal, a disponibilidade dos seis parâmetros de campo. O RobustScaler foi utilizado posteriormente como cenário de sensibilidade.

3.7 Análise de componentes principais

A PCA foi aplicada separadamente a cada campanha após as etapas de seleção, transformação e padronização de variáveis. Para a etapa de agrupamento, foram retidos os componentes necessários para atingir pelo menos 80% da variância acumulada. Os escores correspondentes a esses componentes constituíram o espaço multivariado utilizado nas análises de agrupamento.

Além da redução dimensional, foram examinados os percentuais de variância explicada, os escores dos poços e os *loadings* das variáveis. As representações tridimensionais baseadas em PC1, PC2 e PC3 foram utilizadas para visualização e interpretação, sem pressupor que esses componentes representassem integralmente o espaço empregado na clusterização quando mais componentes fossem necessários para atingir o critério de 80%.

3.8 Análise hierárquica de agrupamentos

A análise hierárquica de agrupamentos foi realizada sobre os escores da PCA por meio de procedimento aglomerativo, utilizando o método de Ward e distância euclidiana. A HCA foi empregada para examinar a estrutura hierárquica dos poços e gerar dendrogramas para cada campanha. Estes, por sua vez, foram utilizados para avaliar a sequência das fusões, identificar subestruturas, observar pontos que permanecessem isolados até maiores distâncias de ligação e comparar diferentes possibilidades de corte. A interpretação foi restrita à similaridade multivariada das observações. Dessa forma, a associação entre dois poços em um mesmo ramo não foi considerada, isoladamente, evidência de conectividade hidráulica, origem comum ou pertencimento à mesma pluma de contaminação.

3.9 Algoritmo K-means e definição do número de agrupamentos

O K-means foi aplicado ao mesmo espaço de escores utilizado pela HCA. Foram avaliadas soluções entre K = 2 e K = 6, utilizando inicialização k-means++ e 100 inicializações independentes com semente 42, sendo que a definição de K não foi baseada em um único indicador. Para cada solução foram calculados o coeficiente de Silhouette, o índice Calinski-Harabasz, o índice Davies-Bouldin e a soma dos quadrados intragrupos, utilizada para avaliação do cotovelo. Também foram considerados tamanho dos grupos, ocorrência de grupos que contivessem exatamente um único ponto de dados, estabilidade por subamostragem, concordância com a HCA e interpretabilidade química.

3.10 Concordância entre HCA e K-means

A concordância entre as classificações produzidas pela HCA e pelo K-means foi

quantificada pelo Adjusted Rand Index (ARI) para cada valor de K. O ARI foi utilizado como medida de concordância entre partições produzidas por métodos distintos.

Essa comparação foi tratada como evidência de consistência metodológica, e não como validação ambiental independente, uma vez que HCA e K-means foram aplicados à mesma matriz de entrada. A interpretação conjunta incluiu o valor do ARI e a inspeção dos dendrogramas.

3.11 Estabilidade dos agrupamentos

A estabilidade das soluções foi avaliada por subamostragem repetida. Para cada K foram executadas 200 repetições, utilizando aleatoriamente 80% das observações em cada ciclo, sem reposição. O modelo obtido na subamostra foi utilizado para classificar o conjunto e sua concordância com a solução de referência foi quantificada pelo ARI.

A distribuição dos valores de ARI foi caracterizada principalmente pela mediana e pelo quantil inferior, permitindo avaliar não apenas o comportamento central, mas também a existência de reamostragens em que a estrutura se mostrasse menos estável.

3.12 Análises de sensibilidade

Foram definidos cenários alternativos para avaliar a dependência dos agrupamentos em relação às principais decisões de pré-processamento. As análises incluíram: substituição do StandardScaler pelo RobustScaler, execução sem PCA, execução sem filtro de redundância, utilização somente da química laboratorial, excluindo os seis parâmetros físico-químicos e tratamento dos resultados censurados por metade do limite analítico quando esse valor estava disponível.

Cada solução alternativa foi comparada à classificação principal por meio do ARI. Essas análises foram utilizadas para avaliar robustez metodológica e identificar decisões de tratamento com maior influência sobre a geometria dos agrupamentos.

3.13 Caracterização química dos agrupamentos

Após a definição dos grupos, sua interpretação química foi realizada por meio de estatísticas descritivas e medidas de diferenciação. Foram examinadas medianas, intervalos interquartis e frequências de detecção das variáveis em cada grupo. Também foi calculado o tamanho de efeito η^2 para quantificar, de forma descritiva, a proporção da variabilidade de cada variável associada à diferenciação entre os agrupamentos.

