

Disponibilidade hídrica superficial do bioma Caatinga na Região Hidrográfica do São Francisco

Fernanda Santana Peiter^{1*}; Ítalo de Oliveira Braga²

¹* Universidade Federal de Alagoas. Doutora em Recursos Hídricos e Saneamento. Av. Lourival Melo Mota, S/N – Tabuleiro do Martins; 57072-970, Maceió, Alagoas, Brasil

² Universidade Federal de Lavras. Mestre em Biotecnologia Vegetal. Empresa Brasileira de Pesquisas Agropecuárias (Embrapa) – Asa Norte; 70770901, Brasília, DF, Brasil

*autor correspondente: peiterfs@gmail.com

Disponibilidade hídrica superficial do bioma Caatinga na Região Hidrográfica do São Francisco

Resumo

O Rio São Francisco e seus afluentes desempenham um papel crucial como fontes de recursos hídricos na região da Caatinga. No entanto, a degradação ambiental desse bioma pode comprometer a disponibilidade de água, especialmente diante das condições climáticas adversas, caracterizadas por longos períodos de estiagem. Este estudo teve como objetivo investigar o impacto dos fatores climáticos e do uso e cobertura do solo nas vazões de referência Q95 na porção da Caatinga dentro da Bacia Hidrográfica do São Francisco. Utilizou-se a análise por componentes principais (PCA) para explorar possíveis relações entre as variáveis explicativas. Além disso, foram aplicados quatro modelos supervisionados de machine learning: Regressão Linear Múltipla, Regressão Não Linear (com Transformação de Box-Cox), Árvore de Regressão e Random Forest. Os resultados revelaram a significância estatística de todas as variáveis explicativas. No entanto, os modelos linear e não linear apresentaram desafios relacionados à não aderência à normalidade dos termos de erro, heterocedasticidade e coeficientes de regressão relativamente baixos (0,33 e 0,35, respectivamente). Embora os modelos de árvore ($R^2 = 0,68$) e Random Forest ($R^2 = 0,80$) tenham demonstrado ajustes mais satisfatórios, ainda não foram considerados ideais devido à presença de diversas fontes potenciais de erros e incertezas nas análises. Por fim, destaca-se a disparidade nas vazões entre o Rio São Francisco e os demais cursos d'água na região, ressaltando a importância da conservação desse rio para a sustentabilidade das comunidades locais e a preservação do bioma Caatinga.

Palavras-chave: segurança hídrica; região semiárida; variáveis climáticas; uso e cobertura do solo; machine learning.

Abstract

The São Francisco River and its tributaries play a crucial role as water sources in the Caatinga region. However, environmental degradation of this biome can jeopardize water availability, especially under adverse climatic conditions characterized by long periods of drought. This study aimed to investigate the impact of climatic factors and land use and cover on the Q95 reference flows in the Caatinga portion within the São Francisco River Basin. Principal component analysis (PCA) was used to explore possible relationships between explanatory variables. Additionally, four supervised machine learning models were applied: Multiple Linear Regression, Nonlinear Regression (with Box-Cox Transformation), Regression Tree, and Random Forest. The results revealed the statistical significance of all explanatory variables. However, both linear and nonlinear models faced challenges related to non-compliance with error term normality, heteroscedasticity, and relatively low regression coefficients (0.33 and 0.35, respectively). Although the tree models ($R^2 = 0.68$) and Random Forest ($R^2 = 0.80$) demonstrated more satisfactory adjustments, they were still not considered ideal due to the presence of various potential sources of errors and uncertainties in the analyses. Finally, the discrepancy in flows between the São Francisco River and other watercourses in the region is highlighted, underscoring the importance of conserving this river for the sustainability of local communities and the preservation of the Caatinga biome.

Keywords: water security; semiarid region; climate variables; land use and cover; machine learning.

1 Introdução

A Caatinga, único bioma exclusivo do Brasil, é encontrada apenas nos estados do Nordeste e no norte de Minas Gerais, em regiões de clima semiárido, caracterizado por altas temperaturas e longos períodos de estiagem (Freire et al., 2018). Esse ecossistema apresenta condições climáticas severas e baixos índices de disponibilidade hídrica, que tendem a ser agravados pelo manejo insustentável de seus recursos, ameaçando um dos principais cursos d'água da região: o Rio São Francisco.

A Caatinga é detentora de uma fauna e flora bem diversificada que abriga cerca de aproximadamente 1.182 espécies de animais e 4.963 vegetais, incluindo plantas cactáceas, arbustivas, arbóreas de pequeno porte, bromélias e herbáceas (Drumond et al., 2016; MMA, 2022). Entretanto, a conservação dessa biodiversidade é dificultada pela expansão urbana desordenada, crescimento descontrolado da agricultura e pecuária, desmatamento, mineração, incêndios florestais e pela caça e pesca predatórias (Drumond, 2004; Freire et al., 2018).

