

## XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS

### **ANÁLISE EXPLORATÓRIA DA QUALIDADE DA ÁGUA SUPERFICIAL NA BACIA DO ALTO TIETÊ (UGRHI-6)**

*Natália de Souza Pelinson<sup>1</sup>; Lorena Ferrari Secchin<sup>2</sup>; Fernanda Santana Peiter<sup>3</sup>; Cláudia Marisse dos Santos Rotta<sup>4</sup>; Mélory Maria Fernandes Araújo<sup>5</sup>; Franciane Mendonça dos Santos<sup>6</sup>; & Ana Júlia Righetto<sup>7</sup>*

**Abstract:** In Brazil, water resource management relies on monitoring networks capable of establishing reference parameters and detecting trends in water quality and sediment load across river basins. This study analyzed a Water Resources Management Units (UGRHI) using both supervised (linear and nonlinear regressions) and unsupervised techniques (principal component analysis, clustering, and geospatial tools). Among the 22 UGRHIs in the state of São Paulo, UGRHI-6 (Alto Tietê)—the most densely populated region in the country, with over 22 million inhabitants—was selected as a case study for the application of statistical methods. Empirical analyses were based on data provided by the São Paulo State Environmental Agency (CETESB) and the National Water and Sanitation Agency (ANA). Despite fragmentation and inconsistencies in historical water quality data available through CETESB's InfoÁguas system, machine learning approaches allowed for a satisfactory delineation of the environmental characteristics of the study area. The findings underscore the potential of these analytical strategies to enhance environmental monitoring frameworks in socioeconomically complex regions, contributing to more effective water resource management.

**Keywords:** multivariate statistical analysis; surface water quality assessment; urban watershed.

**Resumo:** No Brasil, a gestão de recursos hídricos depende de redes de monitoramento capazes de estabelecer parâmetros de referência e detectar tendências na qualidade da água e no aporte de sedimentos nas bacias hidrográficas. Este estudo analisou uma unidade de gerenciamento de recursos hídricos (UGRHI) por meio de técnicas supervisionadas (regressões lineares e não lineares) e não supervisionadas (análise de componentes principais, agrupamentos e ferramentas geoespaciais). Entre as 22 UGRHIs do estado de São Paulo, a UGRHI-6 (Alto Tietê), região com a maior concentração populacional do país (mais de 22 milhões de habitantes), foi selecionada como estudo de caso para aplicação dos métodos estatísticos. As análises empíricas basearam-se em dados disponibilizados pela Companhia Ambiental do Estado de São Paulo (CETESB) e pela Agência Nacional de Águas e Saneamento Básico (ANA). Apesar da fragmentação e inconsistência nos dados históricos da qualidade da água no sistema InfoÁguas (CETESB), os métodos de aprendizado de máquina permitiram delinear de forma satisfatória as características ambientais da área avaliada. Constatou-se, portanto, o potencial dessas abordagens para apoiar o aprimoramento de estratégias de monitoramento ambiental em regiões de elevada complexidade socioeconômica, contribuindo para uma gestão mais eficaz dos recursos hídricos.

**Palavras-Chave** – análise estatística multivariada; avaliação da qualidade da água superficial; bacia hidrográfica urbana.

1) Universidade Federal de São Carlos (UFSCar-CCN Campus Lagoa do Sino); Buri – SP, Brasil; nataliasp@ufscar.br

2) Instituto Federal do Espírito Santo, Campus Ibatiba, Ibatiba – ES, Brasil; lorena.secchin@ifes.edu.br

3) Universidade Federal de Alagoas (UFAL); Maceió - AL; peiterfs@gmail.com

4) Universidade Federal de São Carlos (UFSCar-CCN Campus Lagoa do Sino); Buri - SP; claudiarotta@ufscar.br

5) Universidade Federal de São Carlos (UFSCar-CCN Campus Lagoa do Sino); Buri - SP; meloryaraujo@ufscar.br

6) Universidade Federal de São Carlos (UFSCar-CCN Campus Lagoa do Sino); Buri - SP; e-mail: francianems@ufscar.br.

7) Alvaz Agritech, Londrina 86050-268, Brazil; ajrighetto@gmail.com

## INTRODUÇÃO

O incremento das pressões antrópicas sobre os ambientes naturais tem resultado em alterações significativas na composição química das águas que circulam pelos diversos compartimentos do sistema terrestre, comprometendo sua qualidade e restringindo seu aproveitamento em múltiplos usos. Diante da crescente preocupação com a disponibilidade hídrica, tanto em termos qualitativos quanto quantitativos, observa-se uma tendência de revisão normativa e aprimoramento das práticas de conservação e de gestão integrada, com vistas à manutenção do recurso para as gerações futuras.

No contexto brasileiro, a supervisão da qualidade das águas superficiais pode ser realizada de maneira compartilhada entre diferentes esferas institucionais, conforme as demandas específicas das Unidades de Gerenciamento de Recursos Hídricos (UGRHI), que constituem a base territorial da gestão descentralizada e participativa estabelecida pela Política Nacional de Recursos Hídricos (Lei nº 9.433/1997) (Brasil, 1997).

Para que a governança dos recursos hídricos seja efetiva, é imprescindível que as informações referentes à qualidade da água estejam articuladas ao monitoramento de sua disponibilidade, uma vez que a qualidade define o nível de tratamento necessário para cada finalidade de uso. Ressalte-se que a intensificação dos conflitos decorrentes da sobreposição de demandas em um mesmo corpo hídrico permanece como um desafio atual, cuja resolução exige abordagens integradas e coordenadas. Isso reforça a centralidade da gestão eficiente e equitativa dos recursos hídricos, a fim de assegurar o atendimento simultâneo aos múltiplos usos da água.

