Universidade Estadual de Campinas (Unicamp)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Departamento de Engenharia de Computação e Automação (DCA)

Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Relatório Parcial de Atividades
Edital 2022/2023

Mapeamento de zonas homólogas com base em dados hidroquímicos de um aquífero freático sujeito a contaminação por fossas em área peri-urbana de Campinas utilizando algoritmos de aprendizado de máquina

Aluno
DIEGO XAVIER MACHADO
Graduando do curso de Engenharia de Controle e Automação
Faculdade de Engenharia Mecânica (FEM)
d250258@dac.unicamp.br

Orientadora
PROFA. DRA. PAULA DORNHOFER PARO COSTA
Depto. de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
paulad@unicamp.br

Coorientadora
PROFA. DRA. ANA ELISA SILVA DE ABREU
Departamento de Geologia e Recursos Naturais (DGRN)
Instituto de Geociências (IG)
aeabreu@unicamp.br

5 de março de 2024

Sumário

1 Apresentação

O presente documento caracteriza o relatório final de atividades do Programa Institucional de Bolsas de Iniciação Científica (PIBIC), edital 2022/2023, e apresenta um resumo das atividades realizadas no período de setembro de 2022 a agosto de 2023.

2 Introdução

As águas subterrâneas representam a maior parte da água disponível para o consumo humano e, no Estado de São Paulo, o uso dessa fonte vem crescendo substancialmente nos últimos anos [Iritani e Ezaki 2014]. Mediante a presente situação, mostra-se fundamental o monitoramento da qualidade das águas subterrâneas para preservá-las contra possíveis contaminações e evitar situações de risco à saúde da população que usa dessas águas.

A classificação da qualidade da água e de suas naturezas hidroquímicas são, tipicamente, realizadas por especialistas, capazes de interpretar não apenas os resultados de análise químicas de amostras de água mas também de correlacioná-las com as características hidrogeológicas e ambientais da região onde a amostra foi coletada. Porém, essa abordagem apresenta limitações para o estudo de sistemas subterrâneos complexos ou estudos em grande escala, regionais ou nacionais. Por esse motivo, modelos de predição, construídos a partir de algoritmos de aprendizagem de máquina, têm sido adotados como ferramentas auxiliares para a predição da qualidade d'água ao longo de grandes regiões com características geológicas diversificadas. É o caso, por exemplo, do trabalho de Friedel et al. 2020. Nesse trabalho, os autores utilizaram dados oriundos de bases de dados de clima, geologia, hidrologia, ocupação de terra, solo e topografia da Nova Zelândia para a construção de modelos de predição do estado redox de águas subterrâneas em diferentes localizações geográficas, distantes entre si. Conhecer o local onde ocorre a desnitrificação das águas subterrâneas, que tem como variável "*proxy*" o estado redox das águas subterrâneas, possibilita monitorar as fontes de lixiviação agrícola e, por consequência, a qualidade da água nessas regiões.

Nesse contexto, o presente trabalho teve como objetivo explorar a intersecção entre o domínio especialista de hidrogeologia e técnicas clássicas de análise exploratória de dados, caracterizando-se como um trabalho na área de Ciência de Dados.

Em particular, o trabalho foi motivado pela disponibilidade de dados de análises hidroquímicas de amostras de águas subterrâneas coletadas por pesquisadores do Instituto de Geociências (IG) da UNICAMP, na região do Piracambaia II, localizada ao norte do município de Campinas (SP).

Esses dados foram objeto de estudo para o trabalho de Alencar 2021, no qual buscou-se avaliar as características sanitárias dos poços domésticos e a influência das condições de óxido-redução na qualidade das águas na região do Piracambaia II. O local de realização do estudo é caracterizado como uma região periurbana do município de Campinas, no qual o despejo de efluentes domésticos ocorre

através de saneamento *in situ*, que pode acarretar na contaminação das águas subterrâneas da região e causar problemas de saúde (vide Figura 1). . Os dados utilizado no presente trabalho são oriundos de duas campanhas de coleta do trabalho de Alencar 2021, a primeira no mês de abril de 2019 (classificado como um período de chuvas) e a segunda no mês de agosto de 2019 (caracterizado como um período de seca). Para cada campanha, foram coletadas 20 amostras, sendo 18 de águas subterrâneas de poços domésticos e duas amostras de águas superficiais. Em cada ponto foram coletados cinco alíquotas, que foram enviadas para análise laboratorial nos seguintes laboratórios: Laboratório de Geologia Isotópica da UNICAMP (análise de cátions), Laboratório de Hidrogeologia e Hidrogeoquímica da UNESP (análise de ânions), Laboratório de Geologia Analítica da UNICAMP (análise de Carbono Orgânico Dissolvido e alcalinidade), Laboratório de Análises Isotópicas da UNESP (análise de isótopos estáveis).

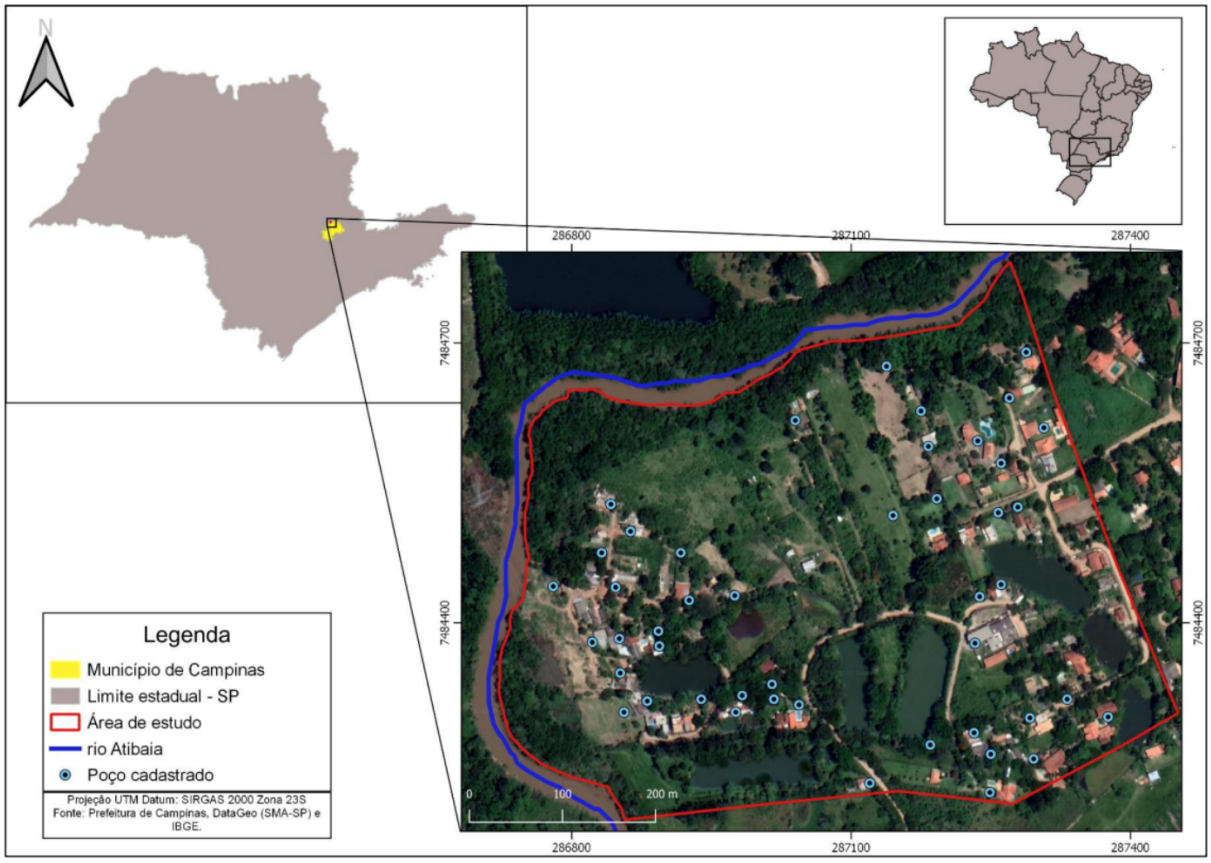

Figura 1: Mapa da região de estudo, Piracambaia II, Localizada ao norte do município de Campinas. Fonte: [Alencar 2021]

Portanto, o presente projeto tem como objetivo usar algoritmos de aprendizado de máquina não-supervisionado de agrupamento para realizar o mapeamento das zonas homólogas com base nos dados hidroquímicos da região estudada, ou seja, busca-se mapear a região de estudo com base nos aspectos hidroquímicos semelhantes entre as amostras coletadas a partir dos agrupamentos feitos pelos algoritmos utilizados.