Ao avaliarem a dinâmica espacial da cobertura vegetal da Caatinga entre 2013 e 2015, Silva et al. (2020) observaram que houve o crescimento de atividades agrossilvopastoris associado à perda da vegetação nativa, com redução de aproximadamente 50% das áreas cobertas por corpos d'água. Os autores ressaltam que os efeitos severos da seca e o uso indiscriminado do solo ao longo do tempo resultaram no aumento das áreas de solo expostas e do déficit hídrico.

Em geral, as áreas vegetadas exercem papel fundamental no ciclo da água, visto que atuam como reservatórios e filtros naturais, além de fornecer umidade atmosférica através da evapotranspiração, elevando potencialmente a ocorrência de chuvas. Contudo, o desflorestamento diminui a capacidade de infiltração da água nos solos e a troca de umidade com a atmosfera, aumenta o risco de erosão e o carreamento de sedimentos, eleva o escoamento superficial causando enchentes e inundações, além de outros problemas que podem comprometer o acesso à água limpa (Mapulanga e Naito, 2019). Sendo assim, o processo de degradação ambiental que atinge a Caatinga interfere também na dinâmica hidrológica de sua área de abrangência, afetando a qualidade e a quantidade dos seus recursos hídricos.

O Rio São Francisco e seus afluentes, importantes fontes de subsistência para a população em seu entorno, têm sido deteriorados pela existência de fontes pontuais e difusas de poluição, pelo assoreamento e pelo avanço da cunha salina. Ademais, o São Francisco comporta um complexo hidroelétrico responsável pelo abastecimento de energia elétrica na região, o que provoca variações constantes em sua vazão (Amorim et al., 2022). Portanto, observa-se que as atividades antrópicas diretas, aliadas às mudanças climáticas globais,

podem acentuar ainda mais a vulnerabilidade dos corpos hídricos presentes na Caatinga, intensificando o risco de escassez e de desertificação (Azevêdo et al., 2017).

Diante dessa problemática, a investigação dos fatores que afetam a disponibilidade de água pode auxiliar os tomadores de decisão da gestão em recursos hídricos na manutenção da segurança hídrica. Nesse sentido, inúmeros estudos buscam estimar a disponibilidade hídrica superficial a partir da análise das vazões de referência em rios, utilizando diferentes métodos de modelagem, a depender da existência de dados medidos (Collischonn et al., 2023).

Modelos físicos são amplamente empregados em estimativas de vazão; porém, demandam uma quantidade substancial de dados, os quais podem ser dispendiosos e escassos, especialmente em regiões em desenvolvimento. Recentemente, o uso de técnicas de *Machine Learning* tem demonstrado notável eficácia e praticidade em comparação com os modelos convencionais, devido à sua habilidade de processar múltiplas fontes de dados simultaneamente e à sua menor exigência de custo e tempo (Ahmed et al., 2024).

Entretanto, mesmo diante dos avanços tecnológicos, a determinação das vazões ainda consiste em um dos principais desafios das ciências hidrológicas, visto que envolvem diversas variáveis não controláveis como a temperatura, a precipitação, o relevo, o tipo de solo e seus usos (Zhang e Wei, 2021). Além disso, o estudo de Impacto da Mudança Climática nos Recursos Hídricos do Brasil, publicado pela Agência Nacional de Águas (ANA, 2024) ressalta que as mudanças climáticas e os padrões de consumo de água pela sociedade também adicionam incertezas às análises dos processos hidrometeorológicos.

Deste modo, considerando a crescente supressão da vegetação da Caatinga e fragilidade de seus recursos hídricos, é importante investigar os diferentes fatores que impactam sua disponibilidade hídrica superficial. Sendo assim, o objetivo desse estudo foi avaliar o desempenho de alguns modelos de *Machine Learning* na predição de vazões de referência, especificamente na porção do bioma que compreende a Bacia Hidrográfica do São Francisco. A abordagem adotada, usando técnicas supervisionadas e não-supervisionadas de aprendizado de máquina, pode auxiliar na compreensão das possíveis relações entre as variáveis estudadas e como estas podem interferir na predição das respostas hidrológicas na região.

2 Material e Métodos

A Figura 1 apresenta um esquema simplificado das etapas do trabalho.