Diante desse panorama, o presente estudo propôs uma abordagem analítica baseada na aplicação de técnicas quantitativas, incluindo métodos de aprendizado de máquina supervisionados e não supervisionados, com o intuito de gerar indicadores regionalizados condizentes com as características hidrográficas das áreas avaliadas. Como objetivo principal, buscou-se aplicar algoritmos de *Machine Learning* (ML) em conjunto com análises geoespaciais, visando otimizar as estratégias de monitoramento da qualidade da água em sistemas fluviais brasileiros, tendo como área de estudo a Unidade de Gerenciamento de Recursos Hídricos do Alto Tietê (UGRHI-06).

## METODOLOGIA

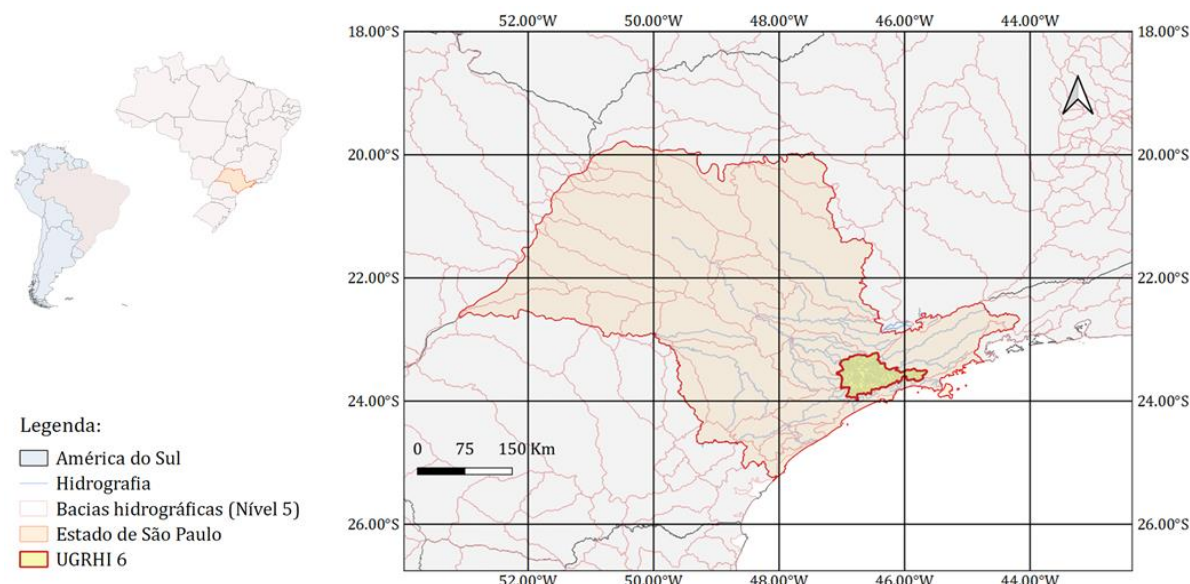
A área de estudo corresponde à Unidade de Gerenciamento de Recursos Hídricos do Alto Tietê (UGRHI-06), situada no estado de São Paulo, Brasil (Figura 1). Esta unidade compreende uma região estratégica por cobrir majoritariamente a Região Metropolitana de São Paulo (RMSP), o principal núcleo urbano do país em termos populacionais e econômicos. A bacia abrange ainda o Rio Tietê, um dos cursos d'água mais impactados do Brasil, historicamente comprometido pelo lançamento de efluentes domésticos e industriais provenientes da RMSP.

A UGRHI-06 é composta por 34 municípios paulistas, dentre eles São Paulo, Guarulhos, Osasco, Santo André e Mogi das Cruzes, englobando tanto áreas urbanas densamente povoadas quanto regiões periféricas sujeitas à ocupação irregular. A complexidade territorial e o uso intensivo dos corpos hídricos refletem-se na degradação da qualidade da água, acentuada pela descarga de esgotos tratados e não tratados em rios e córregos da região.

Para análise da qualidade da água, foram consideradas as diretrizes estabelecidas pela Resolução CONAMA nº 357/2005 (BRASIL, 2005), que classifica os corpos hídricos segundo o uso preponderante e define parâmetros específicos para cada classe (especial, 1, 2, 3 ou 4). Esses

parâmetros são utilizados na construção de índices de qualidade da água, que sintetizam múltiplas variáveis físico-químicas e microbiológicas.

**Figura 1.** Localização da área de estudo no estado de São Paulo, Brasil: Bacia do Alto Tietê (UGRHI-6).



**Fonte:** Autoria própria.

Os dados utilizados nesta pesquisa foram obtidos em plataformas públicas: o sistema InfoÁguas da Companhia Ambiental do Estado de São Paulo (CETESB) e o Atlas de Esgoto da Agência Nacional de Águas e Saneamento Básico (ANA, 2020). Para fins comparativos e integrativos, adotou-se o Índice de Qualidade da Água (IQA), calculado conforme metodologia da CETESB. Este índice é composto por nove parâmetros: pH, coliformes termotolerantes, demanda bioquímica de oxigênio (DBO), nitrogênio total, fósforo total, temperatura da água, turbidez, resíduos totais e oxigênio dissolvido. A fórmula utilizada está expressa na Equação 1:

$$IQA = \prod_{i=1}^n qi^{wi} \quad (\text{Equação 1})$$

Em que:

$qi$  é a concentração ou leitura ajustada da “curva média de variação de qualidade”, valores entre 0 a 100;

$wi$  representa o peso atribuído ao parâmetro em função da sua importância para o valor final de qualidade, então a soma total do peso de cada parâmetro é 1 (ou seja, 100%);

$n$  é o número de variáveis que entram no cálculo de IQA, ou seja, nove variáveis quantitativas.

