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

Diego Xavier Machado*, Paula Dornhofer Paro Costa†, Ana Elisa de Abreu‡
*Faculdade de Engenharia Mecânica (FEM)
†Depto. Eng. de Computação e Automação (DCA), Faculdade de Eng. Elétrica e de Computação (FEEC)
‡Depto. de Geologia e Recursos Naturais (DGRN), Instituto de Geociências (IG)
Universidade Estadual de Campinas (Unicamp)
Campinas, Brasil
e-mail: d250258@dac.unicamp.br, paulad@unicamp.br, aeabreu@unicamp.br

I. 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 [1]. 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 realizadas pelo conhecimento de especialistas. Porém, essa abordagem apresenta limitações para o estudo de sistemas subterrâneos complexos ou estudos em grande escala, regionais ou nacionais.

Nesse contexto de saúde pública, o presente projeto propõe aplicar, comparar e avaliar algoritmos aprendizado de máquina não-supervisionado para fazer um agrupamento dos dados coletados pelo Instituto de Geociências da Unicamp (IG) de águas subterrâneas da região do Piracambaia II, localizada ao norte do município de Campinas (SP), que foram objetos de estudo para a dissertação de mestrado Alencar [2].

Além do estudo feito por Joana Alencar, o presente projeto tem como inpiração o artigo de Friedel et al. [3] a respeito do uso de algoritmos de aprendizado de máquina não supervisionado para prever o caráter redox de águas subterrâneas.

Este trabalho foi financiado pelo Programa Institucional de Bolsas de Iniciação Científica (PIBIC), CNPq.

II. Método

A. Tratamento dos dados: concatenação das bases de dados, análise exploratória e descritiva

Antes da aplicação dos algoritmos de aprendizado de máquina, foi necessário reunir todos os arquivos referentes ao projeto em uma base de dados única. Nesse processo, alguns dados — provenientes de arquivos *.pdf*, *.xmlx* e *.docx* — necessitaram de ser transcritos de forma manual para a base de dados que concatenava todas as amostras e variáveis em um só arquivo *.csv*. Outros dados, no formato *.xlsx*, puderam ser facilmente transcritos para a base de dados completa com o auxílio da biblioteca *pandas*, da linguagem *python*, utilizada em todo o projeto. Após a concatenação, a base de dados era composta de uma coluna referente ao local de coleta da amostra (que majoritariamente eram poços), uma coluna referente a data de coleta das águas para aquela amostra, e as demais colunas referentes as variáveis fisico-químicas que seriam utilizadas para a execução dos algoritmos. Ademais, faz-se presente duas colunas referentes às coordenadas geográficas em formato UTM, referente a cada amostra coletada.

Feito a concatenação das bases de dados e organização do formato da base de dados, foi necessário o tratamento de dados faltantes ou inválidos. Os valores inválidos eram referentes aos dados de amostras em que uma ou algumas das substâncias químicas (que são variáveis) apresentaram valores menores que o limite de detecção (LD) do instrumento utilizado para a análise e, na planilha presente em um dos arquivos originais, era referenciado como "$< LD$". Para tratá-los, esses valores não numéricos foram substituídos por um valor que é duas vezes o valor do LD para a substância coletada. Ao realizar a análise descritiva dos dados percebeu-se que algumas substâncias químicas (as que possuíram muitos valores anotados como "$< LD$") tiveram sua mediana abaixo do limite de quantificação (LQ) do instrumento utilizado para a análise e estas variáveis foram retiradas da base de dados.

A partir desse primeiro tratamento dos dados, foi obtido um *DataFrame* com 39 linhas (amostras) e 59 colunas (variáveis, juntamente com as colunas do nome do local de coleta e da data de coleta). Em seguida, foi organizado um heatmap dos dados, o que permitiu analisar as correlações e decidir pela exclusão das variáveis com alta correlação entre si.

B. Aplicação dos algoritmos de aprendizado de máquina

O *heatmap* com as correlações das 59 variáveis restantes (Figura 1) revelou que algumas variáveis tinham alta correlação com outras, de tal forma que as seguintes variáveis foram removidas: La, Ce, Pr, Sm, Eu, Tb, Dy, Ho, Tm, Yb, Lu, TDS. Assim, restara 47 variáveis na base de dados final para a execução dos algoritmos de aprendizado de máquina, com o *heatmap* mostrado pela Figura 2.