O η^2 foi utilizado como medida de caracterização dos perfis e não como teste de significância ou evidência inferencial. Em soluções contendo grupos com uma única observação, valores elevados foram interpretados com cautela, uma vez que podem refletir o forte afastamento dessa observação em relação ao conjunto majoritário.

A caracterização foi complementada pela análise dos *loadings* da PCA, das distâncias aos centroides do K-means e dos valores individuais de Silhouette. O objetivo dessa etapa foi identificar quais substâncias ou parâmetros físico-químicos estavam associados às diferenças multivariadas observadas, sem atribuir causalidade ambiental exclusivamente a essas associações.

3.14 Avaliação temporal

A análise temporal foi conduzida posteriormente às clusterizações independentes de cada campanha. Foram considerados os poços comuns entre campanhas sucessivas, avaliando-se persistência no padrão majoritário, transições entre perfis e mudanças na condição de observações diferenciadas.

Como os números atribuídos aos clusters são arbitrários entre execuções e campanhas, os grupos não foram comparados diretamente por seus rótulos numéricos, o que certamente geraria confusão. A comparação considerou a correspondência entre os conjuntos de poços e seus perfis. A mudança de classificação entre campanhas foi interpretada como alteração da estrutura química multivariada, e não como evidência automática de migração de contaminantes. Quando pertinente, o teste de Mann-Kendall, já presente na documentação técnica do projeto associado à área PDD, foi utilizado apenas como informação temporal complementar para combinações específicas de poço e contaminante nas avaliações.

3.15 Avaliação espacial

As coordenadas dos poços não foram utilizadas como variáveis na PCA, HCA ou K-means. A distribuição espacial dos agrupamentos foi avaliada somente após a obtenção das classificações químicas, permitindo investigar se a similaridade multivariada apresentava correspondência espacial sem induzir previamente essa estrutura no algoritmo.

A compactação espacial dos grupos foi avaliada por meio de 499 permutações aleatórias dos rótulos, comparando a configuração observada com distribuições obtidas sob aleatorização. A análise espacial foi considerada uma linha complementar de interpretação, não sendo proximidade geográfica utilizada como evidência suficiente de similaridade química, conectividade hidráulica ou pertencimento à mesma pluma de contaminação do solo.

3.16 Confronto com evidências independentes do GAC

A última etapa consistiu no confronto dos padrões multivariados com informações ambientais não utilizadas para definir os agrupamentos. Foram considerados o histórico de contaminação, ocorrências documentadas de LNAPL, resultados de campanhas anteriores, localização dos poços, potenciometria, sentido de fluxo da água subterrânea, contexto hidrogeológico e comportamento temporal de contaminantes selecionados. A documentação técnica da PDD reúne separadamente histórico ambiental, potenciometria, parâmetros físico-químicos e resultados químicos, possibilitando essa avaliação posterior.

O confronto foi estruturado em quatro possibilidades interpretativas: convergência, quando o padrão multivariado mostrou correspondência com evidência ambiental documentada, complementaridade, quando acrescentou informação à interpretação existente e discordância, quando a estrutura estatística não acompanhou a evidência. A classificação estatística não foi utilizada isoladamente para determinar fontes, plumas de contaminação ou conexões hidrogeológicas. Assim, o aprendizado não supervisionado foi avaliado como uma linha complementar de evidência para o GAC, e não como substituto do modelo conceitual da área.

4 Análise dos Resultados e Discussões

Após aplicação da metodologia proposta, os resultados obtidos foram avaliados conforme itens seguintes.

4.1 Estrutura multivariada dos resultados químicos e físico-químicos

A análise dos resultados concentrou-se na identificação dos poços que se diferenciaram do padrão multivariado predominante em cada campanha. A PCA foi utilizada para sintetizar a variabilidade dos dados, enquanto HCA e K-means foram empregados para verificar a estrutura dos agrupamentos. Observou-se nos resultados da PCA que em 2025_1, a maior componente de variabilidade (PC1) foi principalmente associado a PAHs e hidrocarbonetos, em 2025_2, ao BTEX, frações aromáticas e PAHs, já em 2026_1, aos PAHs.

A avaliação das soluções entre K = 2 e K = 6 resultou na seleção de K = 2 nas três campanhas. A escolha não foi baseada isoladamente no coeficiente de Silhouette, mas na avaliação conjunta das métricas internas de agrupamento, incluindo Silhouette, Calinski-Harabasz e Davies-Bouldin, da estabilidade por subamostragem, da distribuição das observações entre os grupos e da concordância entre HCA e K-means. A tabela 1 resume as métricas escolhidas e resultados.