Dada a complexidade na interpretação simultânea de múltiplas variáveis ambientais, técnicas de ciência de dados foram incorporadas como estratégia complementar de análise. Assim, a metodologia envolveu duas etapas principais: (i) análises geoespaciais com base em sistemas de informações geográficas (SIG) e (ii) aplicação de algoritmos de aprendizado de máquina (*Machine Learning*).



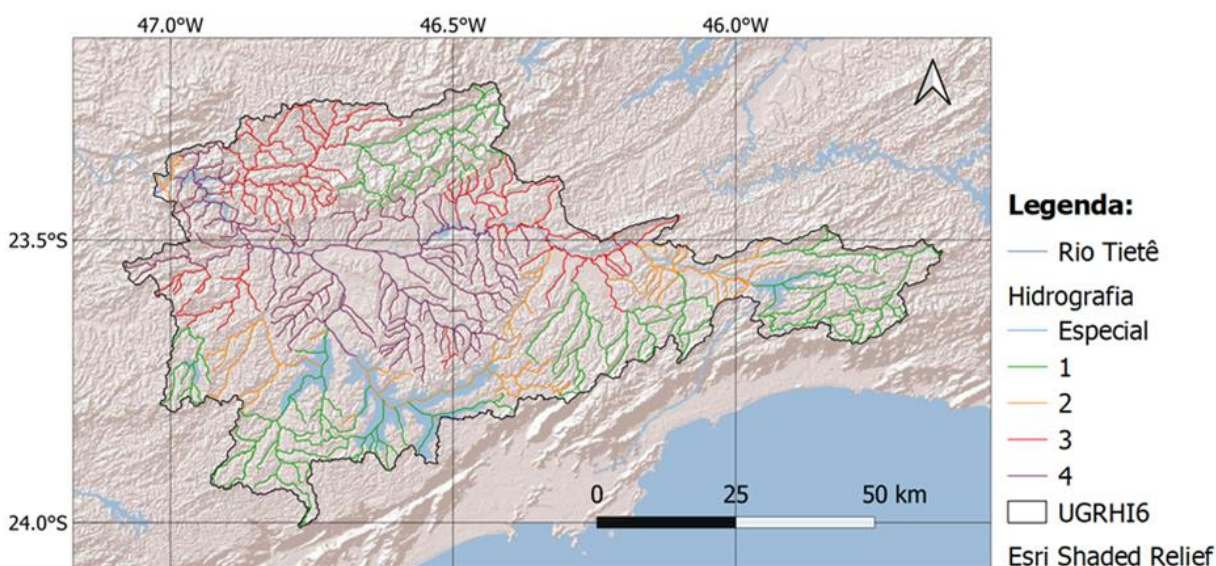
Foram empregados modelos supervisionados e não supervisionados. Entre os primeiros, destacam-se os Modelos Lineares Generalizados (GLM) e modelos não lineares. Entre os métodos não supervisionados, utilizaram-se o agrupamento hierárquico aglomerativo (AHC), o agrupamento por K-means, e a Análise de Componentes Principais (ACP), utilizada também como estratégia de redução dimensional para os modelos GLM. A ACP permitiu resumir a variabilidade dos dados ambientais mantendo sua estrutura informacional essencial. Para desenvolvimento dos códigos foram utilizadas as linguagens Python e R, usando a biblioteca scikit-learn (e suas inclusões: numpy, scipy e matplotlib) e os pacotes estatísticos corrr, factoMineR, factoextra e ggcorrplot, respectivamente.

## RESULTADOS E DISCUSSÃO

As análises desta pesquisa basearam-se em dados da Companhia Ambiental do Estado de São Paulo (via InfoÁguas) e da ANA, com foco na UGRHI-06 (Alto Tietê), região de elevada complexidade ambiental. O avanço tecnológico recente viabiliza o uso de algoritmos de aprendizado de máquina (ML, de *Machine Learning*) para a interpretação desses dados.

Inicialmente, aplicou-se a classificação das águas superficiais segundo a Resolução CONAMA nº 357/2005 (BRASIL, 2005), evidenciando que os trechos com piores classes coincidem com áreas densamente urbanizadas — especialmente a Região Metropolitana de São Paulo (Figura 2), com seus corpos hídricos superficiais identificados como Classe 4 (coloração roxa).

**Figura 2** – Classificação da qualidade das águas superficiais na UGRHI Alto Tietê, com base na Resolução CONAMA nº 357/2005, a partir de dados da Agência Nacional de Águas & Saneamento (ANA) e do Sistema Ambiental Paulista (Datageo).