3 Metodologia

A metodologia adotada pelo projeto inspira-se no *framework* KDD (*Knowledge Discovery in Databases*) descrito por Fayyad et al. 1996. A Figura 2 ilustra o *framework*, que é constituído por etapas de entendimento do domínio da aplicação e do conhecimento já existentes, da escolha do banco de dados disponíveis e pré-processamento de seus dados, da aplicação de algoritmos de *data mining*, da interpretação de padrões por meio de ferramentas estatísticas e, finalmente, da consolidação do conhecimento, visando a condução de novos ciclos de descoberta de conhecimento.

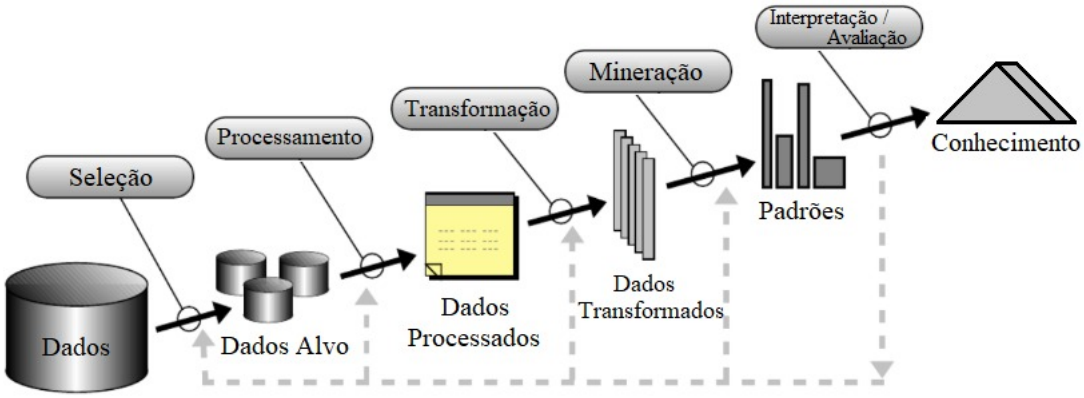

Figura 2: Processo de análise e extração de conhecimento modelado por Fayyad (1996). Fonte: adaptado de [Fayyad et al. 1996]

Partindo-se dessa referência, o trabalho organizou-se nas seguintes etapas:

- Estudo, acesso e seleção de dados de águas subterrâneas disponibilizados pelo Instituto de Geociências (IG), processo descrito na Seção 4;
- Processamento e transformação dos dados extraídos de planilhas e relatórios técnicos para a construção de um *dataset* de estudo, como descrito na Seções 5.1 e 5.2;
- Análise exploratória dos dados por meio de algoritmos de agrupamento não-supervisionado aplicados ao *dataset* de estudo (Seção 5.3);
- Estudo e seleção de métricas de avaliação adequadas ao trabalho para agrupamentos não supervisionados (Seção 5.4);
- Análise e interpretação dos resultados, apresentadas na Seção 6.

4 Aquisição dos dados

A primeira etapa do presente trabalho consistiu na construção de uma base de dados de estudo devidamente formatada para posterior análise exploratória computacional.

Os dados foram extraídos de documentos utilizados e gerados pelo trabalho de Alencar 2021, cedidos pelas autoras do trabalho para a condução deste projeto de iniciação científica.

Originalmente, a maior parte dos dados foi disponibilizada por meio de arquivos com extensão proprietária (*.pdf* e *.docx*), de difícil processamento computacional automatizado, levando à necessidade de serem transcritos manualmente para uma base de dados que concatenou todas as entradas de dados em um só arquivo, posteriormente salva no formato aberto de registro de informações "Comma-separated values" (extensão de arquivo *.csv*). Outros dados, no formato de planilhas (arquivos Excel, com extensão *.xlsx*), puderam ser mais facilmente acessados por rotinas computacionais automatizadas.

No contexto deste trabalho, as rotinas computacionais foram desenvolvidas na linguagem de programação *Python*. Em particular, a biblioteca Python chamada *pandas* foi adotada para as operações de manipulação de dados em planilhas.

Os dados que compõem a base de dados de estudo correspondem a amostras de água coletadas em dois períodos sazonais distintos: um período de chuvas (abril de 2019) e um período de seca (agosto de 2019). Embora o conjunto de dados original contivesse 20 observações para cada período, ao analisar-se todos os resultados laboratoriais, observou-se que a amostra de abril de 2019, referente ao poço identificado como "w45", continha dados faltantes. Por meio de discussões realizadas com a coorientadora deste trabalho, a especialista em hidrogeologia Profa. Dra. Ana Elisa Silva de Abreu, decidiu-se que tal amostra poderia ser descartada. Assim, foram mantidos os dados de 19 amostras referentes ao período de abril de 2019 e 20 amostras associadas ao período de agosto de 2019. Tal informação é relevante pela intenção apresentada de se realizar análises individualizadas para os períodos de chuva e seca.

Ao final da etapa de concatenação informações oriundas de diferentes documentos e planilhas, a base de dados estudos resultante é caracterizada por uma planilha com um total de 39 linhas e 49 colunas, onde:

- A primeira coluna (coluna 0) identifica o ponto de coleta da amostra de água;
- A coluna 1 contém a data da coleta das amostras;
- As colunas 2 até 33 contêm as variáveis dos cátions de cada amostra coletada;
- As colunas 34 até 38 contêm as variáveis dos ânions de cada amostra coletada;
- A coluna 39 contêm o Carbono Orgânico Dissolvido em cada amostra coletada;
- As colunas 40 até 43 contêm as variáveis dos aspectos físico-químicos das amostras coletadas;
- As colunas 44 até 46 contêm as variáveis da análise isotópica das amostras coletadas;
- As colunas 47 e 48 contêm as coordenadas geográficas (no formato UTM) de cada ponto de coleta das águas.

Ademais, as amostras foram nomeadas como nos documentos utilizados para formar a base de dados. Os poços domésticos foram nomeados como "WX´´, onde X é a enumeração dada pelos pesquisadores que fizeram as coletas. O ponto de coleta das águas superficiais do rio foi chamada de "*river´´*. Por fim, o ponto de coleta das águas superficiais do meandro foi nomeado de "*meander´´*.

O dataset criado foi depositado no Repositório de Dados da Unicamp e está disponível publicamente [Alencar et al. 2023].

5 Análise Exploratória de Dados

Na área popularizada como Ciência de Dados, a etapa de análise exploratória (ou EDA, do inglês, *Exploratory Data Analysis*) é caracterizada pela condução de uma série de análises estatísticas, a fim de proporcionar uma maior compreensão do comportamento e das relações dos dados estudados, além de possibilitar a detecção de anomalias (*outliers*) e o teste de hipóteses.

No contexto deste trabalho, a análise exploratória de dados se dividiu em três etapas diferentes.

Na primeira etapa, realizou-se um levantamento de estatísticas descritivas das variáveis hidroquímicas da base de estudos e, juntamente com a especialista em hidrogeologia, avaliou-se a qualidade dos dados, bem como propôs-se recortes da base de estudos original, para a etapa seguinte da análise exploratória. Tais operações estão descritas na Seção 5.1.

Na segunda etapa, realizou-se uma análise de de variáveis altamente correlacionadas. Essa análise possibilitou avaliar relações esperadas advindas do conhecimento especialista de hidrogeologia. A partir dessa análise, apresentada na Seção 5.2, foi possível também sugerir simplificações dos conjuntos de dados estudados.

A terceira etapa da análise exploratória, consistiu num estudo de agrupamentos, como descrito na Seção 5.3.

Como é típico da análise exploratória, as etapas foram revisitadas e repetidas de acordo com cada resultado obtido e discutido à luz do conhecimento especialista. A descrição das etapas, como descrito neste relatório, representam o ciclo final de análise.

