

Avaliação estatística dos fatores cruciais nos parâmetros de qualidade da água no reservatório João Leite, Goiás

Thiago Monteles de Sousa¹, Amanda Bueno de Moraes²,
Warde Antonieta da Fonseca Zang², Regina Célia Bueno da Fonseca²

¹Universidade Federal de Goiás, Instituto de Informática - Goiânia, GO, Brasil

²Instituto Federal de Goiás, Programa de Pós-Graduação
Mestrado Profissional em Tecnologia, Gestão e Sustentabilidade
- Campus Goiânia - Goiânia, GO, Brasil

thiagomonteles@discente.ufg.br, amandamoraes@saneago.com.br

{warde.zang, regina.fonseca}@ifg.edu.br

Abstract. *This study identified determining factors in the water quality of the João Leite reservoir in Goiás to assist in making decisions about the use of resources. Statistical techniques and computational methods were applied to evaluate multicollinearity, co-occurrence and explainability of water quality parameters. Initially, parameter selection reduced from 19 to 7 parameters through Factor Analysis and Principal Component Analysis, condensing them into two main components that explain 86.67% of the total variation in the data. Furthermore, analysis resulted in the creation of a non-linear regression model with XGBoost to predict the turbidity parameter, resulting in a mean squared error of 4.93.*

Resumo. *Este estudo identificou fatores determinantes na qualidade da água do reservatório João Leite em Goiás para auxiliar na tomada de decisões sobre o uso de recursos. Foram aplicadas técnicas estatísticas e métodos computacionais para avaliar multicolinearidade, coocorrência e explicabilidade dos parâmetros de qualidade da água. Inicialmente, a seleção de parâmetros reduziu de 19 para 7 parâmetros por meio de Análise Fatorial e Análise de Componentes Principais, condensando-os em dois componentes principais que explicam 86,67% da variação total dos dados. Além disso, análise resultou na criação de um modelo de regressão não-linear com XGBoost para prever o parâmetro turbidez, obtendo como resultado um erro quadrático médio de 4,93.*

1. INTRODUÇÃO

Os recursos hídricos, especialmente no que diz respeito à qualidade das águas, estão sujeitos a alterações devido às interações com o ambiente. Essas mudanças ocorrem em razão da exploração humana dos recursos naturais em áreas urbanas, industriais e agrícolas, bem como por ação de processos naturais, como erosão, intemperismo das rochas e variações na precipitação. Essas alterações podem resultar na contaminação da água, tornando-a inadequada para consumo humano, recreação, processos industriais e práticas agrícolas [Guedes et al. 2012].

Esta questão representa um desafio de grande relevância para a gestão dos recursos hídricos global, uma vez que envolve diversas questões interconectadas, como a qualidade da água e a necessidade de orientar profissionais na escolha dos parâmetros essenciais para monitorar e preservar esse recurso vital.

No âmbito da exploração de dados multivariados compostos por um alto volume de informações, técnicas estatísticas apresentam ferramentas essenciais para promover tanto a análise como a seleção de parâmetros que tenham uma alta explicabilidade no contexto em questão [Autin and Edwards 2010]. No campo de dados ambientais, técnicas como análise fatorial e análise da componente principal vêm sendo empregadas com o propósito de analisar e extrair informações desses conjuntos de dados, além de permitir a redução da dimensão dos conjuntos, tornando o tratamento e a elaboração de ferramentas computacionais mais ágeis [Andrade et al. 2007b].

Este estudo tem como objetivo apresentar uma proposta para a seleção de parâmetros que desempenhem um papel determinante na avaliação da qualidade da água do reservatório João Leite, um dos principais mananciais da Região Metropolitana de Goiânia, Goiás. Além disso, como experimento utilizar seus resultados como procedimento para a elaboração de um modelo de regressão para previsão do parâmetro físico de qualidade de água turbidez. Para alcançar esse objetivo, será apresentado uma sequência de etapas que visam analisar e identificar evidências estatísticas para a seleção dos parâmetros do conjunto de dados em análise e finalizando com a elaboração de um modelo de regressão não-linear.