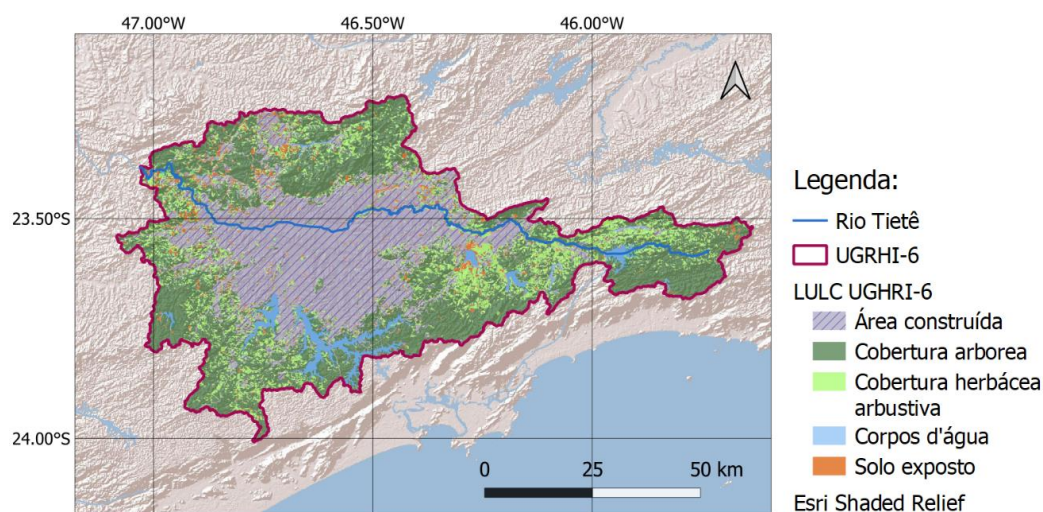


Tal padrão reforça a correlação entre urbanização desordenada, deficiência em saneamento e degradação hídrica (Costa *et al.*, 2021; Galinaro *et al.*, 2021). O uso da classificação da Secretaria Estadual do Meio Ambiente (SMA-SP) permitiu mapear a cobertura da terra, destacando a redução da permeabilidade em zonas urbanizadas (Figura 3).

Com base em séries temporais disponibilizadas pela Companhia Ambiental do Estado de São Paulo (CETESB, 2021), referentes ao período de 2016 a 2020, foi realizada uma Análise de Componentes Principais (ACP) aplicada a um conjunto de nove variáveis de qualidade da água selecionadas por meio de procedimento automatizado de exclusão progressiva. Embora inicialmente

tenham sido considerados onze parâmetros, o carbono orgânico total (COT) foi descartado por ausência de dados anteriores a 2020, enquanto o pH foi omitido em razão da ampla tolerância normativa, que o torna pouco discriminativo em análises multivariadas, mesmo sob condições de impacto ambiental.

Figura 3 - Mapeamento da cobertura do solo na bacia hidrográfica do Alto Tietê, evidenciando cinco classes principais: área construída, vegetação arbórea, vegetação herbácea-arbustiva, corpos d'água e solos expostos, com base na classificação da Secretaria Estadual do Meio Ambiente (SMA-SP).



A composição final dos dados para a Análise das Componentes Principais (ACP) incluiu os seguintes indicadores: condutividade elétrica (EC), turbidez, nitrato ( $\text{N-NO}_3$ ), nitrogênio amoniacal ( $\text{NH}_4^+$ ), oxigênio dissolvido (OD), demanda bioquímica de oxigênio (DBO), fósforo total, *Escherichia coli* (E. coli) e clorofila-a. Essas variáveis representam, os principais índices de qualidade da água utilizados no Brasil, desenvolvidos pela CETESB e pela ANA.

A ACP indicou que os dois primeiros componentes principais (F1 e F2) explicam mais de 70% da variância total dos dados (Figura 4), conforme verificado pelo gráfico de autovalores (*scree plot* da Tabela 1). As cargas fatoriais atribuídas a cada variável refletem sua contribuição para a estrutura latente dos dados, sendo fundamentais para a interpretação ambiental dos padrões observados.

Figura 4 - Gráfico de *scree plot* dos componentes principais, indicando a variância acumulada dos fatores extraídos da ACP para dados de qualidade da água (2016–2020).

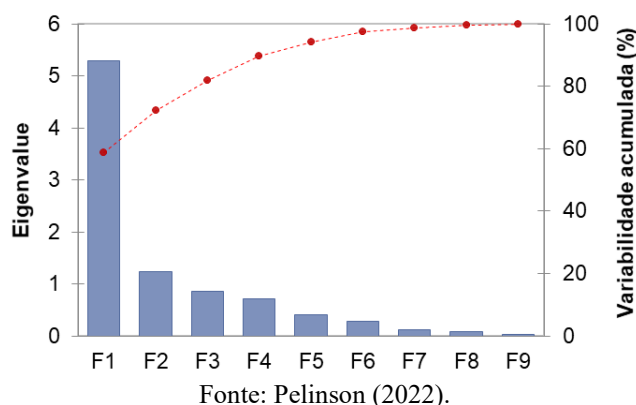


Tabela 1 - Cossenos ao quadrado ( $\cos^2$ ) das variáveis ambientais para os cinco primeiros componentes extraídos da Análise de Componentes Principais (ACP). Em negrito, os fatores com maior representatividade para cada variável, indicando sua contribuição dominante na estrutura fatorial.

Parâmetro	F1	F2	F3	F4	F5
Condutividade	<b>0.682</b>	0.033	0.015	0.086	0.079
Turbidez	<b>0.422</b>	0.085	0.035	0.323	0.120
Nitrato	0.018	<b>0.511</b>	0.433	0.027	0.006
N-Amoniacal	<b>0.933</b>	0.000	0.003	0.000	0.005
OD	<b>0.621</b>	0.087	0.043	0.018	0.086
DBO	<b>0.860</b>	0.002	0.017	0.039	0.005
P-Total	<b>0.903</b>	0.000	0.011	0.042	0.000
E. coli	<b>0.715</b>	0.000	0.028	0.172	0.026
Clorofila a	0.127	<b>0.511</b>	0.267	0.005	0.075

Fonte: Pelinson (2022).