5.1 Análise da Qualidade dos Dados

Como primeira abordagem exploratória na base de dados de estudo, foram levantadas as estatísticas descritivas de suas variáveis hidroquímicas.

A Figura 3 mostra, por exemplo, os boxplots obtidos para os elementos químicos K (potássio), Ge (germânio) e Cd (cádmio). Nos gráficos, as linhas tracejadas em verde e vermelho representam, respectivamente, os valores de Limite de Detecção (LD) e Limite de Quantificação (LQ).

O LD indica a menor concentração de um analito que pode ser detectada (mas não necessariamente

quantificada) e é uma característica do processo ou instrumento utilizado para mensurar o analito. O LQ, por sua vez, corresponde à menor concentração que pode ser quantificada com precisão e exatidão aceitáveis. A literatura esclarece que é possível definir uma relação matemática entre o limites de detecção e de quantificação, no qual $LQ = 3.3 \times LD$ [Commission et al. 2016]. No presente trabalho, adotou-se, como estratégia conservadora, $LQ = 4 \times LD$.

A Figura 3 mostra, como exemplo, que a média das concentrações encontradas para o cádmio é menor que o valor estabelecido para LQ desse analito. O mesmo fenômeno foi observado para outras variáveis da base. Em outras palavras, em média, alguns elementos químicos da base foram encontrados em baixas concentrações nas amostras analisadas. Apesar de detectáveis, tais concentrações encontram-se em regiões de medição pouco precisas dos instrumentos/processos utilizados para mensurá-los.

Por este motivo, decidiu-se remover das análises posteriores de processamento as variáveis cujos valores de *mediana* encontravam-se abaixo do valor de LQ estabelecido para a variável. As variáveis removidas nessa etapa foram: Berílio (Be), Escândio (Sc), Titânio (Ti), Selênio (Se), Zircônio (Zr), Nióbio (Nb), Prata (Ag), Cádmio (Cd), Estanho (Sn), Háfnio (Hf), Tântalo (Ta), Tungstênio (W), Bismuto (Bi), Tório (Th), Amônia (NH_4) e Dióxido de Nitrogênio (NO_2).

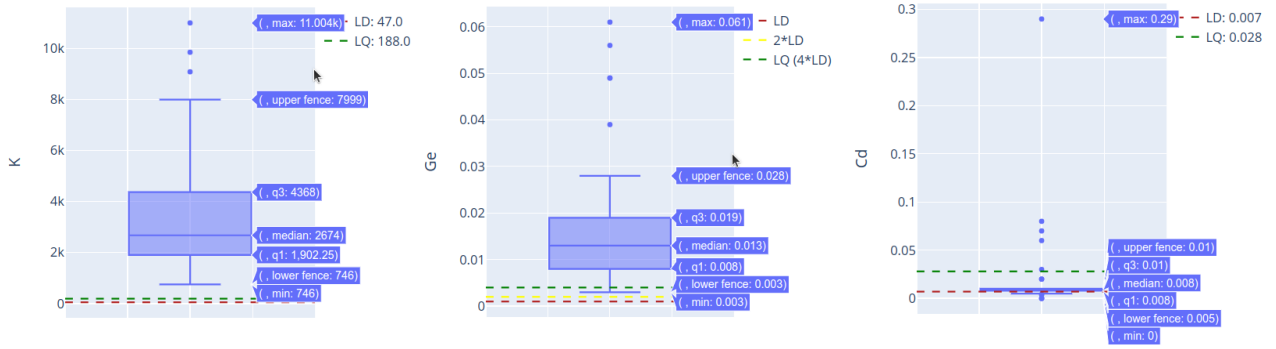

Figura 3: Estatísticas descritivas de três variáveis da base: potássio (K), germânio (Ge) e cádmio (Cd). O potássio ilustra um exemplo de uma distribuição de medidas que encontra-se acima dos limites de detecção e quantificação. Por outro lado, germânio e cádmio apresentam alguns valores de medição abaixo do limite de quantificação, sendo o caso mais grave o caso do cádmio, no qual a mediana das medidas está abaixo do valor de LQ.

5.2 Análise de Variáveis Altamente Correlacionadas

Após a retirada da base de variáveis segundo critérios de qualidade de medição, realizou-se uma análise de correlação entre as variáveis a fim de verificar-se hipóteses oriundas do conhecimento hidrogeológico, bem como eventualmente encontrar-se fenômenos ainda não aparentes.

A análise de correlação foi feita utilizando-se a função *dataframe.corr()* da biblioteca *pandas* do Python, calculando-se, então, a correlação de Pearson para todos os pares de variáveis.

A matriz de correlação resultante pode ser visualizada por meio de um mapa de calor (*heatmap*) como apresentado na Figura 4. Como esperado, uma das características visuais marcantes do *heatmap*

da matriz de correlação é a linha diagonal com altos valores de correlação positiva (representados pela cor azul marinho), por se referir à correlação de cada variável com ela mesma. Ainda, é característica da matriz de correlação sua simetria em relação à diagonal principal. Finalmente, a análise gráfica da imagem gerada permite observar que existe um conjunto específico de elementos químicos que apresentam alta correlação positiva entre si (faixa azul marinha no centro na região central do gráfico). Observa-se que essa região corresponde, essencialmente, aos metais de terras raras.

A partir deste achado, optou-se por uma redução de dimensionalidade da base realizada com base no entendimento de que uma variável A, altamente correlacionada com uma variável B, pode ser considerada uma variável *proxy* da variável B.

Uma análise conduzida pela especialista, resultou na sugestão de retirada da base dos seguintes elementos químicos: Lantânio (La), Cério (Ce), Praseodímio (Pr), Samário (Sm), Európio (Eu), Térbio (Tb), Disprósio (Dy), Hólmio (Ho), Túlio (Tm), Itérbio (Yb) e Lutécio (Lu). Nesta etapa também foi removida a variável Sólidos Dissolvidos Totais (TDS), pois essa variável não é medida por instrumentos, e sim calculada a partir da variável EC.

Um novo *heatmap* foi gerado para as 47 variáveis numéricas restantes, como mostrado na figura 5. As variáveis que permaneceram para a base de dados final foram: Lítio (Li), Boro (B), Sódio (Na), Magnésio (Mg), Alumínio (Al), Silício (Si), Potássio (K), Cálcio (Ca), Vanádio (V), Cromo (Cr), Manganês (Mn), Ferro (Fe), Cobalto (Co), Níquel (Ni), Cobre (Cu), Zinco (Zn), Gálio (Ga), Germânio (Ge), Arsênio (As), Rubídio (Rb), Estrôncio (Sr), Ítrio (Y), Molibdênio (Mo), Antimônio (Sb), Césio (Cs), Bário (Ba), Neodímio (Nd), Gadolínio (Gd), Érbio (Er), Tálio (Tl), Chumbo (Pb), Urânio (U), Flúor (F), Cloro (Cl), Nitrato (NO_3), Sulfato (SO_4), Fosfato (PO_4), Matéria Orgânica Dissolvida (DOC), Temperatura, Condutividade Elétrica (EC), Potencial Redox (Eh), pH, delta-2-H (d2H), delta-18-O (d18O), Excesso de Deutério (d_excess), Coordenadas UTM Leste(UTM_E) e Coordenadas UTM Norte (UTM_N).

Assim, finalizou-se a etapa de tratamento da base de dados com 39 amostras e 47 variáveis numéricas.