2. MATERIAIS E MÉTODOS

Nesta seção, será apresentado os materiais e métodos utilizados para analisar os parâmetros de qualidade da água e estabelecer indicadores estatísticos que poderá servir como apoio para o uso responsável dos recursos hídricos. Através da descrição das ferramentas, técnicas e abordagens empregadas, será fornecido uma base para a compreensão da metodologia adotada neste estudo.

2.1. Região de estudo

A bacia hidrográfica do Ribeirão João Leite, ilustrado na Figura 1, está situado entre os paralelos 16°13' e 16°39' sul e entre os meridianos 48°57' e 49°11' oeste, localizada no Estado de Goiás, abrange diversos municípios, incluindo Campo Limpo de Goiás, Ouro Verde de Goiás, Terezópolis de Goiás, Nerópolis, Anápolis, Goianápolis e Goiânia [Rabelo 2009].

O reservatório João Leite desempenha um papel fundamental no fornecimento de água para a cidade de Goiânia e grande parte de Aparecida de Goiânia. A água captada do reservatório passa por tratamento na estação de tratamento de água (ETA) Mauro Borges antes de ser distribuída para atender às demandas de consumo nos âmbitos doméstico, comercial e industrial em toda a cidade. "

2.2. Parâmetros de qualidade da água bruta

A base de dados utilizada nesta pesquisa foi disponibilizada pela Companhia de Saneamento de Goiás - Saneago, no intervalo compreendido entre janeiro de 2018 e junho de 2023. Os parâmetros avaliados neste trabalho incluem as métricas utilizadas no protocolo

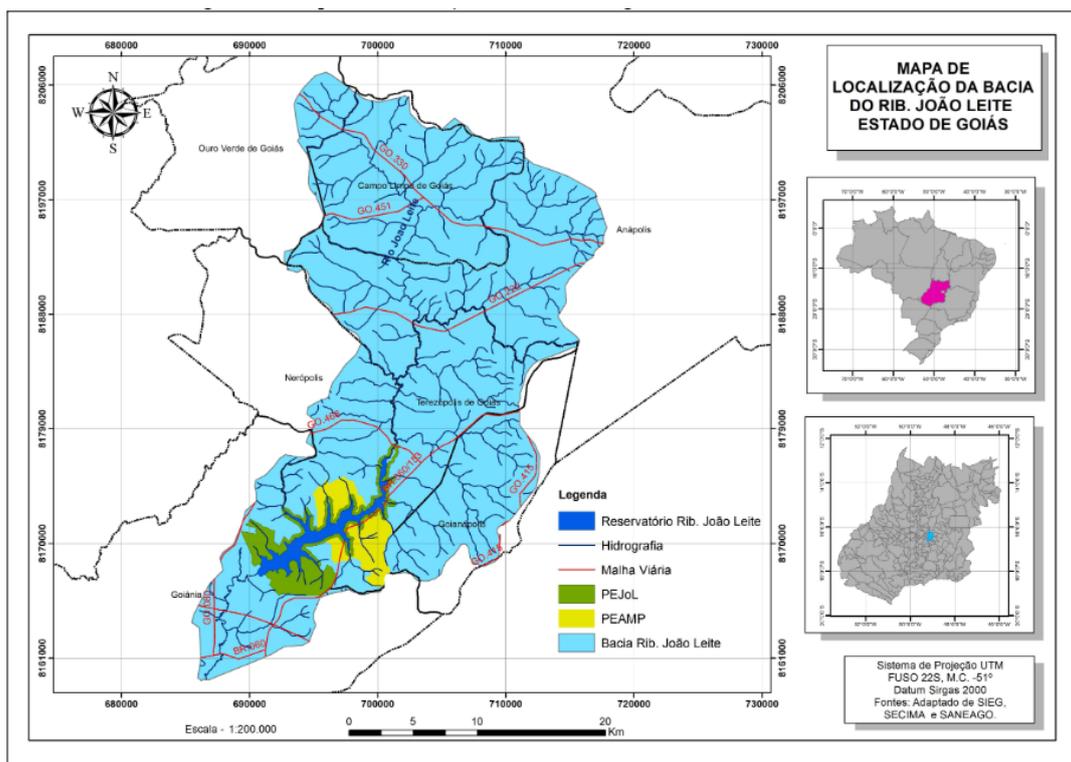


Figura 1. Mapa de localização da bacia hidrográfica do Ribeirão João Leite, onde está marcado em verde o Parque Estadual do João Leite (PEJoL) e amarelo o Parque Estadual Altamiro de Moura Pacheco (PEAMP) [Rabelo 2009].