Tabela 1 - Síntese da PCA e dos agrupamentos por campanha

Campanha	PCs ≥80%	PC1+PC2+PC3	K	Silhouette	ARI HCA / K-means	Pontos diferenciados
2025_1	7	56,14%	2	0,4026	1,000	PM-16; PM-18; PM-36; WLM-PM-40; WLM-PM-42
2025_2	7	53,01%	2	0,5019	1,000	PM-36
2026_1	5	71,49%	2	0,7202	1,000	PM-01

A PCA identificou que os três primeiros componentes explicaram 56,14% da variância em 2025_1, 53,01% em 2025_2 e 71,49% em 2026_1. Embora esses três componentes sejam suficientes para uma representação gráfica tridimensional informativa conforme figura 2, os agrupamentos K-means foram calculados com todos os componentes necessários para representar pelo menos 80% da variância acumulada, sete nas duas campanhas de 2025 e cinco na primeira campanha de 2026, as variáveis que mais contribuem para a diferenciação entre os agrupamentos, ordenadas segundo o tamanho de efeito η^2, estão representados na figura 1.

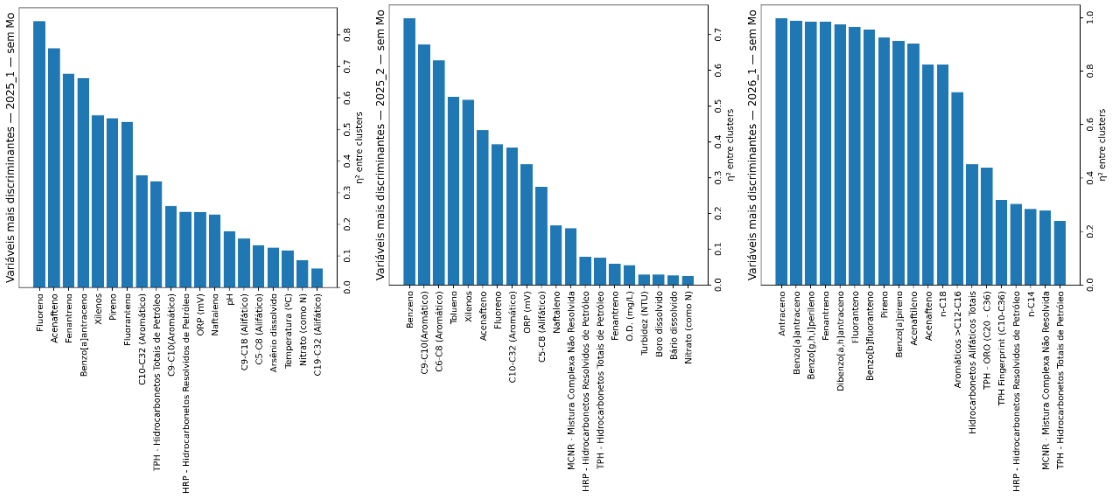


Figura 1 - Ranqueamento de variáveis que mais contribuem para a diferenciação entre os agrupamentos

