XXXI Congresso de Iniciação Científica Unicamp 2023

Classificação e Mapeamento de Sistemas Aquíferos por Aprendizado de Máquina

Ian dos Anjos Melo Aguiar*, Ana Elisa Silva de Abreu***, Paula Dornhofer Paro Costa**
*Instituto de Matematica, Estatística e Computação Científica (IMECC)
**Depto. Eng. de Computação e Automação (DCA), Faculdade de Eng. Elétrica e de Computação (FEEC)
***Instituto de Geociências (IG)
Universidade Estadual de Campinas (Unicamp)
Campinas, Brasil
e-mail: *i172483@dac.unicamp.br, **paulad@unicamp.br, ***aeabreu@unicamp.br

I. Introdução

O hidrogeólogo considera para a classificação e delimitação de uma unidade aquífera a geologia do subsolo, a distribuição e características das camadas rochosas e aquíferas, o fluxo de água subterrânea e as características químicas das águas subterrâneas. Sabe-se por [1] que os diferentes aquíferos do estado têm características hidroquímicas próprias, posto isso, seria possível, com apenas esses dados, treinar a máquina para que ela classifique de forma semelhante a um hidrogeólogo?

Utilizando dos dados hidroquímicos disponibilizados pela *CETESB* (Companhia Ambiental do Estado de São Paulo), o presente trabalho avaliou a seguinte hipótese, *"Os algorítmos de aprendizado de máquina não supervisionado conseguem delimitar corretamente os aquíferos de São Paulo utilizando apenas a parte mais facilmente quantificável de uma unidade aquífera, ou seja, as características hidroquímicas"*.

Foram testados dois algorítmos, o *K-Means*, que é mais simples, e o *Self Organazing Maps* ou em português Mapas Auto-Organizáveis ou ainda Mapas de *Kohonen* [2], inspirado em literaturas como [3] [4] [5] [6] que mostram a capacidade do *SOM* em fazer agrupamentos hidroquímicos.

Ao final do trabalho se espera que as classificações fiquem semelhantes a apresentada pela *CETESB* que pode ser visto na figura 1.

II. Método

A. Extração e Organização dos Dados

A *CETESB* obtém periodicamente dados das águas subterrâneas de São Paulo, com o intuito de avaliar a qualidade das águas. Esses dados são obtidos desde os anos 90, e disponibilizados no *INFOAGUAS*.

O *INFOAGUAS* disponibiliza os dados hidroquímicos da rede de monitoramento de forma gratuita por meio do seu banco de dados e relatórios periódicos. Os relatórios a partir do ano de 2010 tem os dados em planilhas separadas. Estes resultados analíticos são um conjunto de tabelas no formato *XLSX*, que infelizmente não seguem todas o mesmo padrão. Os dados usados no presente trabalho foram os de 2013 a 2018 e eles foram tratados, de forma a gerar uma planilha única a ser disponibilizada no repositório de dados da Unicamp (REDU).

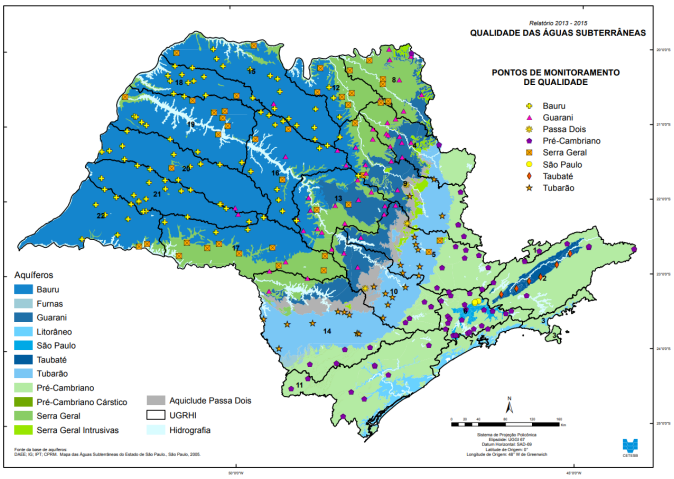

Figura. 1. Pontos de Monitoramento de Qualidade

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

B. Classificação pelo K-Means

Foi aplicado um *K-Means* próprio aos dados (veja em V). A diferença dele para a de pacotes prontos é que ele inicia os K grupos em cima de pontos da amostra, além de tentar minimizar a soma das variâncias dos grupos. Foram feitas várias classificações, considerando K grupos onde K varia entre 2 a 14 grupos. Os resultados foram sempre projetadas em um mapa do estado de São Paulo.