As correlações de Pearson e Spearman (Tabelas 2 e 3) mostraram resultados convergentes, indicando correlação negativa entre oxigênio dissolvido (OD) e indicadores de contaminação por esgoto e matéria orgânica, como demanda bioquímica de oxigênio (DBO) e *E. coli*.

Tabela 2 - Matriz de correlação de Pearson entre os principais parâmetros físico-químicos e biológicos mensurados nas águas superficiais da UGRHI Alto Tietê, considerando linearidade nas relações.

Variáveis	Condutividade	Turbidez	Nitrato	N-Amoniacal	OD	DBO	P-Total	E. coli	Clorofila-a
Condutividade	<b>1.0</b>	<b>0.6</b>	<b>0.2</b>	<b>0.8</b>	<b>-0.6</b>	<b>0.7</b>	<b>0.7</b>	<b>0.5</b>	<b>-0.1</b>
Turbidez	<b>0.6</b>	<b>1.0</b>	<b>0.1</b>	<b>0.6</b>	<b>-0.4</b>	<b>0.5</b>	<b>0.5</b>	<b>0.4</b>	<b>0.0</b>
Nitrato	<b>0.2</b>	<b>0.1</b>	<b>1.0</b>	<b>0.1</b>	<b>0.0</b>	<b>0.1</b>	<b>0.1</b>	<b>0.1</b>	<b>0.2</b>
N-Amoniacal	<b>0.8</b>	<b>0.6</b>	<b>0.1</b>	<b>1.0</b>	<b>-0.7</b>	<b>0.9</b>	<b>0.9</b>	<b>0.8</b>	<b>-0.2</b>
OD	<b>-0.6</b>	<b>-0.4</b>	<b>0.0</b>	<b>-0.7</b>	<b>1.0</b>	<b>-0.6</b>	<b>-0.7</b>	<b>-0.5</b>	<b>0.4</b>
DBO	<b>0.7</b>	<b>0.5</b>	<b>0.1</b>	<b>0.9</b>	<b>-0.6</b>	<b>1.0</b>	<b>0.9</b>	<b>0.9</b>	<b>-0.1</b>
P-Total	<b>0.7</b>	<b>0.5</b>	<b>0.1</b>	<b>0.9</b>	<b>-0.7</b>	<b>0.9</b>	<b>1.0</b>	<b>0.9</b>	<b>-0.2</b>
E. coli	<b>0.5</b>	<b>0.4</b>	<b>0.1</b>	<b>0.8</b>	<b>-0.5</b>	<b>0.9</b>	<b>0.9</b>	<b>1.0</b>	<b>-0.2</b>
Clorofila-a	<b>-0.1</b>	<b>0.0</b>	<b>0.2</b>	<b>-0.2</b>	<b>0.4</b>	<b>-0.1</b>	<b>-0.2</b>	<b>-0.2</b>	<b>1.0</b>

Os valores em negrito são diferentes de 0 com um nível de significância  $\alpha=0,05$

Fonte: Pelinson (2022).

Tabela 2 - Matriz de correlação de Spearman para os mesmos parâmetros da Tabela 2, considerando relações monotônicas, incluindo variáveis com comportamento não linear.

Variáveis	Condutividade	Turbidez	Nitrato	N-Amoniacal	OD	DBO	P-Total	E. coli	Clorofila-a
Condutividade	<b>1.0</b>	<b>0.8</b>	<b>0.2</b>	<b>0.9</b>	<b>-0.8</b>	<b>0.9</b>	<b>0.9</b>	<b>0.9</b>	<b>0.0</b>
Turbidez	<b>0.8</b>	<b>1.0</b>	<b>0.2</b>	<b>0.8</b>	<b>-0.6</b>	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>	<b>-0.2</b>
Nitrato	<b>0.2</b>	<b>0.2</b>	<b>1.0</b>	<b>0.1</b>	<b>0.2</b>	<b>0.1</b>	<b>0.1</b>	<b>0.1</b>	<b>0.1</b>
N-Amoniacal	<b>0.9</b>	<b>0.8</b>	<b>0.1</b>	<b>1.0</b>	<b>-0.8</b>	<b>0.9</b>	<b>0.9</b>	<b>0.9</b>	<b>-0.3</b>
OD	<b>-0.8</b>	<b>-0.6</b>	<b>0.2</b>	<b>-0.8</b>	<b>1.0</b>	<b>-0.8</b>	<b>-0.8</b>	<b>-0.8</b>	<b>0.5</b>
DBO	<b>0.9</b>	<b>0.8</b>	<b>0.1</b>	<b>0.9</b>	<b>-0.8</b>	<b>1.0</b>	<b>1.0</b>	<b>0.9</b>	<b>0.0</b>
P-Total	<b>0.9</b>	<b>0.8</b>	<b>0.1</b>	<b>0.9</b>	<b>-0.8</b>	<b>1.0</b>	<b>1.0</b>	<b>0.9</b>	<b>-0.2</b>
E. coli	<b>0.9</b>	<b>0.8</b>	<b>0.1</b>	<b>0.9</b>	<b>-0.8</b>	<b>0.9</b>	<b>0.9</b>	<b>1.0</b>	<b>-0.4</b>
Clorofila-a	<b>0.0</b>	<b>-0.2</b>	<b>0.1</b>	<b>-0.3</b>	<b>0.5</b>	<b>0.0</b>	<b>-0.2</b>	<b>-0.4</b>	<b>1.0</b>

Os valores em negrito são diferentes de 0 com um nível de significância  $\alpha=0,05$