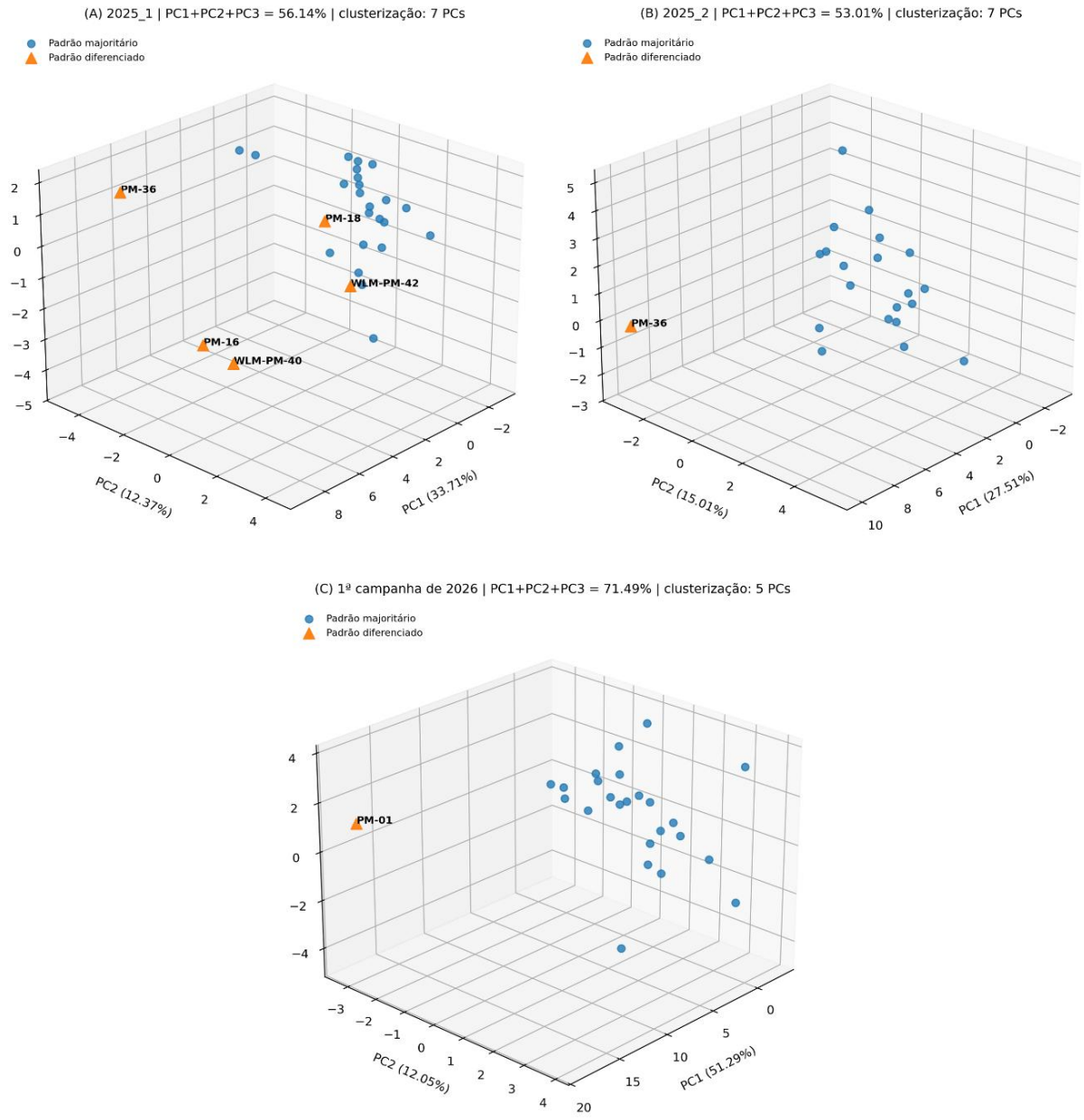


Figura 2 - Representação tridimensional da PCA para as três campanhas avaliadas.

Conforme apresentado nos rótulos dos pontos plotados no espaço tridimensional, foram diferenciados o poço PM-01, para a campanha 2026_1, o poço PM-36 relativo à campanha 2025_2 e os poços PM-16, PM-18, PM-36, WLM-PM-40 e WLM-PM-42 referentes à 2025_1. A figura 3 exibe os dendrogramas da HCA e confirma a mesma estrutura principal de agrupamento identificada pelo K-means em relação aos clusters observados a partir das componentes principais.