C. Classificação pelo Self Organizing Maps

O *SOM* é uma rede neural não supervisionada, em resumo ele leva a topologia dos dados em consideração uma vez que seus neurônios estão em uma espécie de grade onde cada neurônio é influenciado pelo neurônio vizinho. O *SOM* usado na pesquisa é próprio (veja em V). A principal diferença dele para as bibliotecas prontas é que, a grade 2d 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.

Para que o número de grupos fosse menor foi tomada a seguinte abordagem: os neurônios foram classificados de forma hierárquica utilizando um dendrograma que se aproveita da Matriz-U, que é a matriz de distâncias dos neurônios; considerando um número K de grupos, que varia de 2 ao numero total de neurônios que representam algum dado, foram aplicadas as métricas, as quais apontaram o número ideal de agrupamentos.

É importante esclarecer que tanto o *SOM* quanto o *K-Means* são não hierárquicos [7].

D. Métricas

Para avaliar a qualidade das classificações foram escolhidas quatro métricas (veja em V); três delas são internas, ou seja, avaliam a qualidade do agrupamento sem considerar o agrupamento esperado, e uma delas é externa, ou seja, avalia a qualidade do agrupamento usando o agrupamento esperado.

As quatro métricas são: *Silhouette Analysis*, junto ao *Score* gerado qual varia de $(-1, 1)$ e deve ser maximizado; *Calinsk-Harabasz* qual varia de $(0, \infty)$ e deve ser maximizado; *Davies Bouldin* qual varia de $(0, \infty)$ e deve ser minimizado; e por fim a *Normalized Mutual Information* qual varia de $(0, 1)$ e deve ser maximizada.

Em métodos não supervisionados normalmente não se tem o agrupamento esperado, mas neste caso especifico tem-se essa informação e ela é usada para avaliar se o agrupamento segue a lógica esperada ou se ele segue uma lógica paralela, classificando os aquíferos de forma que os fatores considerados para o agrupamentos não são os que respondem a pergunta do trabalho.

E. Coordenadas geográficas como variáveis

Uma vez que para o especialista em hidrogeologia a posição espacial em que é realizada a coleta da amostra para análise hidroquímica é informação importante para a classificação da mesma, decidiu-se realizar classificações utilizando-se as coordenadas geográficas como variáveis e classificações sem as mesmas, para avaliar qual seria a sua influência na classificação não supervisionada.

III. Resultados e Discussão

Após o tratamento dos dados, foram obtidas como produto três planilhas: a principal com mais de 2700 observações, 50 variáveis, cobrindo 360 poços diferentes; a planilha 2, que considera a média de cada variável por poço e tem 360 observações; e a planilha 3, onde também é considerada a média de cada variável por poço mas são mantidas apenas 10 variáveis, escolhidas de acordo com as características principais das unidades aquíferas como constam na documentação da *CETESB*. Esta planilha 3 tem 360 observações.

Para as classificações foi utilizada a planilha 3, pois, de acordo com os diversos testes, quando todas as 50 variáveis são usadas existe um confundimento nos dois algorítmos, isto é, o algorítmo classifica usando uma lógica para os agrupamentos que tem uma acurácia muito baixa além de diminuir a interpretabilidade da classificação no mapa. Desta forma os resultados expressos a seguir referem-se ao processamento dos dados da planilha 3 (10 variáveis selecionadas).