Assim sendo, base de dados filtrada e tratada é composta por 39 amostras e 47 variáveis. É de interesse do projeto dividir essa base de dados baseado na data da coleta das amostras, sendo uma para o período de chuva na região e uma para o período de seca. Dessa forma, essas bases possuem um total de 19 e 20 amostras, respectivamente.

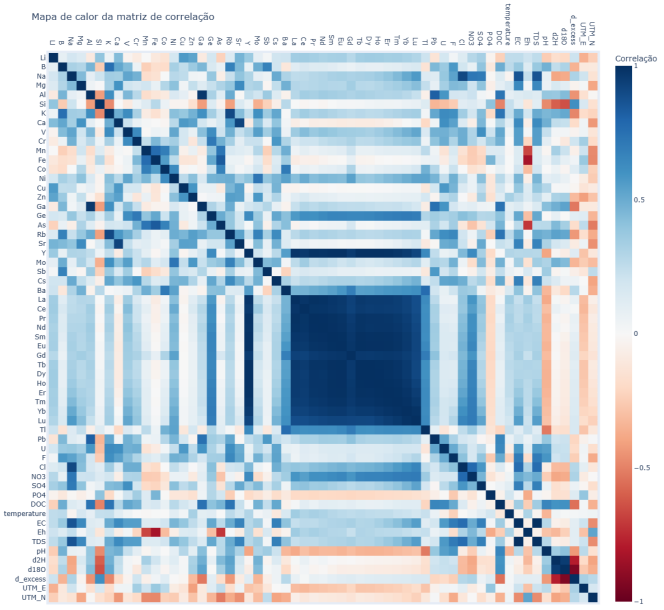

Figura. 1. *Heatmap* com 59 variáveis

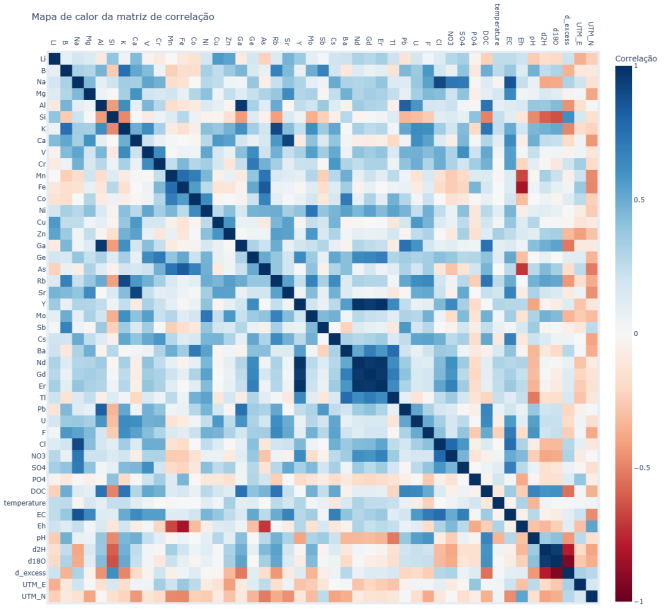

Figura. 2. *Heatmap* da base de dados final, com 47 variáveis

1) K-Means: Para aplicar esse algoritmo, foi utilizado a biblioteca *scikit-learn*. Percebeu-se que a cada vez em que se executava esse algoritmo — tanto para a base de dados do período de seca como para o período de chuva — formavam-se agrupamentos diferente. Esse problema é devido ao grande número de variáveis para poucas amostras. Para contornar esse obstáculo, para cada número K de grupos no qual o algoritmo faria os agrupamentos, o algoritmo foi executado 20.000 vezes e foi escolhido o agrupamento mais frequente para aquele número K. Isso foi feito para $K = 1, 2, ..., 6$ (valores escolhidos) e foi feita uma lista com os agrupamentos resultantes para avaliá-los e compará-los nas métricas de agrupamento. Para cada número de K, foi impresso um mapa — da região analisada — com o local de cada ponto de coleta das águas subterrâneas com a cor respectiva de seu rótulo do grupo a qual faz parte. Para a impressão dos mapas foi utilizado a biblioteca *folium*.