manual para avaliação da qualidade da água, essas métricas são aplicadas pelos profissionais da área e incluem: parâmetros físicos (turbidez, cor verdadeira, pH, condutividade), parâmetros químicos (alcalinidade total, alcalinidade HCO_3^- , alcalinidade CO_3^{2-} , dureza total, cloretos, ferro solúvel, ferro total (FT), fósforo total, nitrito, nitrogênio amoniacal total, manganês (Mn), oxigênio dissolvido (OD), ortofosfato) e parâmetros microbiológicos (índice de coliforme total, índice de escherichia coli).

2.3. Seleção e análise dos parâmetros

A fim de obter recursos para analisar um grande volume de dados complexos gerados pelo monitoramento ambiental, propostas de análise estatística têm sido usadas para obter indicadores consistentes de parâmetros determinantes para a qualidade da água, assim como auxiliar a tomada de decisão de profissionais [Guedes et al. 2012].

Para isso foi elaborada uma sequência de etapas conforme ilustrado na Figura 2, cujo propósito é identificar e avaliar a multicolinearidade, consistência, cocorrência e explicabilidade dos parâmetros. Sendo efetuada através da linguagem de programação *Python*, v.3.10.2, com auxílio de bibliotecas com ferramentas estatísticas *SciPy* e *Pandas*.

O processo se inicia através da (I) filtragem de parâmetros que não possuem relações de consistências e acabam não produzindo valor significativo para etapas posteriores. Sendo que na primeira etapa, é investigada a possível multicolinearidade das variáveis por meio da análise do fator de inflação da variância (FIV), cuja finalidade é estimar se um conjunto de parâmetros possui uma relação derivativa entre si

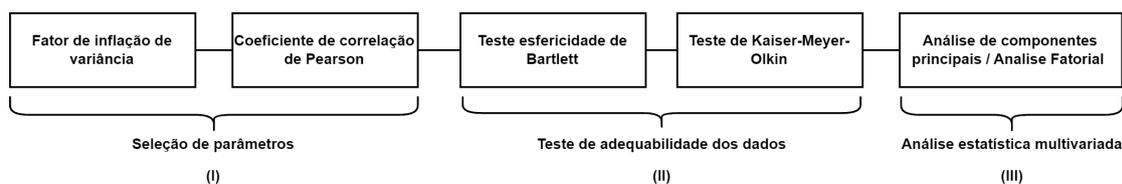


Figura 2. Processo de análise e seleção dos parâmetros de qualidade de água.

[Leite et al. 2022]. Em seguida, há a etapa do coeficiente de correlação de *Pearson*, que mede o grau de correlação entre duas variáveis de escala métrica. Sendo que, valores acima de 0,5 indicam uma forte conexão entre as variáveis-alvo, e serão cruciais para a seleção das variáveis [Guedes et al. 2012].

Em (II) iniciarão os testes para verificar a adequabilidade dos dados resultantes, a qual poderá possibilitar o análise das componentes principais (ACP) e análise fatorial (AF) do conjunto de variáveis. Para isso foi inicialmente realizado o teste de esfericidade de *Bartlett*, que resulta em um valor denominado *p-value* e tem como propósito verificar se a matriz de correlações da base de dados é estatisticamente igual a uma matriz identidade de mesma dimensão. Considerando um nível de significância de 5%, caso o *p-value* seja menor que 0,05, a hipótese nula de que a matriz de correlação não é uma matriz de identidade é rejeitada, e o processo continua. Seguindo a verificação da adequação dos dados, é realizado o teste *Kaiser-Meyer-Olkin* (KMO), que busca identificar se o conjunto de dados possui conformidade e consistência para a realização da ACP/AF. Valores a partir de 0,5 são considerados aceitáveis para prosseguir na análise final [Rocha and Pereira 2016].