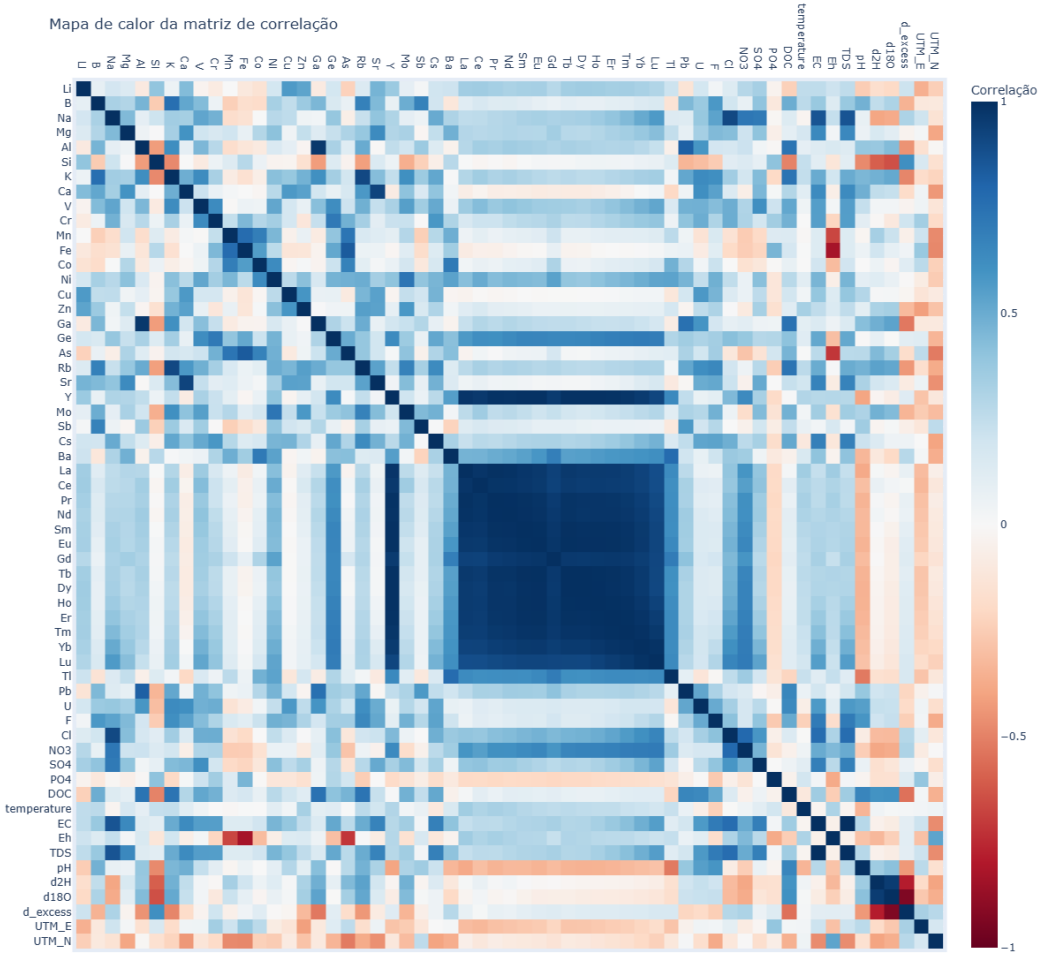

Figura 4: *Heatmap* com 59 variáveis numéricas. É possível observar uma região com uma grande correlação direta entre as variáveis da base de dados. Essas variáveis são referentes aos elementos químicos denominados de “metais de terras raras”.

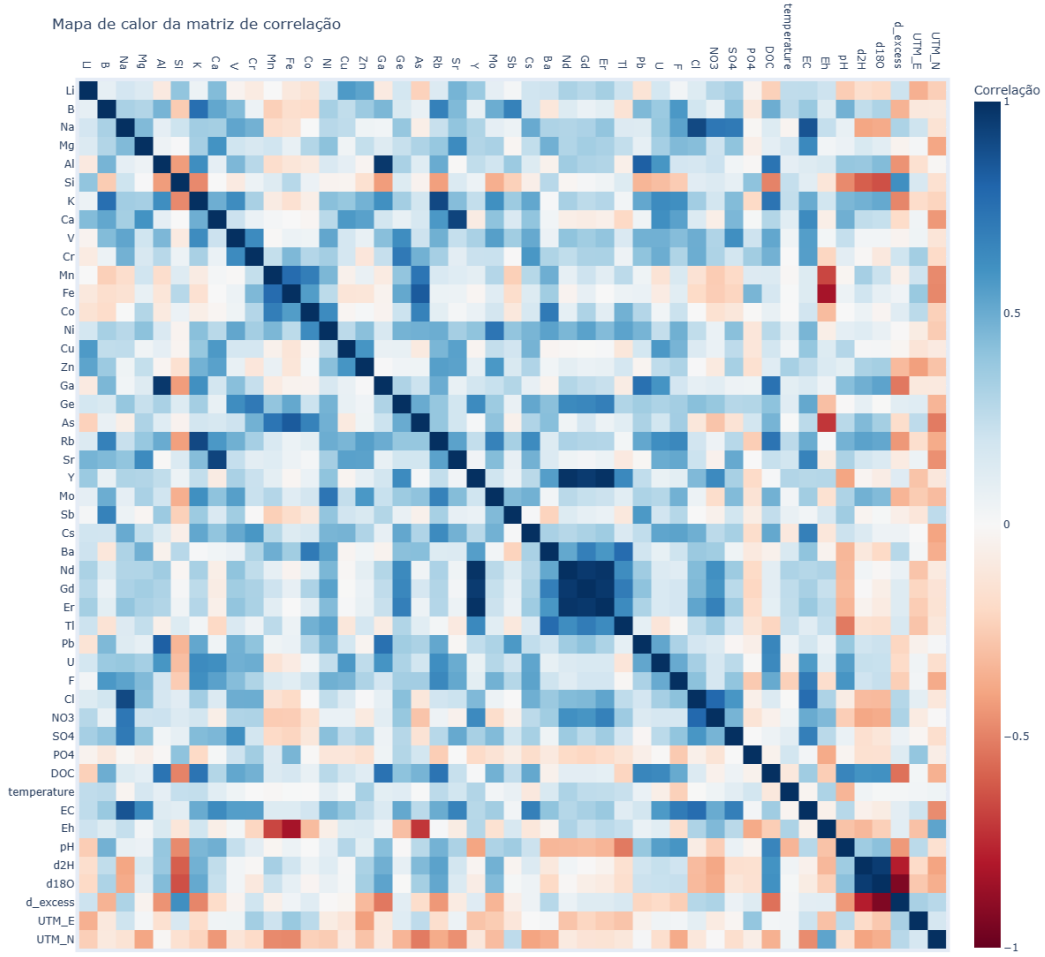

Figura 5: *Heatmap* com 47 variáveis numéricas. Nota-se que o número de variáveis com uma alta correlação entre si teve uma redução considerável.

5.3 Análise de Agrupamentos

Os algoritmos de agrupamento, tipicamente referenciados como algoritmos de aprendizado de máquina não-supervisionados, são usados na etapa de análise exploratória para identificar padrões, nem sempre aparentes. Nesse contexto, o termo não-supervisionado se refere ao fato dos algoritmos de clusterização (do inglês, *clustering algorithms*) não levarem em consideração nenhuma categorização prévia dos dados, deixando a cargo dos modelos implementados a tarefa de identificação de agrupamentos emergentes.

Existem diferentes abordagens estatísticas para a busca por agrupamentos em um conjunto de dados n-dimensional [Xu e Tian 2015]. No contexto deste trabalho, aplicou-se duas abordagens diferentes:

- **Clusterização baseado em partições:** nessa abordagem, os algoritmos clusterização dividem o dataset em grupos de dados mutuamente exclusivos. Nesses algoritmos, o objetivo é agrupar amostras de dados similares, ao mesmo tempo que se busca maximizar as diferenças entre os grupos encontrados. Neste trabalho, aplicou-se o clássico algoritmo *K-means* (vide Seção

5.3.1).

- **Clusterização baseada em modelos:** nessa abordagem, parte-se de uma hipótese de um modelo específico para cada *cluster* e busca-se encontrar os parâmetros do modelo para cada grupo, a partir dos dados. Em particular, neste trabalho adotou-se a premissa de um modelo de rede neural de mapas auto-organizáveis, em inglês, *Self-Organizing Maps* (SOM), que assume a existência de uma topologia dos dados no espaço n-dimensional (vide Seção 5.3.2).
- **Clusterização baseada em densidade:** nessa abordagem, busca-se o agrupamento a partir do agrupamento de pontos que mantém uma vizinhança próxima no espaço amostral e evitar agrupamentos de pontos muito distantes, classificando-os como *outliers*. Nesse projeto em particular, adotou-se o algoritmo DBSCAN (vide Seção 5.3.3).