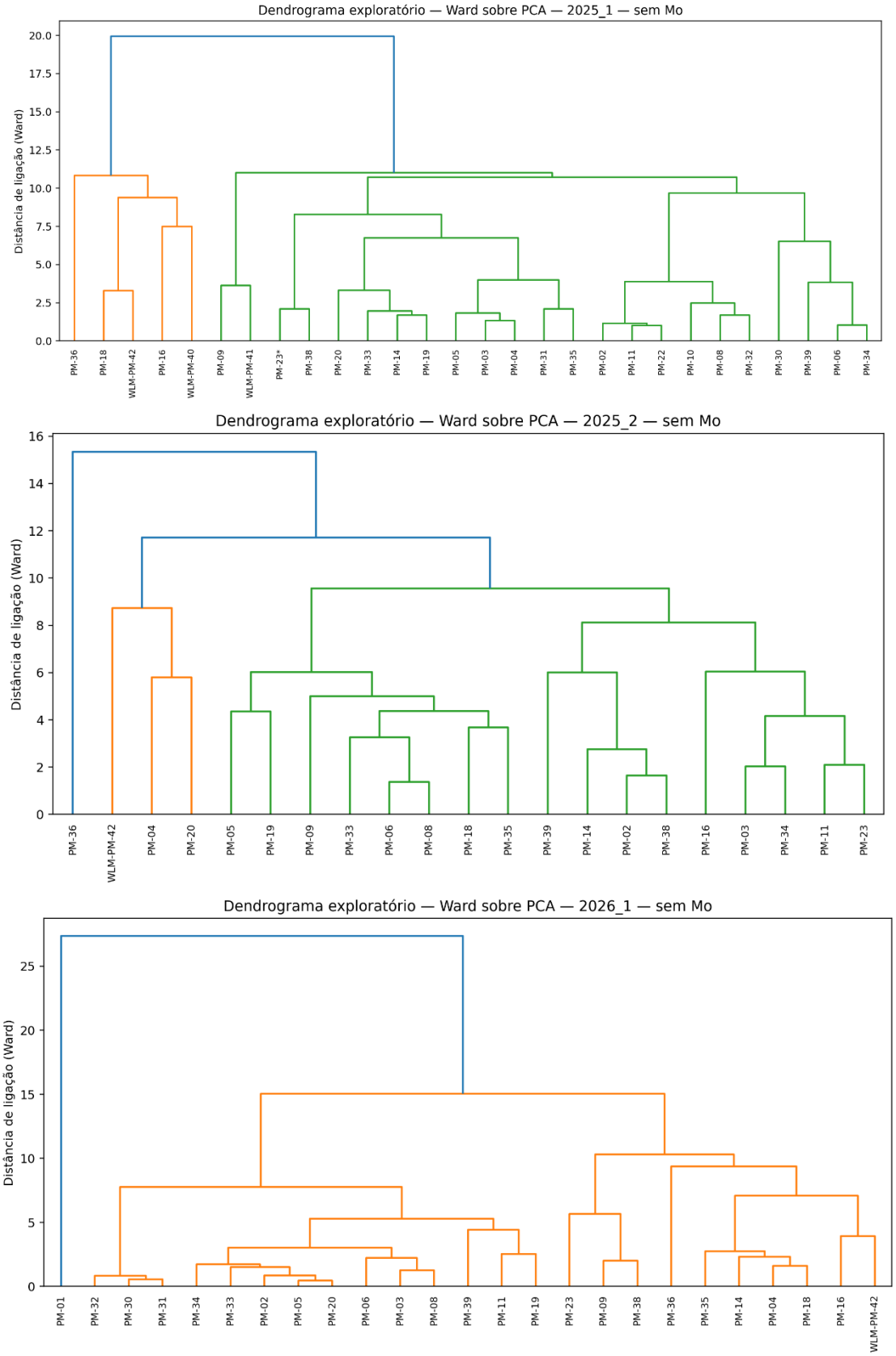


Figura 3 - Dendrogramas obtidos por HCA e método WARD.

Avaliando-se os dendrogramas, observa-se que, em 2025_1, PM-16, PM-18, PM-36, WLM-PM-40 e WLM-PM-42 formaram um ramo diferenciado, enquanto em 2025_2 o PM-36

constituiu o principal isolamento hierárquico. Em 2026_1, o PM-01 permaneceu isolado até a maior distância de ligação. Nos dois últimos casos, a solução K = 2 deve ser interpretada como separação de uma observação multivariadamente anômala, e não como evidência de dois regimes ambientais equivalentes.

4.2 Caracterização química dos padrões identificados

A caracterização dos agrupamentos mostrou que as variáveis associadas à diferenciação variaram entre as campanhas. Em 2025_1, destacaram-se principalmente PAHs e hidrocarbonetos, especialmente fluoreno, acenafteno, fenantreno e benzo[a]antraceno. O grupo diferenciado apresentou ainda maiores concentrações de TPH Total e da fração aromática C10-C32, associadas a valores de ORP inferiores aos observados no padrão majoritário.

Em 2025_2, a diferenciação foi dominada pelo benzeno e por frações aromáticas, com o PM-36 apresentando perfil fortemente distinto do conjunto predominante. Na primeira campanha de 2026, o PM-01 foi diferenciado principalmente por PAHs, com elevados valores de η^2 para diversos compostos e ORP inferior ao padrão majoritário.

Os elevados valores de η^2 em 2026_1 devem ser interpretados de forma descritiva, pois o grupo diferenciado contém exclusivamente o PM-01 e reflete principalmente seu forte afastamento químico em relação aos demais poços, não constituindo evidência inferencial.

Em conjunto, os resultados mostram que os clusters refletem perfis químicos multivariados, e não a ultrapassagem isolada de valores de concentração. Assim, uma excedência individual pode não resultar em diferenciação quando o conjunto das variáveis permanece semelhante ao padrão predominante.

4.3 Robustez e sensibilidade dos agrupamentos

A estabilidade mediana das soluções K = 2 foi igual a 1,000 nas três campanhas. Entretanto, os quantis inferiores de 5% foram de 0,4001 em 2025_1, 0,2805 em 2025_2 e 0,0391 em 2026_1, indicando que as soluções não permaneceram invariantes em todas as reamostragens. Esse comportamento é particularmente relevante em 2025_2 e 2026_1, cujas soluções contêm *singletons*, uma vez que a retirada de uma observação fortemente diferenciada durante a subamostragem pode alterar substancialmente a partição obtida.