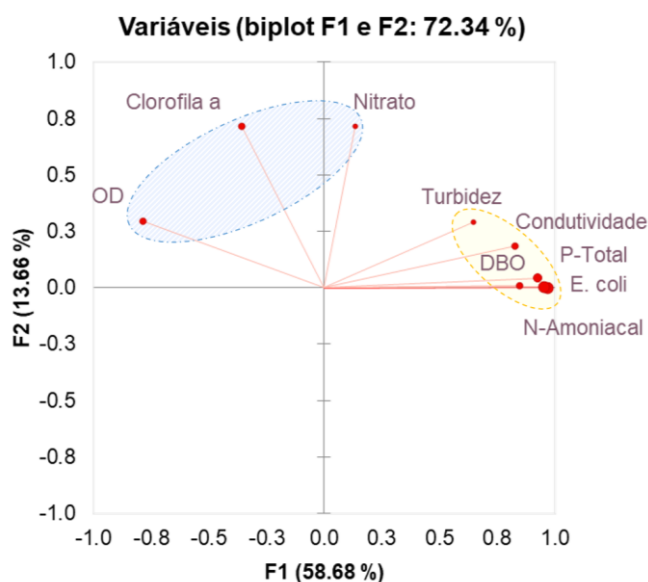
Fonte: Pelinson (2022).



A clorofila-a, apesar de indicar eutrofização, apresentou associação positiva com OD, devido à produção fotossintética de oxigênio esperada. Houve forte correlação positiva entre condutividade elétrica (EC), turbidez, nitrogênio amoniacal ( $\text{NH}_4^+$ ), DBO, fósforo total e E. coli, refletindo possível origem comum dos contaminantes.

A ACP evidenciou dois agrupamentos (Figura 5): o cluster amarelo, mais coeso (similaridade acima de 80%), abrangeu as variáveis com alta correlação e foi explorado por modelos lineares generalizados (LR). O cluster azul, menos coeso, mostrou baixa representatividade nos modelos (abaixo de 40%). Mesmo entre variáveis com correlação negativa, como OD e DBO, os ajustes dos modelos lineares foram limitados ( $R^2 \approx 45\%$ ).

Figura 5 - Distribuição das variáveis ambientais em análise de componentes principais (ACP), com destaque para dois agrupamentos formados por variáveis com padrões similares de correlação.



Fonte: Pelinson (2022).

A PCA revelou que 72% da variabilidade total dos dados é explicada pelos dois primeiros eixos fatoriais, o que justificou a separação das variáveis em dois agrupamentos distintos. O primeiro grupo, mais homogêneo, reúne variáveis com fortes correlações positivas entre si, demonstrando maior coerência interna — o que facilitou sua representação gráfica e posterior exploração estatística por meio de modelos lineares generalizados (GLM). Esse conjunto concentrou alta similaridade interna, superando 80%. Em contraste, o segundo grupo exibiu dispersão mais acentuada entre as variáveis, resultando em baixa coerência estrutural e dificuldade de ajuste por modelos estatísticos, com coeficientes de determinação inferiores a 40%.

Apesar da evidência de associação inversa entre oxigênio dissolvido (OD) e a demanda bioquímica de oxigênio (DBO) nas matrizes de correlação, os modelos estatísticos não refletiram esse comportamento com precisão. Tanto os ajustes lineares quanto os não-paramétricos resultaram em  $R^2$  máximos em torno de 45%, indicando limitada capacidade explicativa.

A avaliação da fração orgânica em águas superficiais costuma ser realizada com base na DBO, na demanda química de oxigênio (DQO) e no carbono orgânico total (COT). Contudo, a DQO — por envolver procedimentos de acidificação em condições extremas — não reflete adequadamente o comportamento natural da matéria orgânica, especialmente em ambientes aquáticos com pH entre 5 e 9. Além disso, esse parâmetro não tem sido amplamente incorporado aos relatórios de

monitoramento da CETESB. Já o COT, introduzido mais recentemente, foi incluído na análise apenas no ano de 2020, único com dados disponíveis.

Nesse contexto, foram construídos três modelos lineares principais: o primeiro relacionando DBO e COT (Fig. 6), o segundo entre fósforo total (Fig 7) e a condutividade elétrica e o terceiro entre DBO e *Escherichia coli* como bioindicador de esgoto doméstico (Fig. 7). A relação DBO versus COT apresentou elevada capacidade explicativa, coerente com o fato de ambos os indicadores representarem frações da matéria orgânica passível de biodegradação em condições naturais. Modelos de regressão simples e múltipla foram aplicados para variáveis indicativas de matéria orgânica, com destaque para a elevada associação entre DBO e COT (Figura 6), explicando quase a totalidade da variabilidade, e para a relação entre fósforo e EC (Figura 7

Figura 6 - Relação linear entre DBO e carbono orgânico total (COT), com alto poder preditivo para quantificação da fração biodisponível da matéria orgânica.