Um aspecto crucial para a análise de resultados de algoritmos de clusterização é a avaliação dos agrupamentos gerados por cada abordagem. Assim, na Seção 5.4 são apresentadas as métricas adotadas no contexto deste trabalho.

5.3.1 Algoritmo K-Means

Para o projeto, foi necessário fazer um estudo sobre algoritmos de aprendizado de máquina não-supervisionados para agrupamento de dados, a começar pelo algoritmo *K-Means*. O *K-Means* é um dos mais simples algoritmos de aprendizado de máquina não-supervisionado para agrupamento de dados. Proposto por MacQueen et al. 1967, a premissa do algoritmo é:

1. Tendo um número n de observações, é distribuído um número $K(\leq$n$)$ de centroides no espaço $\mathbb{R}^n$. Algumas bibliotecas em *Python* distribuem os centroides de forma aleatória sobre o espaço.
2. Para cada dado observado (ponto no espaço), é calculado a distância euclidiana para os K centroides. O dado é ligado ao centroide mais próximo.
3. Feito os agrupamentos dos dados com os centroides mais próximos, é calculado a média da posição de todos os pontos de cada grupo formado. A coordenada do ponto médio de cada grupo é a nova posição do centroide do grupo.
4. Novamente são calculadas as distâncias dos dados observados para os centroides, cada observação é novamente ligada ao centroide mais próximo.
5. O processo se repete até que não haja mais mudanças nas posição dos centroides, ou seja, na convergência dos centroides.

Uma abordagem típica de processamento dos dados antes da aplicação do algoritmo de *clusterização* é o processo de padronização dos dados para que todas as variáveis numéricas sejam expressas na mesma faixa e valores. Neste trabalho, adotou-se uma padronização dos valores utilizando-se o objeto `StandardScaler` da biblioteca `sklearn.preprocessing`.

No presente projeto, o algoritmo *K-Means* utilizado a partir da biblioteca *scikit-learning*. Nessa biblioteca, por padrão, os centroides são distribuídos no espaço em posições aleatórias, caracterizando uma natureza não determinística para o algoritmo. Em resumo, a cada nova execução do algoritmo, pode se chegar em agrupamentos distintos das excussões anteriores, para uma mesma base de dados, como pode se observar na figura 6.

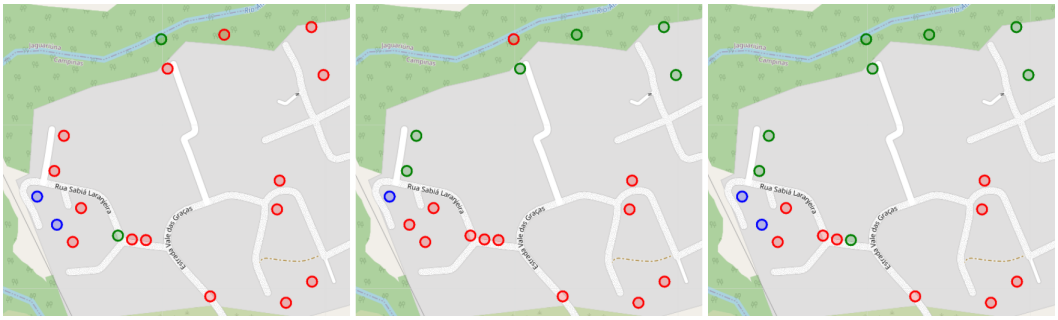

Figura 6: Demonstração da natureza não determinística do algoritmo K-Means. Pode-se observar que para um mesmo número de grupos (3), os agrupamentos possuem alguns elementos distintos a cada execução.

Como o algoritmo resultava em agrupamentos diferentes a cada execução, adotou-se a seguinte estratégia para escolher qual agrupamento seria utilizado para prosseguir com o experimento.

1. A princípio, foi escolhido um intervalo para entre os números mínimo e máximo de grupos a serem formados. Como número mínimo, foi definido dois grupos por ser o menor número possível para formar grupos distintos. Empiricamente, após sucessivos experimentos, definiu-se k=6 como o número máximo de grupos a serem formados.
2. Para cada número de grupo, o algoritmo *K-Means* é executado 20.000 vezes e os rótulos são salvos em uma lista;
3. Na lista formada, calcula-se a frequência dos rótulos formados e o agrupamento mais frequente é escolhido para representar os agrupamentos feitos para aquele número de grupos.
4. O processo é repetido para todos os demais números de grupos.

Essa abordagem para a escolha de grupos torna-se relevante pois, como se trata de um aplicação de aprendizado de máquina não supervisionado, não se conhece, a priori, o número de grupos que melhor ressalta aspectos de interesse dos dados.

5.3.2 Algoritmo Self-Organizing Maps (SOM)

O *Self-Organizing Maps*, também chamado de *SOM* ou Mapa de Konohen, é um algoritmo de aprendizado de máquina não-supervisionado que tem como objetivo a redução de dimensionalidade dos dados e o agrupamento dos mesmos conforme suas relações entre si, criando uma representação espacial efetiva e organizada dos dados [Kohonen 1990].

Esse algoritmo é um tipo de rede neural artificial que utiliza o aprendizado competitivo, isto é, os neurônios competem entre si para que sejam representantes de um subconjunto dos dados de entrada. O SOM leva a topologia dos dados em consideração, uma vez que seus neurônios estão em uma espécie de grade (de duas dimensões, normalmente retangular ou hexagonal) onde cada neurônio é influenciado pelo neurônio vizinho.

Esse algoritmo foi utilizado nos trabalho de [Friedel et al. 2020] para a classificação da natureza hidrogeoquímica de águas subterrâneas, classificando os dados como oxidante ou redutor.

Para a execução do algoritmo, seguem as seguintes etapas:

1. De forma aleatória, são definidos os pesos de todos os neurônios da grade.
2. Para cada dado da base, os neurônios competem entre si e o neurônio vencedor (BMU, do inglês *Best Match Unit*) escolhido para representar aquele dado é aquele cuja a distancia entre entre o conjunto de características do dado e o conjunto de pesos do neurônio é a menor dentre todos os neurônios.
3. O peso do BMU e dos outros neurônios próximos a ele são atualizados (utilizando a distribuição normal ou de *t-student*) de forma a se aproximar do dado de entrada.
4. O processo se repete até que os neurônios cheguem na convergência de seus resultados ou até que o número de iterações máximo seja alcançado.

Na execução do SOM, a biblioteca utilizada [Aguiar 2023] possui uma diferença entre a aplicação padrão do algoritmo. No algoritmo utilizado, a grade bidimensional inicial onde os neurônios são colocados é determinística e sempre quadrada, ela cobre toda amplitude dos dados sendo rotacionada 45 graus para se ajeitar em todas as N dimensões, e a função de atração do neurônio vencedor a um dado é discreta. Na prática isso significa que os neurônios vizinhos são atraídos em velocidades constantes.

5.3.3 Algoritmo DBSCAN

Do inglês *Density-based spatial clustering of applications with noise*, o DBSCAN é um algoritmo de aprendizado de máquina não-supervisionado proposto por [Ester et al. 1996] no qual tem como premissa agrupar pontos que estão bem próximos entre si (região de alta densidade) e classificar como *outlier* os pontos que estão sozinhos no espaço, ou em uma região de baixa densidade.

Além da detecção de *outliers*, esse algoritmo conta com as vantagens de não precisar especificar o número de grupos buscado e também conseguir fazer agrupamentos em padrões não lineares.

Os únicos parâmetros necessários para a execução do algoritmo é o ε e o número mínimo de pontos para formar a região de densidade (minPts). O desenvolvimento do algoritmo segue as seguintes etapas:

1. É escolhido um ponto arbitrário, dentre os diversos pontos do banco de dados no qual ainda não fora escolhido anteriormente.

2. É buscado por pontos vizinhos que se distanciam por um raio ε do ponto inicial escolhido. Caso dentro desse raio haja um número de vizinhos inferior ao minPts, esse ponto inicial escolhi é considerado um ruído, não há formação de grupos, um novo ponto arbitrário é escolhido e as etapas se repetem. Caso contrário, um agrupamento é iniciado e esse ponto é considerado um grupo.