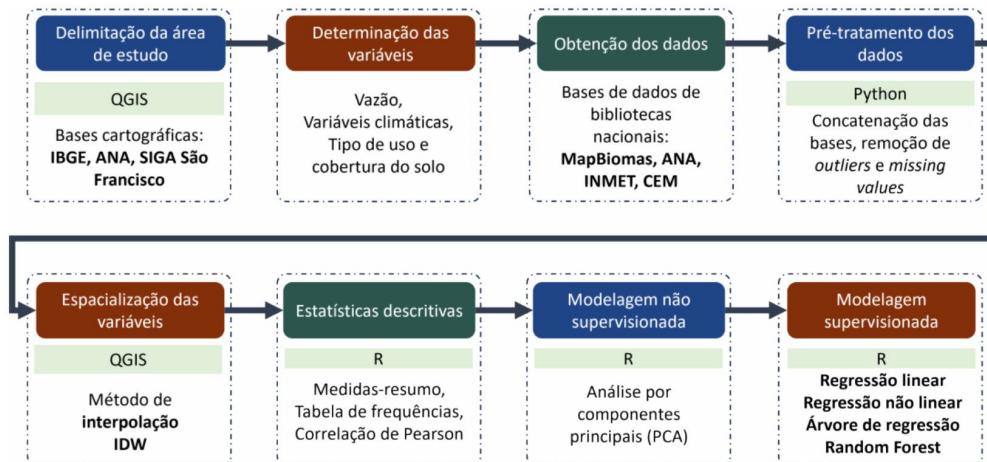


Figura 1. Etapas de desenvolvimento da pesquisa

Fonte: Autoria própria (2024)

2.1 Delimitação da área de estudo

Esse estudo abrangeu a porção do bioma Caatinga localizada na Região Hidrográfica do São Francisco (RH-SF) (Figura 2), englobando quatro sub-regiões (Alto, Médio, Submédio e Baixo São Francisco).

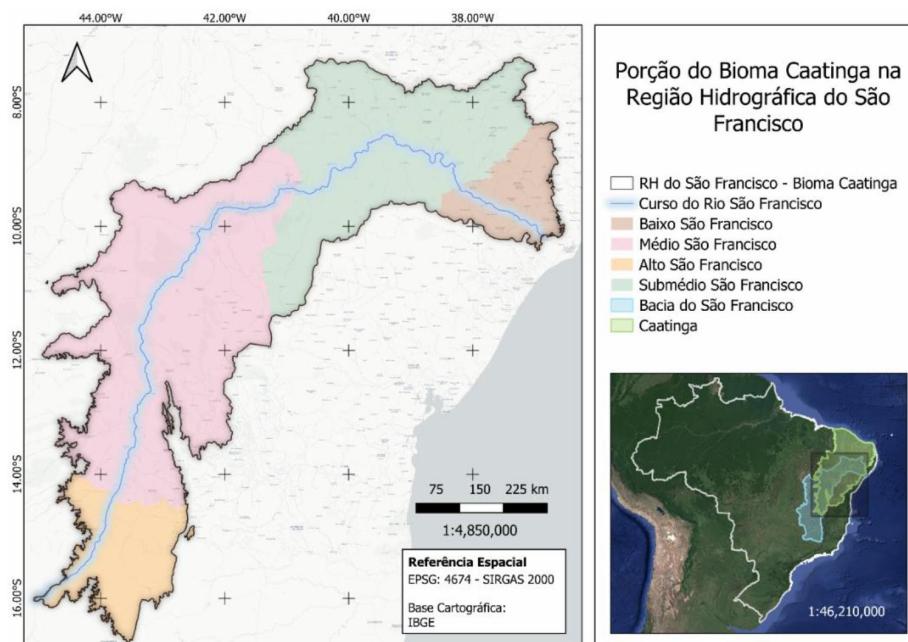


Figura 2. Mapa da região de estudo

Fonte: Autoria própria (2024)

Os dados vetoriais, em formato *shapefile*, para delimitação da área de estudo foram obtidos por meio do banco de dados abertos do Instituto Brasileiro de Geografia e Estatística [IBGE] e do catálogo de metadados da Agência Nacional de Águas e Saneamento Básico [ANA]. Foram utilizados o mapa de Biomas e Sistema Costeiro-Marinho do Brasil (IBGE, 2019) e o mapa digital de Divisão Hidrográfica Nacional (ANA, 2020a).

O vetor referente ao curso do Rio São Francisco foi obtido no módulo SF Map do Sistema de Informações sobre Recursos Hídricos da Bacia Hidrográfica do Rio São Francisco [SIGA São Francisco] (CBH SAO FRANCISCO, 2015). O software *open source* Quantum GIS [QGIS] de sistemas de informações geográficas foi utilizado para manipulação dos dados geoespaciais e elaboração dos mapas usando o sistema de referência de coordenadas EPGS 4674: SIRGAS 2000.

2.2 Obtenção e pré-tratamento dos dados

Foram obtidos dados referentes a cinco variáveis explicativas numéricas (precipitação, temperatura, radiação, umidade do ar e velocidade do vento), uma variável explicativa categórica (tipo de uso e cobertura do solo) e uma variável resposta numérica (vazão). Todas as bases consultadas continham dados ao menos entre os anos de 2012 e 2022, portanto, esse foi o período de análise adotado para todas as variáveis.