Após a conclusão dos procedimentos de avaliação da qualidade e distribuição dos parâmetros, realiza-se em (III) as etapas de análise estatística multivariada. Pela ACP utilizada para analisar as inter-relações entre inúmeras variáveis em termos de suas dimensões inerentes. Para as n variáveis originais, é encontrado um meio de condensar as informações contidas em todos os parâmetros finais em um conjunto menor de variáveis (componentes) com uma perda de informação minimizada [Bernardi et al. 2012]. Por outro lado, a AF é usada para identificar e modelar as relações subjacentes (fatores latentes) entre um conjunto de variáveis observadas, que busca explicar a covariância e a correlação entre as variáveis originais em termos de fatores não observados [Andrade et al. 2007a].

Os métodos ACP/AF são amplamente utilizados na busca de parâmetros fundamentais para a explicação dos dados, tendo em vista que o procedimento para a elaboração de uma AF e ACP pode ser dividido em quatro etapas [Toledo and Nicoletta 2002] resumem em:

1. Cálculo dos autovalores e seus correspondentes autovetores de uma matriz de correlações R . Onde o primeiro autovalor a ser determinado corresponderá à maior porcentagem de variabilidade total presente.
2. Com os autovalores e autovetores calculados, é possível definir a quantidade de fatores para determinar o número de componentes a serem utilizados. Para isso, é utilizado o critério *Kaiser-Guttman*, que propõe utilizar o número de fatores F , onde, para cada fator correspondente a uma métrica, se o valor *Kaiser* para aquele autovalor for superior a 1, então será considerado na ACP e AF final.

3. Após a análise da quantidade de fatores concluída, é então realizada a ACP buscando extrair as combinações lineares das variáveis originais que capturam a maior variância nos dados.
4. Para análise fatorial, será feita a rotação dos eixos relativos aos fatores comuns através da função *varimax*, visando tornar a solução mais simples para a interpretação dos dados.

Após finalizar a sequência de etapas apresentados nesta seção, o qual sua implementação foi realizada na linguagem de programação *Python*, é possível criar experimentos com o conjunto de dados que permitam selecionar parâmetros e analisar os indicadores que influenciam os indicadores de qualidade do conjunto de informações.

2.4. Modelo de regressão não-linear

Visando demonstrar a vantagem das etapas descritas na subseção anterior (2.3) para seleção de parâmetros, será desenvolvido um modelo de regressão não-linear para a previsão da turbidez da água. A escolha do indicador de turbidez é justificada pelo fato de a medida indicar o grau de turvação ou opacidade da água devido à presença de partículas em suspensão, como sedimentos, argila, algas, matéria orgânica e outros sólidos. Sendo assim, é um parâmetro visualmente relevante para a percepção da qualidade da água, que é medida de forma manual. [Leite et al. 2022].

O modelo de previsão para turbidez utilizara o algoritmo *XGBoost*, que combina técnicas de aprendizado de máquina para melhorar a eficácia e a velocidade do modelo. O *XGBoost* utiliza o aprendizado em conjunto e técnicas de otimização para reduzir erros e melhorar as previsões, superando modelos individuais de árvore de decisão ou *Random Forest* [Somvanshi and Chavan 2016]. Os dados de treinamento são baseados nas avaliações manuais dos indicadores de qualidade de água coletados no período de janeiro de 2018 a junho de 2023, com intervalos de medição de 15 a 30 dias. Sendo utilizado como dados de treino 75% do conjunto e os restantes 25% como teste.

Em problemas de regressão, onde se quer prever uma variável contínua com base em um conjunto independente de variáveis auxiliares, as métricas se concentram em volta do erro gerado pelas previsões do modelo. Para isso será utilizado a métrica *root mean squared error* (RMSE) como observado em [Chen et al. 2017].

3. RESULTADOS E DISCUSSÕES

Nesta seção, será apresentado os principais resultados obtidos, observando o impacto das etapas ilustradas na Figura 2, com o intuito de selecionar e entender as relações entre os principais indicadores da qualidade da água do reservatório João Leite. Além de demonstrar os resultados obtidos na criação de um modelo de regressão não-linear com base nas variáveis resultantes do processo descrito.