3. Com o início de um grupo, todos os vizinhos em um raio ε também fazem parte desse grupo. Além disso, todos os pontos desse grupo também buscarão por vizinhos em suas proximidades de raio ε. Se houver novos pontos encontrados, eles são adicionados ao grupo e o processo de buscar por novos vizinhos em um raio ε se repete até que não haja mais vizinhos para serem adicionados ao grupo. Assim, é formado uma região de densidade.

4. Outro ponto, que não faz parte de nenhum de nenhum grupo formado e também não é um ruído é escolhido e as etapas se repetem.

5.4 Métricas

Com a aplicação dos algoritmos de aprendizado de máquina nas bases de dados e com os respectivos grupos formados, é necessário buscar uma forma de avaliar, de forma numérica, a qualidade dos agrupamentos de acordo com o número de grupos para cada algoritmo utilizado sobre uma mesma base de dados.

Para avaliar os agrupamentos feitos, tanto como melhor número de grupos quanto a qualidade dos agrupamentos formados.

As métricas utilizadas nos algoritmos de agrupamento foram *Davies–Bouldin index* [Davies e Bouldin 1979], *Calinski-Harabasz index* [Caliński e Harabasz 1974] e *Silhouette Analisys* [Rousseeuw 1987].

A *Silhouette Analisys*, além de entregar o gráfico da silhueta dos dados que representa o quão próximo um ponto é de seu próprio grupo (coesão) em comparação com os outros grupos (separação), essa métrica entrega a pontuação, no intervalo $(-1, 1)$, dos resultados obtidos. Uma boa pontuação, no qual é o maior valor de retorno da métrica, significa que o método de agrupamento é capaz de separar bem as bordas dos grupos e ter agrupamentos bem concentrados, ou seja, as duas observações mais distantes do mesmo grupo tem que estar mais próxima do que as observações mais próximas entre uma observação desse grupo e de outro grupo.

O *Calinski-Harabasz index* tem uma premissa parecida com a pontuação da *Silhouette Analisys* sobre os conceitos de coesão e separação. Entretanto, nessa métrica é avaliada utilizando a variância dos dados em seus cálculos para obter os resultados da métrica. O *Calinski-Harabasz index* tem o intervalo definido em $(0, \infty)$ e quanto maior o valor, melhor é o agrupamento.

Diferente das métricas anteriores, o *Davies–Bouldin index*, com intervalo de resultados em $(0, \infty)$, o menor valor obtido nessa métrica representa um melhor agrupamento. Essa métrica é calculada como a média dos valores da razão entre a distância *intra-cluster* do ponto mais distante de um grupo e a distância *inter-cluster* entre o mesmo grupo e o grupo mais próximo.

5.5 Impressão dos Mapas

Após a execução dos algoritmos de agrupamento, foram feitos diversos mapas marcando as posições geográfica dos pontos de coleta das águas, no qual cada ponto foi destacado com uma cor referente ao rótulo de seu grupo definido. Como será melhor explicado posteriormente na seção 6, os agrupamentos não foram muitos consistentes entre si nos diversos experimentos feitos com os algoritmos.

Os resultados obtidos nos agrupamentos foram impressos em forma de mapa, utilizando-se a biblioteca *Python* chamada *folium*, para que, com um auxílio de a coorientadora especialista em hidrogeologia, fosse possível avaliar visualmente se algum resultado obtido utilizando os algoritmos de aprendizado de máquina não-supervisionado chegou a algum resultado de interesse hidrogeológico.

6 Resultados

Como primeiro resultado, foi finalizada uma base de dados unificada, formatada e filtrada para os dados coletados na região do Piracambaia II pelo IG. Essa base de dados pode servir como documento auxiliar tanto para a continuação do trabalho de Alencar 2021 quanto para novos estudos hidrogeoquímicos feitos na região. Além disso, tem-se um arquivo escrito em *Python* que, com uma base de dados não filtrada, mas em um formato semelhante a base de dados utilizada como entrada, é capaz de gerar a estatística descitiva dos dados presentes, gerar o *heatmap* de correlação entre as variáveis e também filtrar e tratar os dados para uma forma mais simplificada e ignorando valores ruidosos.

Na execução dos demais algoritmos utilizados e de suas métricas, é possível notar que houveram vários resultados distintos. Esses resultados distinguem entre si ao comparar tanto o período de coleta das águas subterrâneas, quanto o algoritmo utilizado e também ao sob a influência das coordenadas geográficas como variáveis. Para cada um desses casos, serão apresentados os seus respectivos resultados posteriormente.

6.1 Análise dos agrupamentos em períodos diferentes

Como dito anteriormente, a base de dados é dividida em duas, referente aos dois diferentes períodos de coleta das amostras de águas subterrâneas (período de chuva e período de seca). Ao verificar os resultados dos algoritmos, é notável que as métricas não concordaram entre si nem para o mesmo período e nem para o período diferente. Os agrupamentos para ambos períodos obtiveram resultados

totalmente diferentes entre si para o algoritmo SOM (figura 7) e para o algoritmo DBSCAN (figura 8). Para o *K-Means*, ouve uma diferença menor, como mostrado e explicado na figura 9.

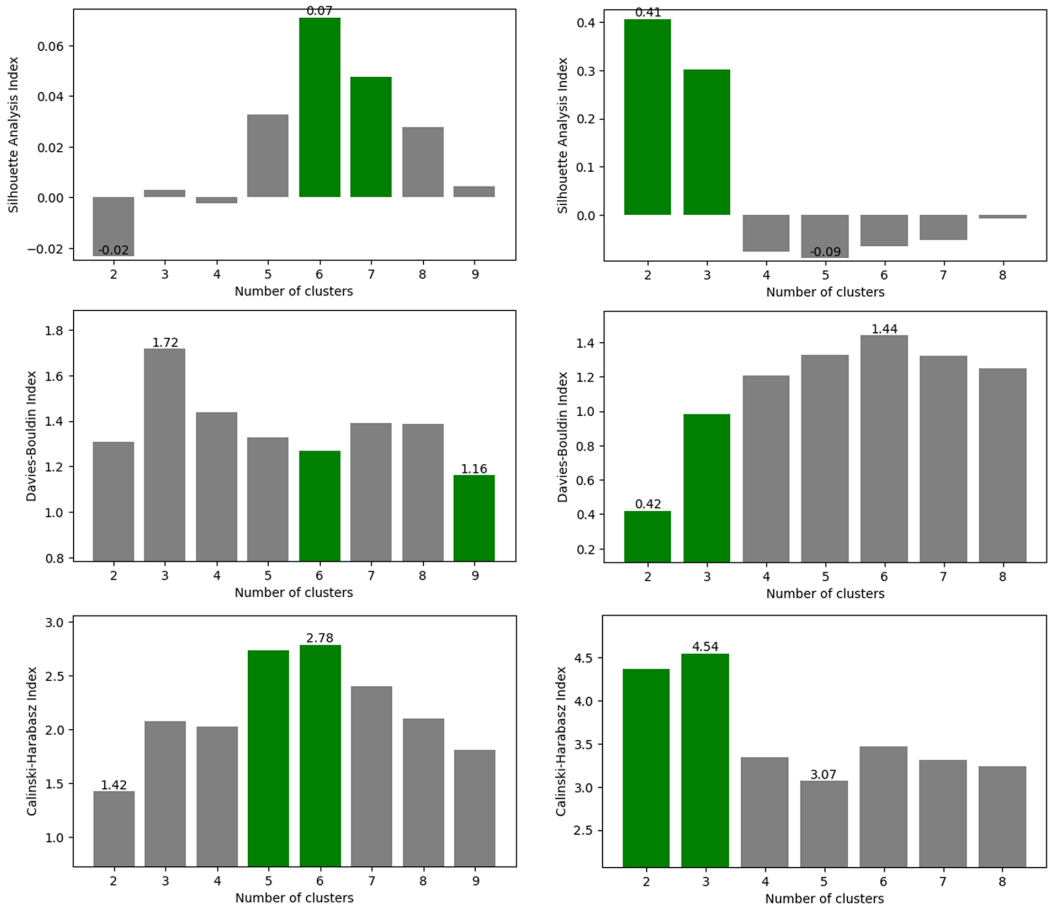

Figura 7: Comparação das métricas dos resultados gerados pelo algoritmo SOM para os dois diferentes períodos analisados (as coordenadas geográficas estão incluídas como variáveis nesse caso). Os gráficos da esquerda referem-se ao período de chuva e os da direita ao período de seca. É possível perceber que para o período de seca, o SOM manteve os melhores números de grupos para ambas as métricas, diferentemente para o período de chuva. Entretanto, ao comparar as métricas nos diferentes períodos, as métricas demonstraram resultados totalmente diferentes.