2) Self-Organizing Maps (SOM): Para aplicar esse algoritmo, foi utilizado a biblioteca *Self_Organizing_Maps* [4], que usa de métodos determinísticos para fazer o agrupamento, ou seja, os agrupamentos feitos para qualquer número de grupos são iguais a cada execução. O resultado também foi impresso, utilizando a biblioteca *folium*, o mapa de cada agrupamento formado e seus respectivos grupos.

C. Aplicação das métricas dos algoritmos de agrupamento

Nessa etapa, foi utilizada a biblioteca *metrics* [5] para avaliar e comparar os dois algoritmos de agrupamento testados, para ambos os períodos de coleta. Dentre as métricas utilizadas, estão o índice de silhueta [6], índice de Calinski-Harabasz [7] e o índice de Davies–Bouldin [8].

D. Avaliação da importância das coordenadas geográficas como variáveis de agrupamento

Foram testados agrupamentos pelos método K-means e pelo SOM com e sem as coordenadas geográficas como variáveis. A métrica de Informação Mútua foi testada para comparar se os agrupamentos permanecem iguais ao remover as coordenadas das variáveis da base de dados.

III. Resultados

Como primeiro resultado, foi finalizada uma base de dados formatada e filtrada para os dados coletados na região do Piracambaia II pelo IG.

De primeiro momento, pode-se perceber que a presença das coordenadas geográficas do locais de coleta das águas subterrâneas nas variáveis da base de dados teve uma pequena influência no treinamento de ambos os algoritmos. Numericamente, para cada métrica analisada comparando a respectiva base de dados do período com e sem as coordenadas, os índices mostravam valores similares entre si, podendo ter uma pequena alteração em qual o número de grupos ideais para formar o agrupamento, como mostra a Figura 3.

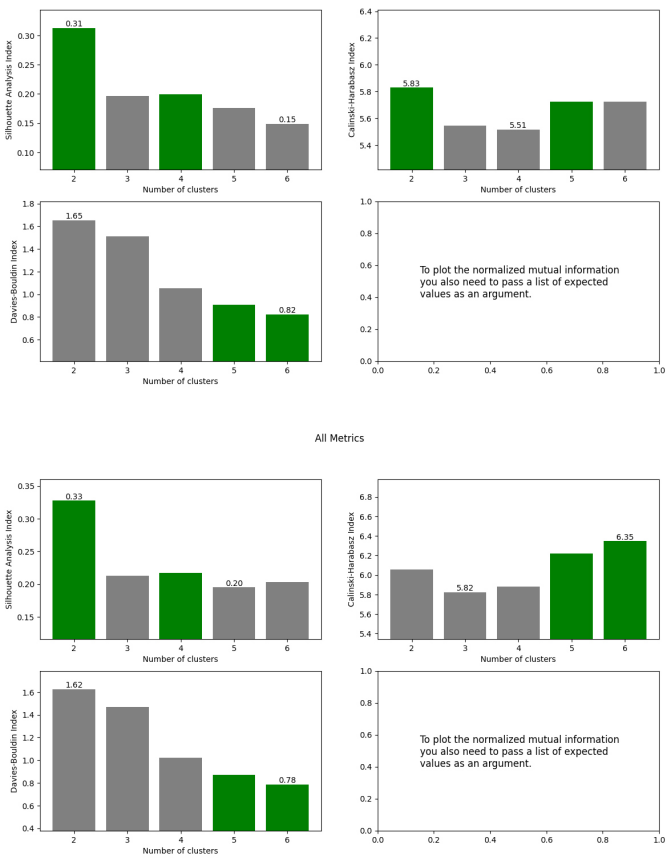

Figura. 3. Comparação das métricas do algoritmo *K-Means* para a base de dados do período de chuva. Os gráficos superiores representam as métricas com as coordenadas geográficas incluídas na base de dados para o treinamento do algoritmo. Os gráficos inferiores não contém as coordenadas para o treinamento do algoritmo. O primeiro gráfico de cada conjunto de gráficos representa o índice da análise de silhueta; o segundo gráfico (a direita do primeiro gráfico) representa o índice de Calinski-Harabasz; o terceiro gráfico representa o índice de Davies-Bouldin; o quarto gráfico não está disponível pois não temos os rótulos verdadeiros para avaliá-los. As barras destacadas em verde representam os dois melhores números de grupos para aquela métrica.