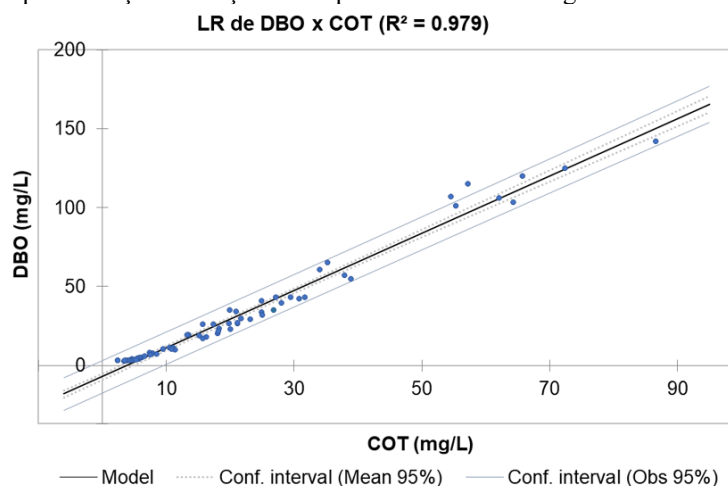
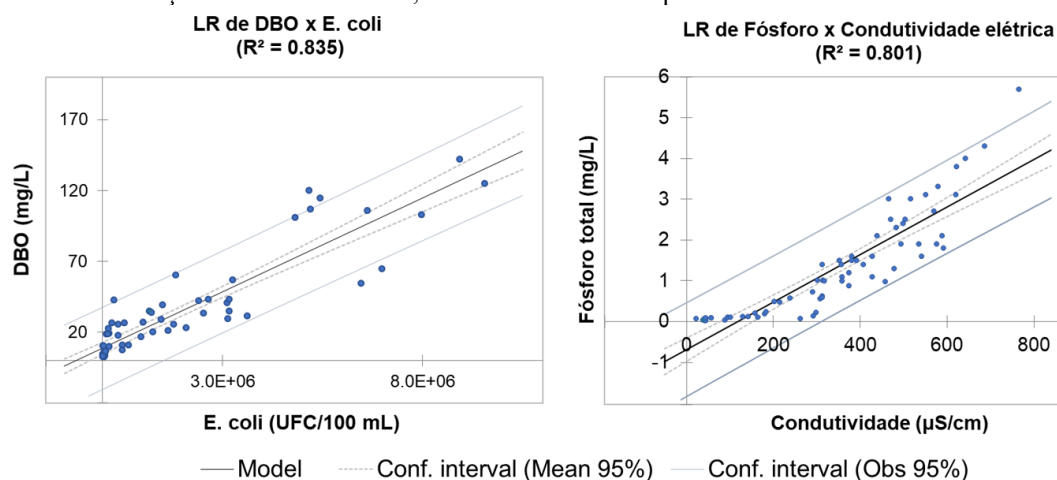


Figura 7 - Modelos lineares representando: (i) fósforo total em função da condutividade elétrica e (ii) DBO em função de *Escherichia coli*, ambos com elevada representatividade estatística.



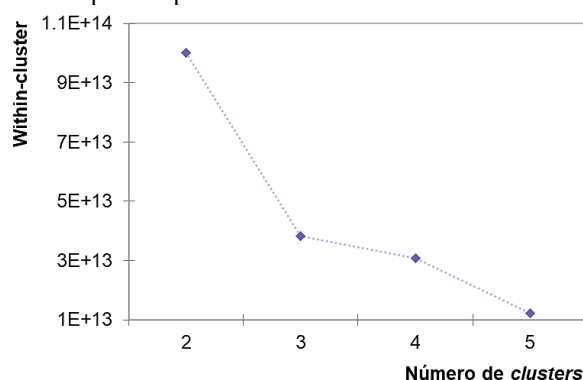
Complementarmente, a clusterização por *k-means* corroborou a divisão em dois grupos, acrescentando pontos de amostragem ao agrupamento definido por AHC (Figura 8). Para representar espacialmente os agrupamentos obtidos pelo método *k-means*, foi elaborado um mapa com a



distribuição das amostras (Figura 9). A região central da bacia, onde os Índices de Qualidade da Água (IQAs) são mais baixos, apresenta eficiência intermediária na remoção de carga orgânica (entre 50% e 75%). Em contraste, sete municípios — São Roque, Paraibuna, Cajamar, Francisco Morato, Franco da Rocha, Caieiras e Itaquaquecetuba — registram remoções inferiores a 5%, apesar de gerarem esgotos significativos e não estarem incluídos na rede estadual de monitoramento.

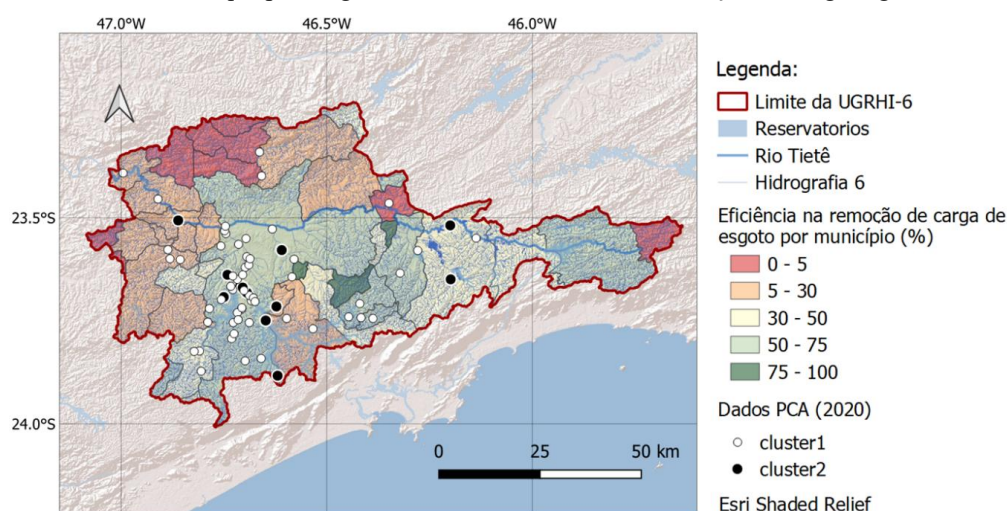
Observa-se que os agrupamentos não apresentam padrão regional claramente definido, tampouco evidências consistentes que orientem sua otimização prática. Ainda assim, destaca-se certa coerência nas medidas associadas aos reservatórios Guarapiranga e Billings, ambos estratégicos para o abastecimento da Região Metropolitana de São Paulo.