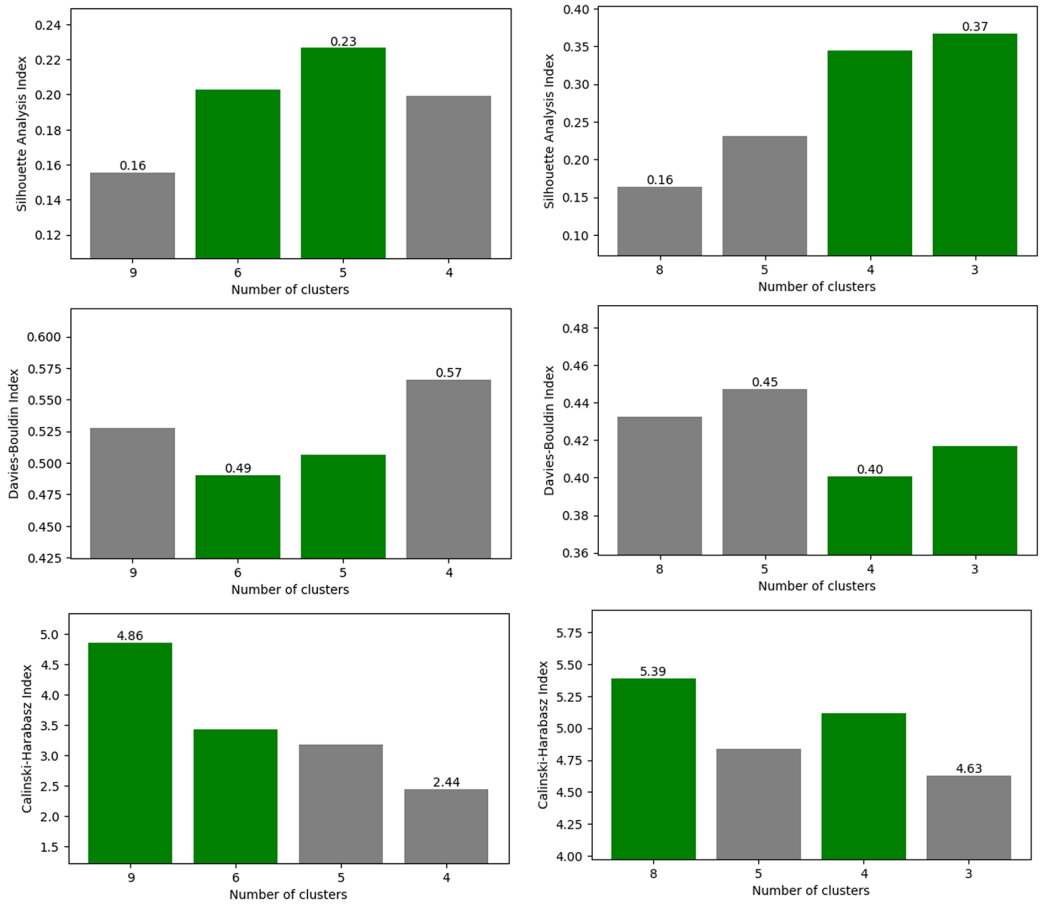

Figura 8: Comparação das métricas dos resultados gerados pelo algoritmo DBSCAN para os dois diferentes períodos analisados (as coordenadas geográficas estão incluídas como variáveis nesse caso). Os gráficos da esquerda referem-se ao período de chuva e os da direita ao período de seca. Embora haja uma diferença numérica entre os valores, o melhor número de grupos, analisando métrica por métrica, permanece a mesma para os períodos diferentes. Entretanto, para um mesmo período e para o mesmo algoritmo de aprendizado de máquina, as métricas não concordam entre si.

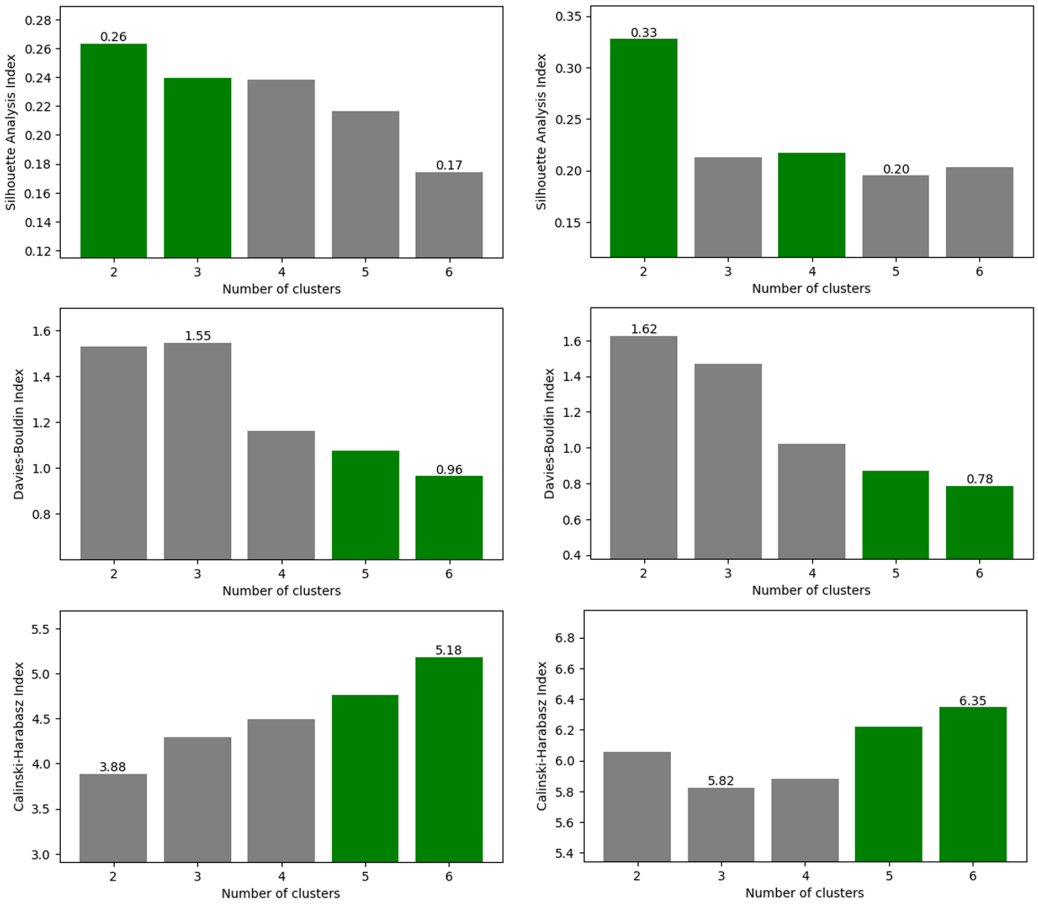

Figura 9: Comparação das métricas dos resultados gerados pelo algoritmo *K-Means* para os dois diferentes períodos analisados (as coordenadas geográficas estão removidas como variáveis nesse caso). Os gráficos da esquerda referem-se ao período de chuva e os da direita ao período de seca. Embora haja uma diferença numérica entre os valores, o melhor número de grupos, analisando métrica por métrica, permanece a mesma para os períodos diferentes. Entretanto, para um mesmo período e para o mesmo algoritmo de aprendizado de máquina, as métricas não concordam entre si.

6.2 Rotulação dos dados nos diferentes algoritmos

Como visto anteriormente, em certos momentos duas métricas concordam entre si para definir qual o melhor número de grupos para o agrupamento, em outros momentos nenhuma das métricas utilizadas concordam entre si. Entretanto, ainda é importante analisar os rótulos indicados aos dados ao fazer um agrupamento específico pois, dados um certo número K de grupos, caso ambos algoritmos apontem o grupos similares, é um sinal positivo para o funcionamento dos algoritmos.