A. Classificação com as 10 Variáveis Hidroquímicas apenas (K-Means, SOM)

A figura 2 ilustra o resultado obtido com o *K-Means*, para K = 7 grupos. A tabela I (em *K-Means*) resume as métricas para avaliar a qualidade destes agrupamentos (métricas internas), e a tabela II (em *K-Means*) avalia a qualidade da classificação do aquífero levando em consideração o resultado esperado (métrica externa).

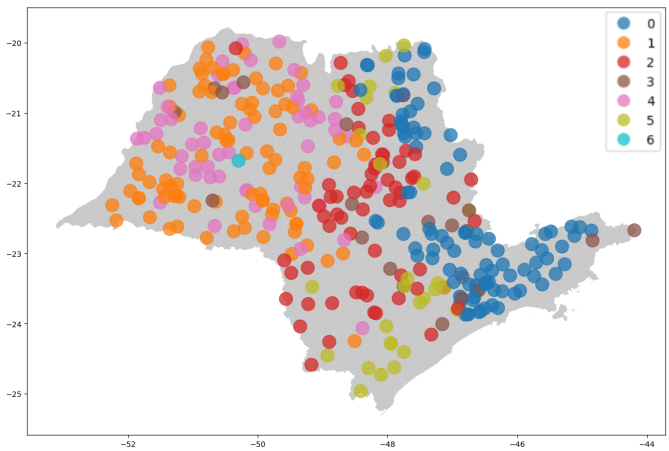

Figura. 2. Classifiação pelo *K-Means* Considerando 7 Grupos

O *SOM* cria um grupo para cada neurônio, no exemplo da figura 3 e nas tabelas I e II (*SOM*) foi usado uma grade de 20 por 20 neurônios, no total 400 neurônios/grupos. Como tem-se os rótulos das unidades aquíferas esperadas para cada observação, é possível enxergar com o *Self Organazing Maps*

Número de Grupos	Silhouette Analysis			Calinsk-Harabasz			Davies Bouldin		
	K-Means	SOM	SOM Final	K-Means	SOM	SOM Final	K-Means	SOM	SOM Final
2	0,00	0,641*	0,593*	0,617	257,811*	146,714*	17,408	0,589*	0,622*
3	0,004	0,308	0,043	1,738	148,974	97,848	8,559	1,070	0,972
4	0,006*	0,299	0,087	8,061	101,672	138,063	10,602	0,882	5,604
5	-0,051	-0,118	0,051	5,927	80,978	104,476	12,069	1,175	5,165
6	-0,089	-0,036	0,008	8,656	77,483	84,236	4,959*	1,024	6,592
7	-0,108	-0,085	-0,008	12,170	65,116	74,923	5,804	2,003	5,015
8	-0,096	-0,056	-0,100	11,351	70,419	64,367	18,793	1,933	4,614
9	-0,133	-0,082	-0,088	7,197	62,621	63,777	10,533	3,591	4,304
10	-0,188	-0,094	-0,116	13,240	56,258	57,396	11,121	3,566	4,064
11	-0,230	-0,077	-0,156	15,836*	63,516	51,800	7,243	3,541	4,359

Tabela I
RESULTADOS DAS MÉTRICAS INTERNAS, (*) É O NÚMERO ÓTIMO DE AGRUPAMENTOS DE ACORDO COM A MÉTRICA

Número de Grupos	Normalized Mutual Information		
	K-Means	SOM	SOM Final
2	0,323	0,055	0,059
3	0,281	0,062	0,119
4	0,283	0,064	0,172
5	0,246	0,082	0,178
6	0,232	0,094	0,386
7	0,297	0,147	0,374
8	0,306	0,170	0,387
9	0,337	0,180	0,379
10	0,303	0,181	0,378
11	0,312	0,205	0,367

Tabela II
RESULTADOS DA MÉTRICA EXTERNA

a estrutura/semelhança dos aquíferos, que pode por exemplo ainda na mesma figura, ser interpretada da seguinte forma, a unidade aquífera PC(Pré-Cambriano) é em parte semelhante com SG(Serra Geral) e GU(Guarani), que são seus vizinhos bem definidos.