Nos testes de sensibilidade, a utilização de LQ/2 para os resultados censurados, a retirada da PCA, a retirada do filtro de redundância e a utilização exclusiva das variáveis químicas mantiveram ARI = 1,000 nas três campanhas. A principal alteração ocorreu com a substituição do StandardScaler pelo RobustScaler, que resultou em ARI de 0,675 em 2025_1, 0,404 em 2025_2 e 1,000 em 2026_1. Esses resultados indicam dependência parcial das soluções em relação ao método de escalonamento nas campanhas de 2025, enquanto a diferenciação do PM-01 em 2026_1 permaneceu inalterada nesse cenário de sensibilidade.

A manutenção dos agrupamentos quando considerados exclusivamente os resultados químicos mostra ainda que os parâmetros físico-químicos, embora contribuam para a caracterização e interpretação dos perfis, especialmente por meio do ORP, não foram isoladamente responsáveis pela estrutura dos agrupamentos.

4.4 Confronto com evidências independentes do gerenciamento da área

Após a identificação dos padrões multivariados, os resultados foram confrontados com informações ambientais que não participaram da formação dos agrupamentos, incluindo histórico de contaminação, ocorrência de LNAPL, comportamento temporal, localização dos

poços e contexto hidrogeológico. Dessa forma, o confronto foi realizado posteriormente à clusterização, preservando a independência entre a identificação estatística dos padrões e sua interpretação ambiental.

O PM-36 apresentou uma das convergências mais consistentes. O poço integrou o grupo diferenciado em 2025_1 e constituiu a única observação diferenciada em 2025_2, campanha na qual apresentou 347,19 µg/L de benzeno, além de outros compostos aromáticos. O histórico ambiental já registrava ocorrências recorrentes de benzeno nesse ponto, indicando convergência entre a diferenciação multivariada e uma condição previamente reconhecida no GAC.

O PM-01 apresentou comportamento igualmente relevante. Na primeira campanha de 2026, foi a única observação diferenciada, com perfil fortemente associado a PAHs. Independentemente da clusterização, o histórico registra presença recorrente de fase livre desde 2020 e, em outubro de 2025, espessura aparente de 0,02 m (20 mm) de LNAPL. O resultado não significa que o algoritmo tenha identificado diretamente a fase livre, mas demonstra convergência entre um perfil químico multivariado singular e uma condição ambiental também diferenciada. O relatório técnico confirma a presença recorrente de fase livre no PM-01.

O PM-16, por outro lado, evidencia a distinção entre anomalia univariada e diferenciação multivariada. Embora possua histórico de benzeno e tenha apresentado concentração superior ao valor orientador considerado no relatório em outubro de 2025, permaneceu no padrão majoritário da análise 2025_2. Assim, a excedência de uma substância isolada não implica necessariamente diferenciação do perfil multivariado em relação ao conjunto da rede.

Outras correspondências apresentaram caráter complementar. PM-18 e WLM-PM-42 foram diferenciados em 2025_1 e possuem registros relacionados a hidrocarbonetos ou PAHs. O WLM-PM-40 também integrou o grupo diferenciado nessa campanha e apresenta histórico relacionado a PAHs e TPH; o ponto não esteve disponível nas campanhas posteriores, tendo sido registrado como não localizado em outubro de 2025. Essas associações devem ser interpretadas com cautela, pois não demonstram persistência do padrão, continuidade de pluma, origem comum ou conexão hidráulica.

Em conjunto, os resultados mostram que os agrupamentos não reproduzem simplesmente uma classificação entre poços contaminados e não contaminados. Eles expressam a posição relativa dos perfis químicos multivariados em cada campanha, podendo apresentar convergência, complementaridade ou discordância em relação às demais informações do GAC. Nesse sentido, sua principal contribuição reside na utilização como linha complementar de evidência, coerentemente com a abordagem de integração posterior de informações independentes discutida por Costanzo et al. (2026).

4.5 Comportamento temporal

A comparação temporal foi restrita aos poços comuns entre campanhas sucessivas. O ARI foi de 0,2735 entre 2025_1 e 2025_2, considerando 20 poços comuns, e igual a zero entre 2025_2 e 2026_1, para 21 poços comuns. Apesar desses valores reduzidos, 85% e 95,2% dos poços comparáveis, respectivamente, permaneceram na mesma classe estrutural, indicando influência das poucas observações diferenciadas e do forte desbalanceamento entre os grupos.