Nos experimentos feitos, os grupos nunca se coincidiam a um agrupamento em comum, evidenciando que os dados não apresentam padroes claramente definidos.

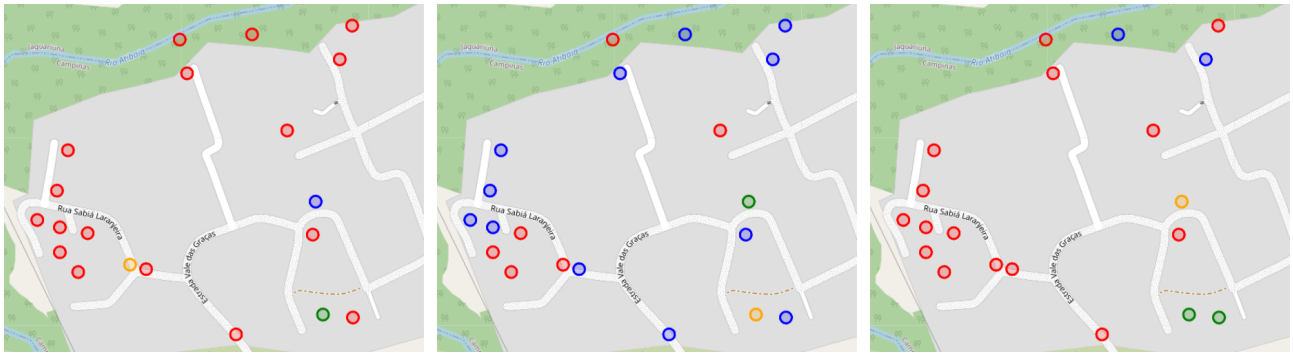

Figura 10: Agrupamentos feitos pelos três algoritmos de aprendizado de máquina utilizados. Da esquerda para a direita, tem-se o DBSCAN, o *K-Means* e o SOM. Os círculos marcados no mapa são referentes aos pontos de coleta das amostras de água na campanha feita no período de chuva. As cores representam os rótulos obtidos para um agrupamento feito com um número de grupos igual a 4 para o período. Nesses mapas, estão sendo considerados as coordenadas geográficas como variáveis.

7 Discussão

Como apresentado anteriormente, os agrupamentos encontrados pelos diferentes modelos de clusterização não apresentaram grande consistência entre si, indicando que não existem padrões de clara separabilidade em grupos dos dados analisados.

Uma hipótese levantada foi a respeito do baixo espaço amostral, no qual se trata de 39 amostras coletadas divididas em duas outras bases de dados, com 19 e 20 amostras cada, para um número alto de dimensões do espaço, de 47 variáveis (que pode abaixar para 45 se desconsiderar as coordenadas geográficas). Com o baixo número de observações, os algoritmos poderiam ter dificuldades de encontrar padrões estatísticos para executar o agrupamento não-supervisionado de forma coerente entre si.

Outra hipótese discutida é sobre a complexidade do problema, no qual uma análise hidrogeológica apenas pelas características físico-químicas das amostras não é suficiente para uma execução eficiente dos algoritmos. Talvez, com outras variáveis do ambiente como topologia, condições climáticas, altitude e influência humana no ambiente.

Ao apresentar os diversos mapeamentos feitos pelos algoritmos testados para a especialista em hidrogeologia Profa. Dra. Ana Elisa de Abreu, foi discutido que alguns grupos formados pelo *K-Means* realmente fazem algum sentido hidrogeoquímico. Entretanto, ao alterar o número de grupos ou período analisado, o *K-Means* apresenta resultados incongruentes, como os demais algoritmos.

Em geral, não houve um algoritmo cujo resultados foram superiores em todos os casos, podendo, então, considerar que ambos obtiveram desempenho semelhantes. Entretanto, vale destacar que o SOM utilizado segue um agrupamento hierárquico, ou seja, ao aumentar o número de grupos, há apenas a fragmentação de um grupo maior em um grupo menor, sem grandes alterações como nos demais algoritmos.

8 Conclusão

Baseando-se nos experimentos feitos, é notável que a aplicação dos algoritmos não produz resultados que se confirmem entre si, ao utilizar a base de dados das coletas de águas subterrâneas feitas pelo IG na região do Piracambaia II.

Ao comparar os resultados entre os três algoritmos utilizados, o *K-Means*, DBSCAN e o SOM, ambos algoritmos apresentaram resultados consideravelmente diferentes para todos os experimentos. Além disso, os rótulos gerados pelos algoritmos, em geral, mostraram-se pouco significativos para um especialista na área da hidrogeologia. Dessa forma, conclui-se que a aplicação desses algoritmos para essa base de dados não apresentou contribuições significativas para o especialista.

9 Considerações Finais

O presente trabalho proporcionou a aproximação dos grupos de pesquisa em hidrogeologia do Instituto de Geociências (IG) e o grupo de inteligência artificial multimodal da Faculdade de Engenharia Elétrica e de Computação (FEEC) da UNICAMP.

No contexto dessa aproximação, possibilitou que o aluno de iniciação científica e as pesquisadoras envolvidas pudessem se apropriar dos jargões e parte dos conhecimentos especialistas de cada área.

Partindo-se do conhecimento adquirido de que os dados hidroquímicos de poços não fornecem informações suficientes, por exemplo, para a identificação de aquíferos, trabalhos futuros podem envolver a busca por séries temporais mais longas e com maior número de poços e o estudo de quais variáveis são as mais relevantes na caracterização da qualidade d'água desses poços. Tais estudos estão alinhados com abordagens estado-da-arte da aplicação de modelos de aprendizagem de máquina na hidrogeologia.

Referências Bibliográficas

[Aguiar 2023]AGUIAR, I. *Self-Organizing-Maps - A Python implementation of Self-Organizing Maps (SOM)*. 2023. GitHub repository. Disponível em: <https://github.com/IanAguiar-ai/Self_Organizing_Maps>.

[Alencar 2021]ALENCAR, J. d. M. *Características sanitárias dos poços domésticos e a influência das condições de redução-oxidação na qualidade das águas subterrâneas do aquífero aluvial do rio Atibaia*. 130 p. Monografia (Mestrado) — Universidade Estadual de Campinas, Campinas, SP, 2021.

[Alencar et al. 2023]ALENCAR, J. M. de et al. *Hydrochemical Data from Underground Water Samples from Piracambaia II, Campinas, Brazil*. Repositório de Dados de Pesquisa da Unicamp, 2023. Disponível em: <https://doi.org/10.25824/redu/WNGM07>.

[Caliński e Harabasz 1974]CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974.

[Commission et al. 2016]COMMISSION, E. et al. *Guidance document on the estimation of LOD and LOQ for measurements in the field of contaminants in feed and food*. [S.l.]: Publications Office, 2016.

[Davies e Bouldin 1979]DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979.

[Ester et al. 1996]ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. [S.l.]: AAAI Press, 1996. (KDD'96), p. 226–231.

[Fayyad et al. 1996]FAYYAD, U. M. et al. Knowledge discovery and data mining: Towards a unifying framework. In: *KDD-96 Proceedings*. [S.l.: s.n.], 1996. v. 96, p. 82–88.

[Friedel et al. 2020]FRIEDEL, M. J. et al. Comparison of four learning-based methods for predicting groundwater redox status. *Journal of Hydrology*, Elsevier, v. 580, p. 124200, 2020.

[Iritani e Ezaki 2014]IRITANI, M.; EZAKI, S. *As águas subterrâneas do Estado de São Paulo*. Governo do Estado de São Paulo, Secretaria do Meio Ambiente, 2014. (Cadernos de educação ambiental). ISBN 9788562251306. Disponível em: <https://books.google.com.br/books?id=_N6SxgEACAAJ>.

[Kohonen 1990]KOHONEN, T. The self-organizing map. *Proceedings of the IEEE*, Institute of Electrical and Electronics Engineers (IEEE), v. 78, n. 9, p. 1464–1480, 1990. Disponível em: <https://doi.org/10.1109/5.58325>.

[MacQueen et al. 1967]MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

[Rousseeuw 1987]ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.

[Xu e Tian 2015]XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, Springer, v. 2, p. 165–193, 2015.