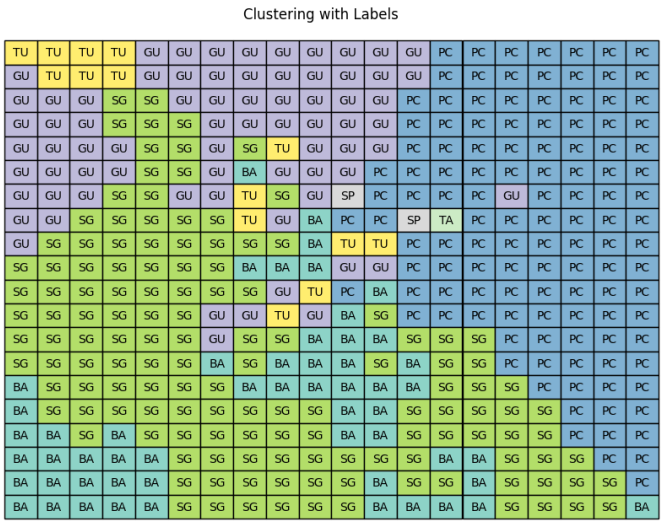

Figura. 3. Topologia dos Dados observando os Neurônios

Surpreendentemente, mesmo as métricas indicando que o *SOM* tem uma qualidade maior nos agrupamentos em comparação ao *K-Means*, ele não conseguiu classificar de forma condizente com a realidade. A Informação Mútua Normalizada aponta que tanto no SOM quanto no K-means a classificação tem uma distribuição independente da esperada.

B. Classificação Ótima, SOM com 7 Variáveis Aliado às Posições Geográficas (SOM Final)

O cenário muda se as posições geográficas dos poços forem consideradas. O *K-Means* tem mudanças muito sutis caso a latitude e a longitude dos poços forem consideradas, enquanto, em contraste, o *SOM* é bastante beneficiado. Mesmo assim é importante notar que essa abordagem tem um problema, dois poços podem estar localizado na mesma latitude e longitude, sendo distintos pela profundidade.

Aliado à adição das coordenadas, após alguns testes, as variáveis cálcio, nitrogênio-nitrato e sulfato foram retiradas, pois todas elas causavam uma piora na performance dos métodos, o *SOM* ficou então com 7 variáveis e foi denominado SOM Final.

O número ótimo de neurônios para maximizar a semelhança da classificação com a realidade foi entre 7^2 e 10^2, nas tabelas I e II podem ser vistas as métricas (em SOM Final) usando 8^2 neurônios de treinamento e considerando as mudanças propostas. Observando as métricas, o *SOM* após as mudanças (em *SOM* Final) conseguiu ter todas as métricas superiores ao *K-Means*.

C. Discussão global

Utilizando a proposta original da pesquisa, isto é, utilizando apenas os dados hidroquímicos, o *K-Means* foi superior ao *SOM* na classificação de unidades aquíferas. Algumas hipóteses do porquê do método utilizado nesta pesquisa não conseguir classificar as unidades são:

1) Os aquíferos podem ter distinções subjetivas;
2) Para classificar uma unidade aquífera precisamos de mais do que dados hidroquímicos e as coordenadas geográficas dos poços. Talvez sejam necessários os tipos de rochas predominantes, fluxo da água, dentre outras

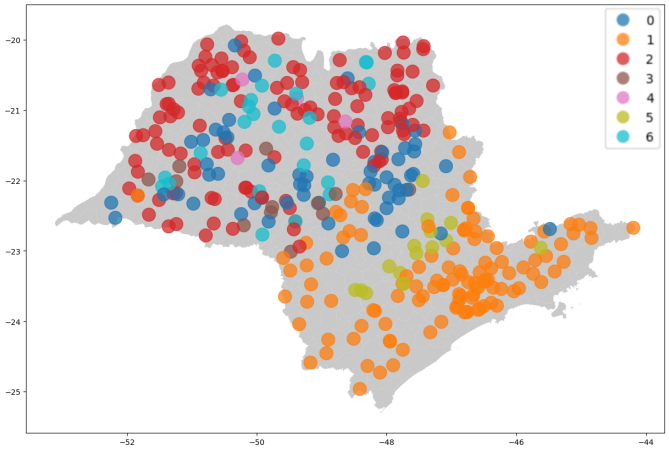

Figura. 4. Classificação pelo *SOM* Considerando 7 Grupos e a Latitude e Longitude para Comparação com os Pontos de Monitoramento da Figura 1