O PM-36 apresentou trajetória “Diferenciado”,“Diferenciado”, “Majoritário”. PM-16, PM-18 e WLM-PM-42 passaram de “Diferenciado” em 2025_1 para “Majoritário” nas duas campanhas seguintes. O PM-01 esteve disponível apenas na primeira campanha de 2026, quando foi classificado como “Diferenciado”. Essas trajetórias representam mudanças de posição relativa no espaço multivariado e não devem ser interpretadas automaticamente como migração, atenuação ou expansão de pluma.

A variação de grupo também deve ser distinguida da tendência de uma substância específica. No PM-36, por exemplo, a série histórica de TPH Total apresentou oscilações entre campanhas, mas o teste de Mann-Kendall disponível na documentação técnica não indicou tendência monotônica significativa. Assim, tendência univariada e diferenciação multivariada representam linhas de evidência distintas.

A influência de condições temporais sobre padrões hidroquímicos também foi observada por Peng et al. (2025). No caso da PDD, os resultados permitem reconhecer mudanças na estrutura química relativa entre campanhas, mas não são suficientes, isoladamente, para atribuir os mecanismos responsáveis por essas alterações.

4.6 Relação espacial e hidrogeológica

A análise espacial, realizada por teste de permutação dos rótulos dos clusters com base na distância média intracluster, resultou em p = 0,860 para 2025_1, p = 0,322 para 2025_2 e p = 0,342 para 2026_1. Como todos os valores foram superiores a 0,05, não houve evidência estatística de compactação espacial global dos agrupamentos.

Esse resultado reafirma a premissa de que análise de agrupamentos não reflete isoladamente informações do mundo real: a similaridade química multivariada e proximidade espacial não são equivalentes. Assim, poços pertencentes ao mesmo agrupamento não podem ser considerados, apenas por essa razão, integrantes da mesma pluma de contaminação ou hidraulicamente conectados através do solo.

A interpretação hidrogeológica deve, portanto, permanecer como linha independente de evidência. A ausência de compactação espacial reforça que os agrupamentos devem ser confrontados com potenciometria, hidroestratigrafia, localização das fontes e histórico ambiental, atuando como informação complementar ao GAC, em linha com a abordagem integrada discutida por Costanzo et al. (2026).

4.7 Limitações

As principais limitações relacionam-se ao reduzido número de poços por campanha, à presença de *singletons* em 2025_2 e 2026_1 e à dependência parcial das soluções em relação ao pré-processamento. Embora a maioria dos testes de sensibilidade tenha reproduzido os agrupamentos principais, o uso do RobustScaler evidenciou dependência do escalonamento, especialmente em 2025_2, indicando influência das escolhas de escalonamento.

O tratamento das não detecções também constitui uma limitação, apesar de a substituição por LQ/2 não ter modificado os agrupamentos. Além disso, a concordância entre HCA e K-means representa consistência algorítmica, mas não validação ambiental independente.

Assim, PCA, HCA e K-means mostraram-se úteis para reduzir a complexidade dos dados e identificar padrões e observações diferenciadas, com destaque para PM-36 e PM-01. Entretanto, seus resultados devem ser interpretados como linha complementar de evidência, em conjunto com o modelo conceitual, a avaliação hidrogeológica e o histórico ambiental da área.

5 Considerações finais e conclusões

Os resultados permitem responder à questão de pesquisa ao demonstrar que a combinação de PCA, HCA e K-means foi capaz de identificar padrões químicos multivariados e observações diferenciadas na rede de monitoramento, acrescentando informação complementar ao gerenciamento de áreas contaminadas. O objetivo do trabalho foi, portanto,

atingido, assim como os objetivos específicos apresentados na metodologia proposta relacionados à redução da dimensionalidade, identificação dos agrupamentos, avaliação de sua consistência, caracterização das variáveis associadas à diferenciação e confronto posterior com informações ambientais independentes.

A aplicação da PCA permitiu sintetizar a estrutura multivariada dos dados, enquanto HCA e K-means recuperaram a mesma estrutura principal de agrupamento nas três campanhas. Foram identificados um conjunto diferenciado de cinco poços em 2025_1 e observações individualmente diferenciadas, representadas pelo PM-36 em 2025_2 e pelo PM-01 em 2026_1. As diferenças estiveram associadas principalmente a combinações de PAHs, BTEX, frações de hidrocarbonetos e parâmetros físico-químicos. Esses resultados são compatíveis com aplicações de técnicas multivariadas e de aprendizado não supervisionado em dados ambientais e de águas subterrâneas descritas por Nguyen et al. (2025) e Panneerselvam et al. (2026).