A Tabela 1 mostra os resultados da análise do fator de inflação da variância (FIV) com as 19 métricas iniciais do conjunto de dados. O primeiro teste avaliou a possibilidade de multicolinearidade entre os parâmetros. Os números FIV (Tabela 1) maiores que 10 sugere uma influência derivativa dessas métricas em relação às outras do conjunto. Observando os demais parâmetros, é possível notar que o subconjunto (alcalinidade Total, alcalinidade HCO_3^- , alcalinidade CO_3^{2-}) apresenta uma relação de dependência, em

que a alcalinidade total agrupa informações dos demais parâmetros. Portanto, será utilizada apenas a alcalinidade total como parte do conjunto final de parâmetros. O mesmo raciocínio se aplica ao subconjunto (ferro total, ferro solúvel), onde a opção solúvel será descartada. Após a primeira etapa de filtragem, uma nova avaliação do FIV foi realizada, tendo como principais alterações os valores anteriores superiores a 10, onde nesse novo experimento foram obtidos 1,77 para alcalinidade total e 9,46 para ferro total.

Tabela 1. Tabela da análise do fator de inflação da variância (FIV)

Parâmetros	FIV
Alcalinidade Total	∞
Alcalinidade HCO_3^-	∞
Alcalinidade CO_3^{-2}	∞
Ferro Total	12,92
Ferro Solúvel	8,74
Cor Verdadeira	7,97
Turbidez	6,80
Ortofosfato	5,25
Nitrito	3,52
Nitrogênio Amoniacal Total	2,77
Manganês	2,76
Condutividade	1,66
Índice de Coliforme Total	1,65
Índice de <i>Escherichia coli</i>	1,53
Cloretos	1,46
pH	1,46
Dureza Total	1,29
Oxigênio Dissolvido	1,23
Fósforo Total	1,08

Após a remoção dos três parâmetros (alcalinidade HCO_3^- - alcalinidade CO_3^{-2} e ferro solúvel) da etapa anterior, o teste de correlação de *Pearson* será conduzido para identificar pares de variáveis com alto grau de correlação, com valores maiores ou iguais a 0,5. Variáveis com valores abaixo desse limiar serão eliminadas, conforme realizado por [Rocha and Pereira 2016]. Após essa análise, verificou-se que nove parâmetros não apresentaram uma correlação forte entre as variáveis de qualidade da água, resultando em sete variáveis com correlações destacadas na Tabela 2.

A Tabela 2 mostra a matriz de correlação de *Pearson* entre os sete parâmetros com maiores correlações em negrito. Há uma alta correlação em pelo menos um par de parâmetros, como turbidez, nitrogênio amoniacal total (NAT), cor verdadeira (Cor), manganês (Mn), ferro total (FT), nitrito e ortofosfato. Sendo destaque para um grau muito forte nos pares turbidez e FT ($r = 0,9$), Cor e NAT ($r = 0,85$), assim como também FT e Cor ($r = 0,84$).

A relação entre ferro e turbidez é importante em sistemas de tratamento de água, pois altos níveis de turbidez e ferro podem afetar a eficácia dos processos de tratabilidade. A remoção eficiente desses parâmetros são objetivos críticos no tratamento de água para garantir a qualidade da água potável. Em muitos casos, a presença de ferro na água pode

Tabela 2. Tabela de correlação de *Pearson* entre os parâmetros (melhores valores em negrito)

Parâmetros	Turbidez	NAT	Cor	Mn	FT	Nitrito	Ortofosfato
Turbidez	1,00						
NAT	0,52	1,00					
Cor	0,85	0,57	1,00				
Mn	0,45	0,60	0,62	1,00			
FT	0,90	0,71	0,84	0,62	1,00		
Nitrito	0,08	0,27	0,17	0,23	0,18	1,00	
Ortofosfato	0,09	0,06	0,13	0,23	0,12	0,65	1,00

levar ao aumento da turbidez devido à formação de partículas sólidas de óxido de ferro.