A disponibilidade hídrica superficial é um valor de vazão mínima que representa a oferta de água a ser considerada no Balanço Hídrico, o qual consiste na relação entre a oferta de água superficial e a demanda por essa água em diversas atividades humanas. A ANA adota a vazão mínima de 95% (vazão Q95), para determinado curso d'água, que corresponde à vazão média diária com 95% de permanência. Desta forma, é importante compreender que apesar de existirem dados reportando as vazões máximas de determinado curso d'água, o valor de referência para determinar a disponibilidade hídrica superficial foi a Q95 (ANA, 2020b).

Assim, para as variáveis precipitação e vazão, foram utilizados dados de 332 estações pluviométricas e 124 estações fluviométricas, obtidos por meio do plugin *ANA Data Acquisition*, disponível no QGIS. Para as variáveis temperatura, radiação, umidade do ar e velocidade do vento, utilizou-se os dados de 79 estações automáticas disponibilizados pelo Instituto Nacional de Meteorologia (INMET, 2024). A limpeza e organização desses dados foram feitas utilizando as bibliotecas OS, *pandas* e *numpy* da linguagem de programação Python.

Após o tratamento dos dados referentes às variáveis numéricas, foi realizada a interpolação espacial dos dados de cada variável por ano utilizando a ferramenta IDW,

também disponível no QGIS, que realiza a ponderação pelo inverso da distância de uma camada vetorial de pontos.

Quanto à variável categórica, dados do tipo *raster* apresentando os tipos de uso e cobertura do solo, desenvolvidos pelo projeto MapBiomas Brasil (MapBiomas, 2024), foram obtidos por meio dos Toolkits preparados no Google Earth Engine (GEE). Além disso, foi utilizado o arquivo de estilo QGIS (formato QML) da coleção 8 do MapBiomas para classificação da camada raster, definindo as classes (Floresta, Formação natural não florestal, Agropecuária, Área não vegetada e Corpo d’água) e subclasses a partir dos diferentes valores de pixel.

A partir dos mapas gerados pelo método IDW para as variáveis numéricas e dos mapas classificados para a variável categórica, realizou-se a amostragem de dados utilizando as camadas *raster* de cada variável por ano. Os pontos de amostragem foram determinados por meio da ferramenta de criação de pontos regulares do QGIS, onde foram selecionados 2678 pontos com intervalos regulares distribuídos dentro da região de estudo.

Posteriormente, com finalidade de caracterização da área de estudo, um mapa de classificação climática de Köppen-Geiger foi elaborado para verificar os tipos de clima predominantes. Para isso, utilizou-se a base cartográfica digital em formato shapefile das zonas climáticas disponibilizado pelo Centro de Estudos da Metrópole da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (CEM, 2021).

Por fim, a partir de dados geoespaciais da ANA (2021), realizou-se a categorização das vazões Q95 nos cursos d’água da região de estudo utilizando o QGIS. O mapa da disponibilidade hídrica superficial referente ao ano de 2017 (correspondente aos dados mais recentes para essa variável) da ANA foi utilizado para comparação com o mapa gerado para o mesmo ano por meio da interpolação IDW com os dados coletados, visando verificar a coerência na distribuição espacial da variável.

2.3 Estatística descritiva

Os dados foram explorados inicialmente por meio de estatísticas descritivas, que utilizam métodos numéricos e gráficos para procurar padrões em um conjunto de dados, resumir informações e apresentá-las de uma forma conveniente. Para as variáveis numéricas foram utilizadas algumas medidas-resumo (medidas de tendência central e dispersão) e medidas de correlação (coeficiente de correlação de Pearson). No caso das variáveis categóricas, a verificação dos dados consistiu na análise das frequências relativas e absolutas de cada classe estabelecida (McClave e Sincich, 2018).

A variável qualitativa “uso do solo” disponibilizada pelo MapBiomas engloba 5 categorias principais (Floresta, Formação natural não florestal, Agropecuária, Área não

vegetada e Corpo d'água) e mais 32 subcategorias. Entretanto, considerou-se apenas as classes principais na determinação das variáveis do tipo *dummy* aplicadas nos modelos supervisionados.

2.4 Aplicação de técnicas de Machine Learning

Inicialmente, visando identificar correlações entre as variáveis originais e possivelmente obter uma redução estrutural, foi realizada a análise por componentes principais (PCA) das variáveis numéricas. A PCA é um método de aprendizagem não supervisionado que consiste em uma análise multivariada capaz de reduzir a complexidade dos conjuntos de dados enquanto preserva sua covariância. Essa redução é obtida criando-se combinações lineares de variáveis, chamadas de componentes principais, que caracterizam os objetos estudados (Fávero e Belfiore, 2017).

Em seguida, para avaliação da capacidade de predição do conjunto de dados, foram analisados os seguintes modelos supervisionados de regressão utilizando a linguagem de programação R: Regressão Linear e Não Linear com transformação de Box-Cox, Árvore de Regressão e Random Forest.