O confronto com informações independentes mostrou convergências ambientalmente relevantes, especialmente para PM-36, associado ao histórico recorrente de benzeno, e PM-01, diferenciado em uma área com ocorrência histórica de LNAPL. Por outro lado, o comportamento do PM-16 demonstrou que excedências individuais não implicam necessariamente diferenciação multivariada. Esse resultado reforça que os clusters representam similaridades relativas entre perfis químicos, e não uma classificação direta entre poços contaminados e não contaminados. Essa interpretação é coerente com Costanzo et al. (2026), que destacam o maior significado dos agrupamentos quando confrontados posteriormente com linhas independentes de informação ambiental.

As análises de estabilidade e sensibilidade indicaram consistência das soluções em diferentes cenários, embora tenha sido observada dependência do escalonamento, especialmente em 2025_2. As análises temporal e espacial também estabeleceram limites importantes: mudanças entre clusters não podem ser interpretadas automaticamente como migração ou atenuação, e a ausência de compactação espacial significativa mostrou que similaridade química não equivale a pertencimento à mesma pluma ou conexão hidráulica. Essa cautela encontra paralelo em Peng et al. (2025), que evidenciam a influência de fatores temporais, espaciais e hidrogeológicos sobre padrões hidroquímicos.

Conclui-se, assim, que o aprendizado de máquina não supervisionado pode constituir uma linha complementar de evidência no GAC, especialmente para sintetizar bases multivariadas complexas, reconhecer padrões relativos e direcionar a atenção para observações que demandem investigação adicional. Sua aplicação, entretanto, deve permanecer integrada ao modelo conceitual, à avaliação hidrogeológica, ao histórico ambiental e à interpretação convencional dos resultados analíticos, não constituindo ferramenta autossuficiente para determinação de fontes, plumas ou conexões hidráulicas.

6 Referências

CETESB. (2024). Manual de gerenciamento de áreas contaminadas (3ª ed. rev. e ampl.). Companhia Ambiental do Estado de São Paulo.

Costanzo, V. L., Izarra-Jacome, A., Costanzo-Alvarez, V., Jacome, M., Trigo-Ferre, R., Jeffrey, M. A., Cheng, Y., Aldana, M., & Amon, C. H. (2026). Machine-learning-derived hydrogeochemical facies to support geoelectrical interpretation of contaminant plumes. Environmental Processes, 13, Article 3.

Fávero, L. P., & Belfiore, P. (2025). Manual de análise de dados: Estatística e machine learning com Excel®, SPSS®, Stata®, R® e Python® (2ª ed.). LTC.

Guan, Y., Lu, H., Dong, J., Ge, Y., Zhang, W., & Deng, Y. (2025). A novel groundwater pollution risk assessment method for producing-enterprises sites: Integrating scenario-specific pollution evaluation with Gaussian mixture model clustering. Process Safety and Environmental Protection, 195, 106760.

Hirata, R., Suhogusoff, A. V., Marcellini, S. S., Villar, P. C., & Marcellini, L. (2019). A revolução silenciosa das águas subterrâneas no Brasil: Uma análise da importância do recurso e os riscos pela falta de saneamento. Instituto Trata Brasil.

International Organization for Standardization. (2017). ISO/IEC 17025:2017: General requirements for the competence of testing and calibration laboratories (3rd ed.).

Mohammed, M. A. A., Szabó, N. P., Mikita, V., & Szűcs, P. (2025). Tracking the spatiotemporal evolution of groundwater chemistry in the Quaternary aquifer system of Debrecen area, Hungary: Integration of classical and unsupervised learning methods. Environmental Science and Pollution Research, 32, 6884–6903.

Nguyen, D., Cadle, S., Verstraet, T., Kammer, L., Van Scoy, T., Anding, M., & Anderson, R. (2025). Novel insights from multivariate analysis of metadata from a large PFAS remedial investigation dataset. Journal of Hazardous Materials, 500, 140539.

Panneerselvam, M., Govindan, V., & Sakthivel, L. P. (2026). A novel framework for groundwater quality evaluation in industrial zone using unsupervised machine learning methods. Environmental Geochemistry and Health, 48, Article 161.

Peng, J.-Y., Zhang, S., Wang, Y., Lan, J., Wang, S., Zhou, Y., Wu, L., & Zhan, L. (2025). Time-space patterns of groundwater contamination in a closed MSW landfill: Insights from long-term monitoring and SOM-based clustering analysis. Environmental Pollution, 382, 126659.

Zhang, Y., Zhang, Q., Chen, W., Shi, W., Cui, Y., Chen, L., & Shao, J. (2023). Hydrogeochemical analysis and groundwater pollution source identification based on self-organizing map at a contaminated site. Journal of Hydrology, 616, 128839.