Outro resultado notável é que, para todos os experimentos feitos, os agrupamentos gerados entre os períodos de chuva e seca são diferentes, tanto para o valor do melhor número de grupos, quanto para os índices na métrica e os rótulos, como mostra a Figura 4.

Ademais, ao comparar os resultados entre os dois algoritmos utilizados, o *K-Means* e o SOM, ambos algoritmos apresentaram resultados consideravelmente diferentes para todos os experimentos. Em valores numéricos nas métricas de avaliação, o *K-Means* teve melhores resultados para o período de chuva. Para os períodos de seca, o SOM teve melhores valores para duas das três métricas aplicadas.

Por fim, ao utilizar a Informação Mútua para verificar a mudança nos agrupamentos de acordo com a presença das coordenadas geográficas na base de dados do treinamento dos algoritmos, chegamos na seguinte conclusão:

- Para o *K-Means* no período de chuva, os número de grupos no qual os dados poderiam se agrupar sem

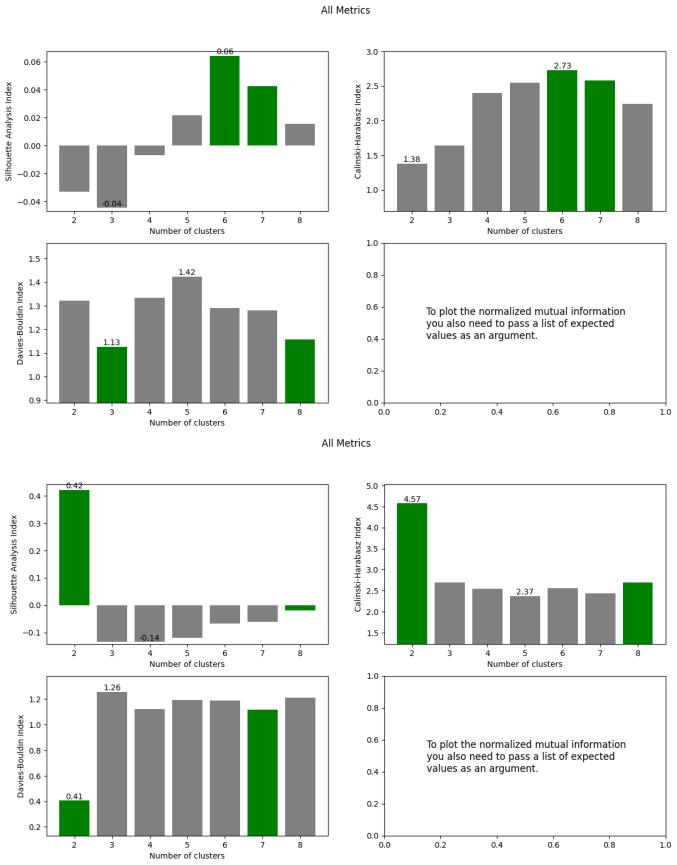

Figura. 4. Comparação dos agrupamentos feitos pelo algoritmo SOM (ambos sem as coordenadas geográficas entre as variáveis da base de dados). Os gráficos superiores representam as métricas para a execução do algoritmo para o período de chuva e os gráficos inferiores representam para o período de seca. A ordem dos gráficos é a mesma descrita na legenda da Figura 3.