Um modelo de regressão linear (Figura 3) assume que uma função de regressão é linear nas variáveis de entrada. Esses modelos são simples e muitas vezes fornecem uma descrição adequada e interpretável de como variáveis explicativas afetam os valores previstos. Em geral, o objetivo da regressão linear é encontrar os parâmetros α (coeficiente linear) e β (coeficiente angular) para os quais o termo de erro é minimizado. A técnica mais usual de estimação desse modelo é o método dos mínimos quadrados (Hastie et al., 2009).

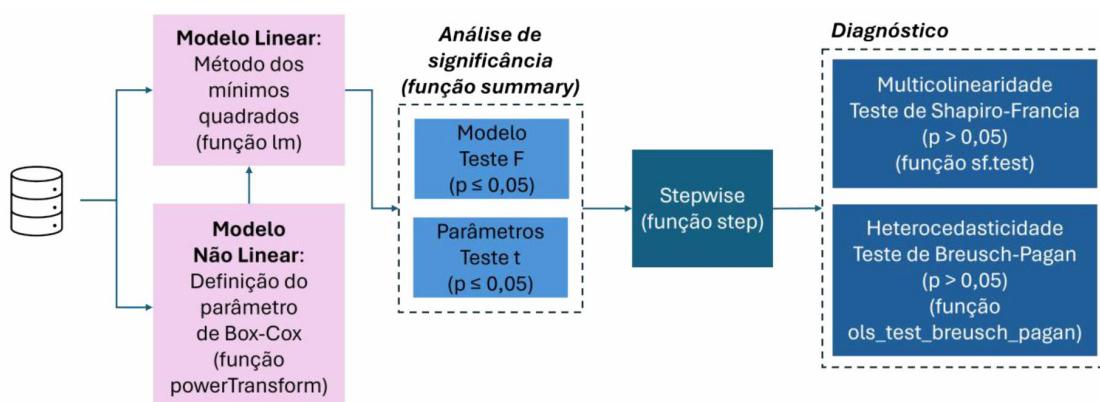


Figura 3. Etapas dos modelos de Regressão Linear e Regressão Não Linear no R

Fonte: Autoria própria (2024)

No entanto, há situações em que as relações entre variáveis podem se manifestar de várias formas funcionais não lineares, as quais devem ser consideradas durante a estimativa

de modelos de regressão para uma compreensão mais precisa do comportamento dos diversos fenômenos. Nestes casos, a partir do modelo de regressão linear, é possível derivar um modelo de regressão não linear através da transformação de Box-Cox (Figura 3). Esse método implica na substituição da variável resposta Y por parâmetros de transformação, conforme determinada forma funcional (linear, semilogarítmica, logarítmica, inversa, quadrática ou cúbica) (Fávero e Belfiore, 2017).

Uma abordagem alternativa à regressão linear é a Árvore de Regressão (Figura 4), adequada para conjuntos de dados com muitos recursos interagindo de maneiras não lineares. Ao invés de um único modelo global, a árvore de regressão subdivide o espaço de dados em regiões menores através de particionamento recursivo. Cada nó terminal da árvore representa uma célula da partição e possui um modelo simples associado, aplicável apenas àquela célula. A previsão para uma observação é feita utilizando a média dos dados de treinamento na região correspondente. Esse método simplifica a interpretação do modelo e permite lidar com interações complexas de forma mais gerenciável (Hastie et al., 2009).

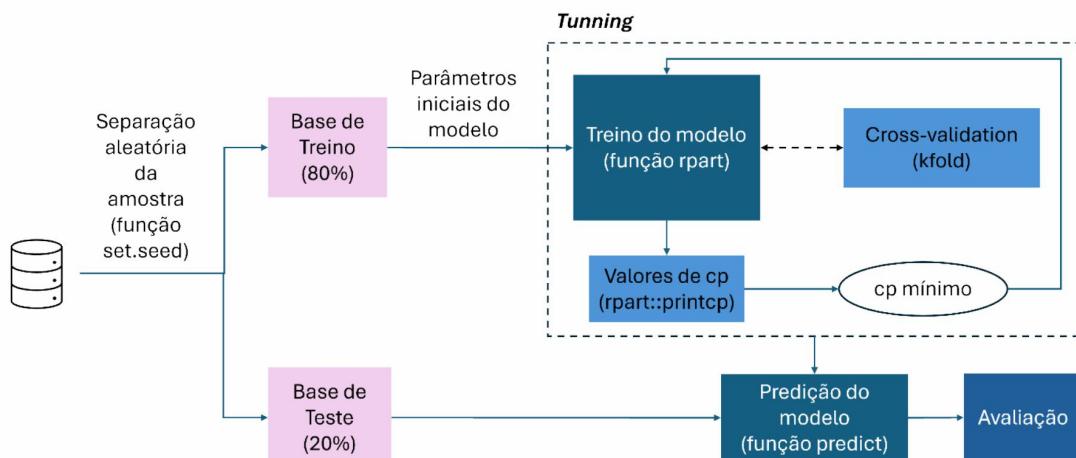


Figura 4. Etapas do modelo de Árvore de Regressão no R

Fonte: Autoria própria (2024)