No contexto específico do Reservatório João Leite, é esperado que ocorra um aumento nos níveis de ferro e turbidez durante as estações chuvosas. Isso ocorre devido ao carregamento de solos e à ocorrência de processos de erosão nas margens do reservatório. Vale ressaltar que o ferro é um dos constituintes naturais da bacia de drenagem do Reservatório João Leite, devido à composição rochosa do Complexo Granulítico Anápolis-Itaçu, que contém rochas ricas em óxidos de ferro e alumínio [Rodrigues et al. 2013].

Em muitos casos, a presença simultânea de ferro total e manganês pode apresentar desafios adicionais. Esses elementos podem interagir e formar compostos insolúveis, contribuindo para o aumento da turbidez. Compreender essas interações é essencial para desenvolver estratégias de tratamento eficazes, otimizando a clarificação da água e garantindo a pureza necessária para o consumo humano.

A correlação de ($r = 0,84$) entre a cor e o teor de nitrogênio amoniacal total (NAT), bem como a possível interligação desses fatores, pode ser explicada pelo fato de que tanto a cor da água quanto o NAT podem estar relacionados à presença de matéria orgânica na água. Ao se decompor, essa substância pode liberar compostos orgânicos que agem como antioxidantes na água, além de contribuir para a produção da amônia (NH_4^+), que é parte do NAT.

A presença dessa matéria orgânica na bacia do João Leite pode ter diversas origens, refletindo as atividades humanas na região, tais como propriedades rurais, estradas, áreas urbanas, esgoto e efluentes industriais. Essa variedade de fontes orgânicas pode influenciar significativamente a composição e a qualidade da água na bacia. É relevante considerar os ácidos húmicos e fúlvicos de origem natural, que desempenham papéis cruciais no equilíbrio e na saúde dos ecossistemas aquáticos. Nesse contexto, a presença desses ácidos pode ser benéfica ou desafiadora, dependendo da quantidade e da interação com outros elementos presentes na água. Eles podem melhorar a fertilidade do solo já que promovem a disponibilidade de nutrientes essenciais para as plantas, influenciando diretamente a qualidade da água. Por outro lado, a matéria orgânica pode atuar como veículo de transporte de substâncias poluentes, especialmente quando há excesso de nutrientes no solo. Isso pode resultar em problemas como a eutrofização em corpos d'água, interferindo nos processos de tratamento de água.

Com a finalização das etapas de seleção de parâmetros, o experimento voltou-se

para iniciar os testes estatísticos para verificar a possibilidade de utilizar AF e ACP para a finalização da seleção dos indicadores de qualidade de água. Para isso, foi necessário realizar dois testes, cujo primeiro denominado como esfericidade de *Bartlett*, o qual resultou em um $p\text{-value} \approx 0,0$. Isso indica que há evidências estatísticas para rejeitar a hipótese nula de que a matriz de correlação é uma matriz idêntica, podendo, assim, haver uma correlação significativa entre as variáveis resultantes. Outro indicador complementar foi realizado, desta vez através do teste KMO, o qual é usado para avaliar a adequação dos dados para a análise fatorial. O teste KMO resultou em um valor de 0,7, indicando que o conjunto de dados possui adequação moderada para a AF.

Tabela 3. Tabela de peso fatorial das variáveis da qualidade de água nos dois componentes principais selecionados (melhores valores em negrito)

	PC1	PC2
Turbidez	0,86	-0,25
Nitrogênio Amoniacal Total	0,79	-0,05
Cor Verdadeira	0,91	-0,16
Manganês	0,77	0,07
Ferro Total	0,95	-0,17
Nitrito	0,34	0,85
Ortofosfato	0,27	0,87
Alto valor	3,86	1,59
% variância explicada	54,4	22,43
% variância acumulada	54,4	76,83

Concluída a etapa de verificação da adequabilidade dos dados, foi então realizada a análise do critério *Kaiser-Guttman*, que seleciona o número de componentes a serem executadas na ACP/AF, resultando em 2 componentes das 7 possíveis que obtiveram uma explicabilidade superior a 1, indicando que o número de componentes recomendado pelo critério é 2.

A tabela 3 mostra que, para a análise de componentes principais com dois componentes (PC1, PC2), explicam 76,83% da variância total das variáveis selecionadas, sugerindo que os fatores em negrito possuem a maior significância em cada fator, com o primeiro componente explicando 54,4% de todo o conjunto de variáveis.