coisas que um especialista da área de hidrogeologia usaria;

3) Algorítmos não supervisionados tendem a classificar coisas mais obvias e menos complexas.

A primeira hipótese pode ser descartada. Dentre a segunda e terceira hipóteses, existem indícios de que a segunda esteja correta uma vez que, usando as coordenadas geográficas, o *SOM* obteve uma grande melhora na classificação em comparação ao *SOM* com apenas os dados hidroquímicos. Todavia, não é possível dizer de forma conclusiva que, se contornada a segunda hipótese, estaria garantido que um dos dois algorítmos não supervisionados de máquina pudesse classificar as unidades aquíferas.

IV. Conclusão

Os testes de classificação realizados nesta pesquisa utilizando o *K-Means* e o *SOM* revelaram que os dois algorítmos não supervisionados ainda não conseguem classificar diferentes unidades aquíferas de forma satisfatória baseando-se apenas em dados hidroquímicos. A adoção das coordenadas geográficas dos poços como variáveis melhora significativamente o desempenho do *SOM* para a classificação. Nesta pesquisa, o melhor desempenho, considerando-se a métrica externa foi obtido com o *SOM* Final e foi de 38,7%.

Os resultados sugerem que a hipótese inicial formulada para a pesquisa não e verdadeira, ou seja, não é possível classificar com aprendizado de máquina não supervisionado unidades aquíferas usando apenas dados hidroquímicos.

V. Contribuições

- Dados da *CETESB* de 2013 - 2018 mesclados e tratados;
- Biblioteca para a geração dos gráficos das métricas junto a implementação própria da Informação Mútua Normalizada, pode ser visto em [8];
- Implementação própria do K-Means, pode ser visto em [9];
- Implementação própria do SOM, pode ser vista em [10].

Referências

[1] F. F. T. e. V. S. Modesto, "Qualidade das águas subterrâneas no estado de são paulo," vol. 978-65-5577-021-6, 2021.

[2] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[3] A. S. Rahman, Y. Kono, and T. Hosono, "Self-organizing map improves understanding on the hydrochemical processes in aquifer systems," *Science of the Total Environment*, vol. 846, p. 157281, 2022.

[4] E. A. Varouchakis, D. Solomatine, G. A. C. Perez, S. Jomaa, and G. P. Karatzas, "Combination of geostatistics and self-organizing maps for the spatial analysis of groundwater level variations in complex hydrogeological systems," *Stochastic Environmental Research and Risk Assessment*, pp. 1–12, 2023.

[5] F. Iwashita, M. J. Friedel, and F. J. Ferreira, "A self-organizing map approach to characterize hydrogeology of the fractured serra-geral transboundary aquifer," *Hydrology Research*, vol. 49, no. 3, pp. 794–814, 2018.

[6] K. Nakagawa, H. Amano, A. Kawamura, and R. Berndtsson, "Classification of groundwater chemistry in shimabara, using self-organizing maps," *Hydrology Research*, vol. 48, no. 3, pp. 840–850, 2017.

[7] F. K. Gülagiz and S. Sahin, "Comparison of hierarchical and non-hierarchical clustering algorithms," *International Journal of Computer Engineering and Information Technology*, vol. 9, no. 1, p. 6, 2017.

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

[9] ——, "K-means - a python implementation of k-means," GitHub repository, 2023. [Online]. Available: https://github.com/IanAguiar-ai/K_Means_Variance

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