A validação cruzada é uma técnica crucial na avaliação de modelos de machine learning, como as árvores de regressão. O método k-fold, uma forma comum de validação cruzada, divide o conjunto de dados em k subconjuntos, usando k-1 para treinamento e o restante para validação. Isso é repetido k vezes, alternando os subconjuntos de validação, permitindo uma avaliação robusta do desempenho do modelo. O parâmetro cp (complexidade do custo) em árvores de regressão controla a complexidade do modelo, evitando o sobreajuste (*overfitting*). A seleção adequada do parâmetro cp é crucial para equilibrar a complexidade do modelo e sua capacidade de generalização (Hastie et al., 2009).

O método de Random Forest (Figura 5) é uma técnica de aprendizado de máquina que utiliza um conjunto de árvores de decisão para realizar a regressão ou classificação. Cada árvore é construída de forma independente, com uma amostra aleatória dos dados de treinamento e com um subconjunto aleatório dos recursos. A previsão final é feita pela média das previsões de todas as árvores (no caso da regressão). Essa abordagem apresenta vantagens sobre a árvore de regressão, pois reduz a tendência ao *overfitting*, resultante da diversidade das árvores construídas com diferentes subconjuntos de dados. Além disso, a Random Forest lida melhor com a dimensionalidade dos dados e é mais robusta a outliers, tornando-se uma escolha preferencial em muitos cenários de modelagem preditiva.

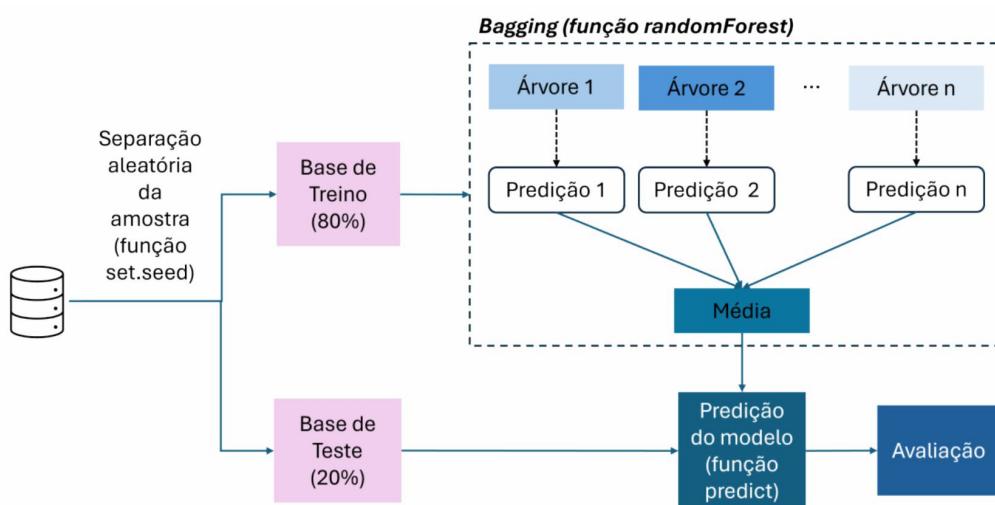


Figura 5. Etapas do modelo de Random Forest no R

Fonte: Autoria própria (2024)

3 Resultados e Discussões

3.1 Caracterização da região da Caatinga na Bacia do São Francisco

O bioma Caatinga ocupa cerca de 51% do território da Bacia Hidrográfica do São Francisco, que abrange os estados de Alagoas, Sergipe, Pernambuco, Bahia e Minas Gerais. A região de estudo possui aproximadamente 323.175 km² e engloba 334 municípios, dos quais 88% estão situados na região Nordeste.

A cobertura do solo no ano de 2022 (Figura 6) era constituída por aproximadamente 51% de florestas (formação florestal e formação savânica), 42% de atividades agropecuárias (silvicultura, pastagem, cana, soja, café, algodão, mosaico de usos, outras lavouras temporárias, outras lavouras perenes), 5% de formação natural não florestal (campo alagado e área pantanosa, formação campestre e afloramento rochoso), 2% por corpos d'água (rio,

lago e oceano) e 1% por área não vegetada (praia, duna e areal, área urbanizada, outras áreas não vegetadas, mineração). Dentre esses, destacam-se a formação savânicas e as pastagens, que ocupam 50 e 26% da extensão territorial, respectivamente. O ano de 2022 foi escolhido como referência inicial por englobar os dados disponíveis mais recentes acerca dos tipos de uso e cobertura do solo.