Figura 8 - Evolução da inércia intragrupo no método *k-means*, indicando a melhor performance com dois agrupamentos para os pontos de monitoramento analisados.



Fonte: Pelinson (2022).

Figura 9 - Mapa dos agrupamentos gerados por *k-means*, com distribuição espacial dos pontos monitorados por *cluster* e destaque para regiões com baixa eficiência de remoção de carga orgânica.



Fonte: Pelinson (2022).

O uso de técnicas de aprendizado de máquina para definição de comportamentos de parâmetros de qualidade da água pode contribuir significativamente para otimizar o processo de monitoramento. A partir da identificação de padrões e relações entre variáveis por meio de análises como correlação de Spearman e Pearson, além de métodos de redução de dimensionalidade como a Análise de Componentes Principais (PCA), torna-se possível selecionar os parâmetros mais representativos. Isso permite reduzir o número de variáveis a serem monitoradas sem comprometer a

representatividade dos dados. Modelos preditivos baseados nesses parâmetros-chave podem ser aplicados para estimar valores faltantes em séries temporais, assim como prever a qualidade da água em intervalos mais curtos. Essa abordagem viabiliza o aumento da frequência das coletas com menor custo operacional. Recomenda-se a validação dos modelos por meio de novas campanhas de amostragem, integrando os dados atuais com os históricos para garantir robustez e confiabilidade nas previsões.

## CONSIDERAÇÕES FINAIS

A análise da UGRHI-06 demonstrou forte associação entre urbanização desordenada e degradação da qualidade da água, especialmente nas áreas densamente ocupadas da Região Metropolitana de São Paulo. A classificação por classes de uso e o mapeamento da cobertura do solo confirmaram que os trechos mais críticos concentram os menores índices de qualidade e maior comprometimento ambiental. A aplicação da Análise de Componentes Principais (ACP) permitiu identificar dois agrupamentos de variáveis. O primeiro, mais coeso, apresentou alto poder explicativo nos modelos lineares, com destaque para a relação entre DBO e COT. Já o segundo grupo mostrou baixa correlação entre variáveis e desempenho estatístico limitado. A condutividade elétrica demonstrou potencial como indicador substituto em modelos de predição de matéria orgânica. Por fim, os agrupamentos por *k-means* reforçaram a separação identificada na ACP e evidenciaram municípios com baixa eficiência na remoção de carga poluidora, muitos dos quais não estão cobertos pelo monitoramento estadual. Os resultados indicam que modelos simples de aprendizado de máquina podem aprimorar o diagnóstico ambiental e orientar ações corretivas em áreas prioritárias.

## REFERÊNCIAS

- ANA - Agência Nacional de Águas e Saneamento (Brasil). *Atlas esgotos: despoluição de bacias hidrográficas: atualização*, Secretaria Nacional de Saneamento Ambiental. Brasília-DF: ANA, 2020. Disponível em: <http://atlasesgotos.ana.gov.br>.
- BRASIL (2005) *Resolução do Conselho Nacional do Meio Ambiente (CONAMA) N° 357, 2005*. Classificação dos corpos de água. Disponível em: [https://www.icmbio.gov.br/cepsul/images/stories/legislacao/Resolucao/2005/res\\_conama\\_357\\_2005\\_classificacao\\_corpos\\_agua\\_rtfcd\\_altrd\\_res\\_393\\_2007\\_397\\_2008\\_410\\_2009\\_430\\_2011.pdf](https://www.icmbio.gov.br/cepsul/images/stories/legislacao/Resolucao/2005/res_conama_357_2005_classificacao_corpos_agua_rtfcd_altrd_res_393_2007_397_2008_410_2009_430_2011.pdf).
- BRASIL (1997) *Lei Federal N° 9.433, 1997. Política Nacional de Recursos Hídricos e criação do Sistema Nacional de Gerenciamento de Recursos Hídricos*. Disponível em: [http://www.planalto.gov.br/ccivil\\_03/leis/19433.htm](http://www.planalto.gov.br/ccivil_03/leis/19433.htm). Acesso em 09 nov. 2021.
- COSTA, D.A.; SILVA JUNIOR, L.C.S.; AZEVEDO, J.P.S.; SANTOS, M.A.; ASSUMPÇÃO, R.S.F.V. From Monitoring and Modeling to Management: How to Improve Water Quality in Brazilian Rivers? A Case Study: Piabanha River Watershed. *Water* 2021, 13, 176. Disponível em: <https://doi.org/10.3390/w13020176>.
- GALINARO, C. A.; SPADOTO, M.; AQUINO, F. W. B.; PELINSON, N. S.; VIEIRA, E. M. Environmental risk assessment of parabens in surface water from a Brazilian river: the case of Mogi Guaçu Basin, São Paulo State, under precipitation anomalies. *Environmental Science and Pollution Research*, v. 29, p. 8816-8830, 2022.
- PELINSON, N. P. *Ciência de dados aplicada à caracterização ambiental de bacias hidrográficas brasileira: um estudo de caso*. Trabalho de Conclusão de Curso de especialização MBA em Data Science e Analytics. Programa de Educação Continuada em Economia e Gestão de Empresas [PECEGE]; Universidade de São Paulo – Piracicaba, SP. 2022. 24 p.