Visando complementar as análises das relações subjacentes entre os parâmetros, aplicou-se a análise fatorial utilizando o algoritmo *Varimax*, que altera a raiz característica sem afetar a proporção da variância total explicada pelo conjunto [Andrade et al. 2007a]. Os resultados podem ser observados na Tabela 4 a seguir:

A primeira componente rotacionada (AF1) apresenta uma explicabilidade de 49,26%, tendo um total junto com o AF2 uma explicabilidade acumulada de 86,67%. Observou-se que os componentes turbidez, cor verdadeira, ferro total e nitrito apresentam um peso maior que 0,75 (ou menor que -0,75), o que indica que essas variáveis são as mais significativas na definição da qualidade da água.

Comparando as Tabelas 3 e 4, é possível observar uma interseção entre os parâmetros turbidez, cor verdadeira e ferro total no primeiro componente de ambos os testes, evidenciando uma alta contribuição dessas variáveis na distribuição de pesos sobre

Tabela 4. Tabela de Fatores dos componentes da matriz transformada pelo algoritmo Varimax (melhores valores em negrito)

	AF1	AF2
Turbidez	0,96	0,30
Nitrogênio Amoniacal Total	0,31	0,86
Cor Verdadeira	0,86	0,45
Manganês	0,30	0,64
Ferro Total	0,83	0,55
Nitrito	-0,75	-0,52
Ortofosfato	-0,60	-0,78
Alto valor	3,44	2,61
% variância explicada	49,26	37,41
% variância acumulada	49,26	86,67

a qualidade da água. Essa tríade de parâmetros está relacionada a uma sequência de processos em que o ferro dissolvido na água pode estar associado à presença de particulado em suspensão. A turbidez é uma medida da quantidade de material suspenso, como sedimentos, argila, matéria orgânica e minerais, presentes na água. Tal fator acarreta uma água turva e com aumento da cor, uma vez que essas partículas podem conter ferro ou promover sua liberação na água devido à interação com o ambiente aquático.

Observou-se também um aumento na explicabilidade acumulada, de 76,83% para ACP e 86,67% para AF, devido ao significativo aumento do fator AF2, de 37,41% em relação a 22,43% do CP2. Uma possibilidade é que a em AF pressupõe a existência de fatores latentes que influenciam as variáveis observadas, enquanto a ACP busca componentes que maximizam a variância dos dados. Sendo ACP melhor utilizado em encontrar representações a um conjunto de dados já estabelecido, enquanto AF visa simplificar a interpretação dos resultados, reforçando a relação entre as variáveis e os fatores extraídos.

Para a realização do experimento de um modelo de regressão não-linear, foram selecionadas as variáveis que demonstraram maior eficácia em termos de explicabilidade, conforme observado anteriormente. A variável alvo escolhida foi a turbidez, enquanto os dados de entrada consistiram em quatro variáveis: nitrogênio amoniacal total, cor verdadeira, manganês e ferro total. Os resultados do modelo de previsão para a turbidez, utilizando o algoritmo XGBoost, indicaram um RMSE de 4,93. Em comparação, ao utilizar as 18 variáveis originais como entrada para o modelo com as mesmas configurações, obteve-se um RMSE de 5,39. Os resultados são similares, com uma diferença de 0,46, no entanto, há uma diferença na complexidade da inferência do modelo, uma vez que a dimensão do conjunto é menor. Ao selecionar as variáveis que possuem uma elevada relação de explicabilidade com o conjunto geral, é possível aumentar a eficiência na administração de recursos para garantir a preservação e manutenção da qualidade da água.

4. Conclusões

Este trabalho apresentou um estudo sobre o uso de ferramentas estatísticas para promover a análise e seleção de parâmetros que apresentem uma relação determinante no indicativo da qualidade de água do reservatório João Leite em Goiás. Para isso, foi apresentada uma sequência de etapas que passam pela seleção de parâmetros, teste de adequabilidade

dos dados e análise multivariada dos dados. Além disso, foram realizados experimentos com a criação de modelos de regressão não-lineares visando demonstrar a utilização do processo de filtragem das variáveis.