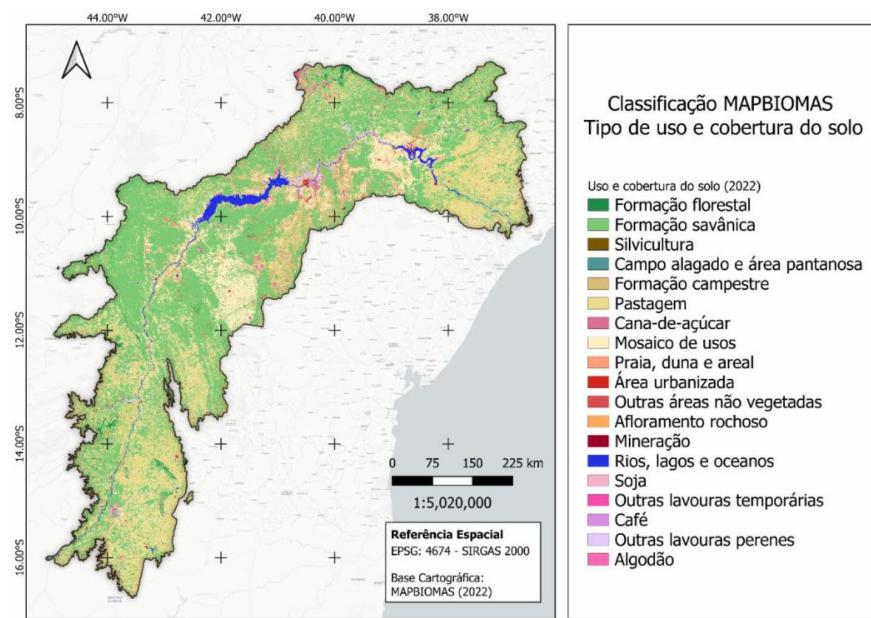


Figura 6. Tipos de uso e cobertura do solo na porção de Caatinga da RHSF em 2022 (Dados: MapBiomas, 2024)

Fonte: Autoria própria (2024)

O clima predominante é o semiárido (55,3%), seguido pelo tropical de savana verão seco (27,9%), tropical de savana inverno seco (16,3%) e o subtropical (0,5%) (Figura 7a).

A Figura 7b apresenta o comportamento histórico da precipitação diária média (mm) entre 2012 e 2022, classificado de acordo com os quantis 20, 40, 60 e 80. Os menores índices pluviométricos são observados nas regiões do Submédio e em parte do Médio e do Baixo São Francisco, que apresentam clima semiárido.

Correia et al. (2011) ressaltam que a região do semiárido exibe grande variabilidade quanto à distribuição de chuvas, com períodos longos de estiagem (8 a 9 meses) intercalados com estações poucas chuvosas e episódios pontuais de chuvas intensas. Esse comportamento pode estar associado à ocorrência de fenômenos como o *El Niño* e às variações de padrões de temperatura da superfície do mar (TSM) sobre os oceanos tropicais.

Ao avaliar os padrões temporais de uso da terra e precipitação sob desertificação na região semiárida do Brasil, Silva et al., (2023) demonstram que mesmo com eventos extremos de chuva, a região experimenta um aumento da seca, influenciando o uso dos recursos

hídricos e impactando a agricultura. Porém, os autores relatam que ainda há predominância de vegetação nativa em comparação com regiões semiáridas de outros países, permitindo a construção de políticas públicas para restaurar a vegetação na área.