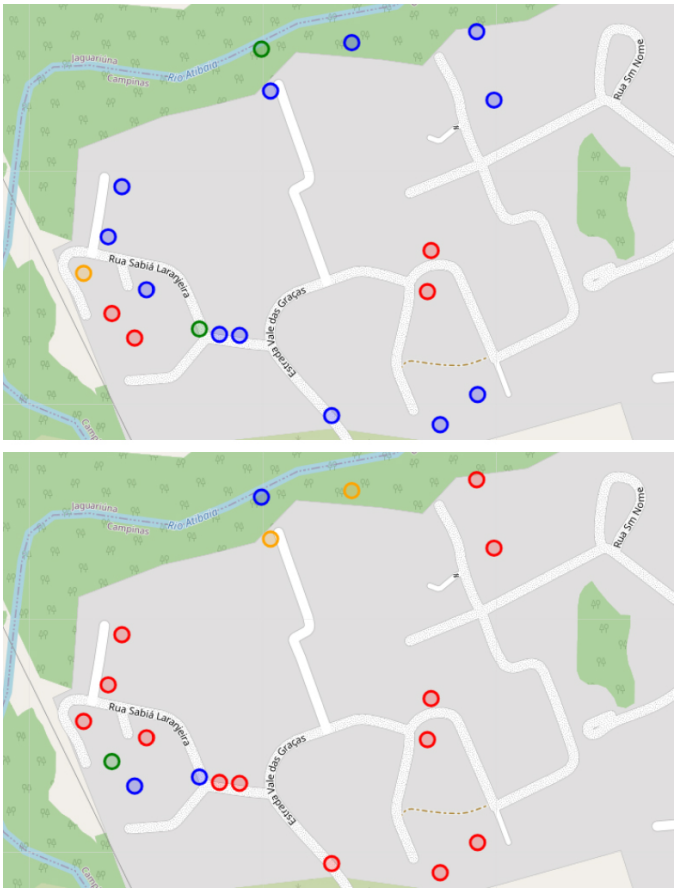

Figura. 5. Mapeamento feito com a biblioteca *folium* a partir dos rótulos gerados pelo *K-Means* e SOM, respectivamente. Ambos os agrupamentos de 4 grupos referem-se ao período de chuva sem a presença das coordenadas geográficas nas base de dados de treinamento dos algoritmos.

que houvesse resultados diferentes independentemente da presença das coordenadas geográficas são 2, 4 e 6 grupos.
- Para o *K-Means* no período de seca, esses números são de 2, 3, 4, 5 grupos.
- No algoritmo SOM, para ambos os períodos, apenas o agrupamento de dois grupos permanecem iguais independente da presença das coordenadas geográficas entre as variáveis (mas o agrupamento de dois grupos são diferentes entre os dois períodos).

Ao observar os diversos mapeamentos feitos a partir da execução dos algoritmos de aprendizado de máquina (Figura 5), além da rotulação e do mapeamento dos dados resultarem em agrupamentos diferentes, não foi possível determinar, o significado dos grupos formados, ou seja, os resultados não puderam ser interpretados por um especialista. Isso é um indicador de que a aplicação desses algoritmos para essa base de dados não se mostra muito eficiente.

IV. Conclusão

Baseando-se nos experimentos feitos, é notável que a aplicação dos algoritmos não produz resultados que se confirmem entre si, ao utilizar essa base de dados das coletas de águas subterrâneas feitas pelo IG. Uma hipótese para esses resultados incongruentes é devido a um tamanho amostral baixo para um número muito grande de variáveis, sendo necessário usar técnicas no contexto de problemas de alta dimensão.

Por fim, ao comparar os resultados entre os dois algoritmos utilizados, o *K-Means* 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.

Referências

[1] M. Iritani and S. Ezaki, *As águas subterrâneas do Estado de São Paulo*, ser. Cadernos de educação ambiental. Governo do Estado de São Paulo, Secretaria do Meio Ambiente, 2014. [Online]. Available: https://books.google.com.br/books?id=_N6SxgEACAAJ

[2] J. d. M. Alencar, "Características sanitárias dos poços domésticos e a influência das condições de redução-oxidação na qualidade das águas

subterrâneas do aquífero aluvial do rio atibaia," Campinas, SP, p. 130, 2021.

[3] M. J. Friedel, S. Wilson, M. Close, M. Buscema, P. Abraham, and L. Banasiak, "Comparison of four learning-based methods for predicting groundwater redox status," *Journal of Hydrology*, vol. 580, p. 124200, 2020.

[4] I. Aguiar, "Self-organizing-maps - a python implementation of self-organizing maps (som)," GitHub repository, 2023. [Online]. Available: https://github.com/IanAguiar-ai/Self_Organizing_Maps

[5] ——, "Metrics - a repository with metrics functions for machine learning," GitHub repository, 2023. [Online]. Available: https://github.com/IanAguiar-ai/metrics

[6] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[7] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[8] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.