O protocolo conduziu à seleção de 7 parâmetros, reduzindo um conjunto inicial de 19, destacando variáveis altamente correlacionadas e explicativas. O destaque foi a redução para 2 componentes principais que explicaram 86,67% da variação total dos dados, onde se destacaram os parâmetros turbidez, cor verdadeira, ferro total, nitrogênio amoniacal total e manganês como indicadores-chave na simplificação do conjunto de dados. Utilizando essas variáveis com destaque, sendo a turbidez como alvo, foi construído um modelo de regressão não-linear que apresentou um erro quadrático médio de 4,93, consideravelmente abaixo do desvio padrão de 15,25 da variável alvo.

A seleção de parâmetros de monitoramento da qualidade da água é relevante para o gerenciamento de recursos hídricos, tendo em vista as particularidades de cada ambiente apresentado neste estudo poderá ser usado para auxiliar na tomada de decisões sobre o uso de recursos e também para criar modelos computacionais para previsão de métricas.

Como trabalhos futuros, se concentrará na investigação das relações entre os parâmetros de qualidade da água e as medidas de temperatura, chuva e desmatamento. Essa análise visará, além de examinar as interações entre esses indicadores, criar modelos computacionais para auxiliar os profissionais de interesse no processo das tomadas de decisões.

Referências

- Andrade, E. M. d., Araújo, L. d. F. P., Rosa, M. d. F., Gomes, R. B., and Lobato, F. A. d. O. (2007a). Fatores determinantes da qualidade das águas superficiais na bacia do alto acaraú, ceará, brasil. *Ciência Rural*, 37(6):1791–1797.
- Andrade, E. M. D., De, L., Araújo, F. P., Rosa, M. F., Disney, W., and Alves, A. B. (2007b). Seleção dos indicadores da qualidade das águas superficiais pelo emprego da análise multivariada. *Aprovado pelo Conselho Editorial em*, 27:683–690.
- Autin, M. A. and Edwards, D. (2010). Nonparametric harmonic regression for estuarine water quality data. *Environmetrics*, 21:588–605.
- Bernardi, J. V. E., Lacerda, L. D., Dórea, J. G., Landim, P. M. B., Gomes, J. P. O., Almeida, R., Manzatto, A. G., and Bastos, W. R. (2012). Aplicação da análise das componentes principais na ordenação dos parâmetros físico-químicos no alto rio madeira e afluentes, amazônia ocidental. *Geochimica Brasiliensis*, 23(1).
- Chen, C., Twycross, J., and Garibaldi, J. M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLoS ONE*, 12.
- Guedes, H. A. S., Silva, D. D. d., Elesbon, A. A. A., Ribeiro, C. B. M., Matos, A. T. d., and Soares, J. H. P. (2012). Aplicação da análise estatística multivariada no estudo da qualidade da água do rio pomba, mg. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 16(5):558–563.
- Leite, B. J., Zimmermann, N. E. K., Lúcia, V., Ferreira, D., and Dias, F. L. (2022). Modelagem da dependência do consumo de coagulante dos parâmetros brutos da água: ajuste de um modelo de regressão linear.

- Rabelo, C. G. (2009). Mapeamento de áreas vulneráveis para a qualidade das águas superficiais na bacia hidrográfica do ribeirão João Leite, Goiás utilizando técnicas de geoprocessamento goiânia-go.
- Rocha, C. H. B. and Pereira, A. M. (2016). Análise multivariada para seleção de parâmetros de monitoramento em manancial de Juiz de Fora, Minas Gerais. *Revista Ambiente Água*, 11(1):176–187.
- Rodrigues, K. C., Rios, C., Duane, ., Barbosa, I., Oliveira, W. N. D., Nilson, ., Ferreira, C., and Kopp, K. (2013). Caracterização exploratória-espacial da bacia hidrográfica do ribeirão João Leite, Goiás, Brasil. 2:1519–8634.
- Somvanshi, M. and Chavan, P. (2016). A review of machine learning techniques using decision tree and support vector machine.
- Toledo, L. G. d. and Nicolella, G. (2002). Índice de qualidade de água em microbacia sob uso agrícola e urbano. *Scientia Agricola*, 59(1):181–186.