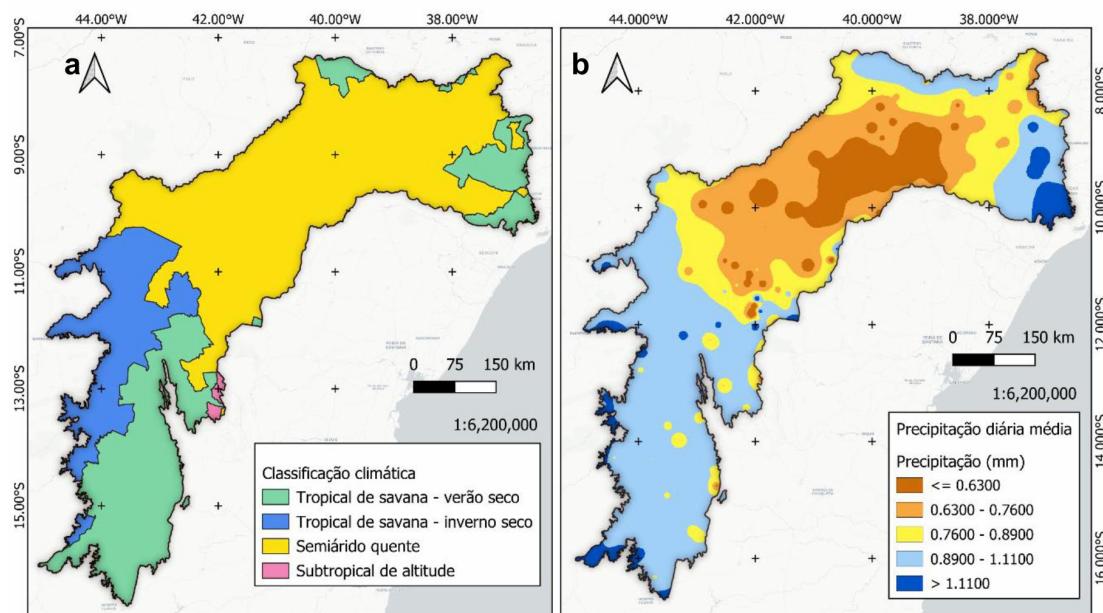


Figura 7. (a) Classificação climatológica (Dados: CEM, 2021); (b) Precipitação diária média na Caatinga na RHSF (2012 a 2022) (Dados: ANA, 2024)
Fonte: Autoria própria (2024)

3.2 Análise inicial dos dados

A Tabela 1 apresenta as medidas-resumo referentes aos dados brutos coletados para as variáveis numéricas. As variáveis radiação, temperatura, umidade e vento apresentam uma distribuição mais próximo à normalidade, com valores da média próximos à mediana.

No caso da precipitação, o 3º quartil indica que pelo menos 75% dos dados são iguais a zero, mostrando os baixíssimos índices de ocorrência de chuva nos postos pluviométricos observados. A distribuição desses dados está concentrada em valores mais baixos, com uma média maior que a mediana.

Tabela 1. Dados brutos referentes às variáveis numéricas

	Prec (mm)	Rad (kJ/m ²)	Temp (°C)	Umid (%)	Vento (m/s)	Q95 (m ³ /s)
Mínimo	0	0	10,35	0,20	0	0
1º Quartil	0	1278,71	23,89	49,67	1,44	0
Mediana	0	1628,97	25,88	61,79	2,14	1,38
Média	1,83	1575,48	25,75	61,17	2,65	250,75
3º Quartil	0	1901,42	27,68	73,12	3,012	92,08
Máximo	502,9	23105,90	90,75	100	288,33	52893,00
Desvio padrão	7,93	477,76	3,00	15,74	7,72	622,74

Prec.: Precipitação diária; Rad.: Radiação diária; Temp.: Temperatura diária; Umid.:
Umidade do ar diária; Vento: Velocidade do vento diária; Q95: vazão Q95 diária.

Fonte: Autoria própria (2024)

A vazão também está concentrada em valores mais baixos, com uma média superior proeminente em relação à mediana. A mediana indica que pelo menos 50% dos valores são menores ou iguais a 1,38 m³/s, ressaltando novamente a problemática da baixa disponibilidade hídrica na região.

O conjunto de dados inicial possui outliers bem discrepantes, principalmente para as variáveis precipitação, radiação, temperatura, velocidade do vento e vazão Q95, o que poderia prejudicar a eficiência de predição dos modelos. Deste modo, utilizou-se o método IQR (Intervalo entre Quartis) para remoção desses valores.

Após a exclusão dos outliers, o método de interpolação IDW foi aplicado para estimar as distribuições espaciais dos parâmetros na superfície de interesse. Essa etapa é fundamental porque os dados para cada tipo de variável foram obtidos em diferentes estações de monitoramento, localizadas em pontos distintos. Sendo assim, é importante utilizar dados representativos de pontos em comum, possibilitando o estabelecimento de correlações entre os parâmetros de interesse.

A princípio, verificou-se que o método IDW não foi adequado para estimar a distribuição da vazão Q95. A Figura 8 apresenta uma análise inicial dos mapas, considerando a amostra com o Rio São Francisco (ARSF) e a amostra sem o Rio São Francisco (AsRSF), que foram comparados com o mapa de referência da ANA para disponibilidade hídrica no ano de 2017, sendo os dados publicados mais recentes.

A alta disponibilidade hídrica vazões do Rio São Francisco, quando comparada à sua região hidrográfica como um todo, exerceu grande influência na distribuição nas áreas sem a presença de postos fluviométricos (Figura 8a), mostrando a predominância de altos valores de Q95 (> 119 m³/s).

Enquanto isso, a amostra AsRSF (Figura 8b) apresentou um comportamento mais próximo aos dados da ANA, com maior cobertura entre 0 e 10 m³/s (Figura 8c). Observa-se que a região do Alto São Francisco, na faixa mais à oeste, apresenta índices fluviométricos mais elevados em comparação às demais regiões. Além disso, as áreas com menor vazão estão localizadas no Semiárido, região com índices críticos de precipitação baixa, indicando, a princípio, a existência de relação entre os fatores climáticos e a disponibilidade hídrica.

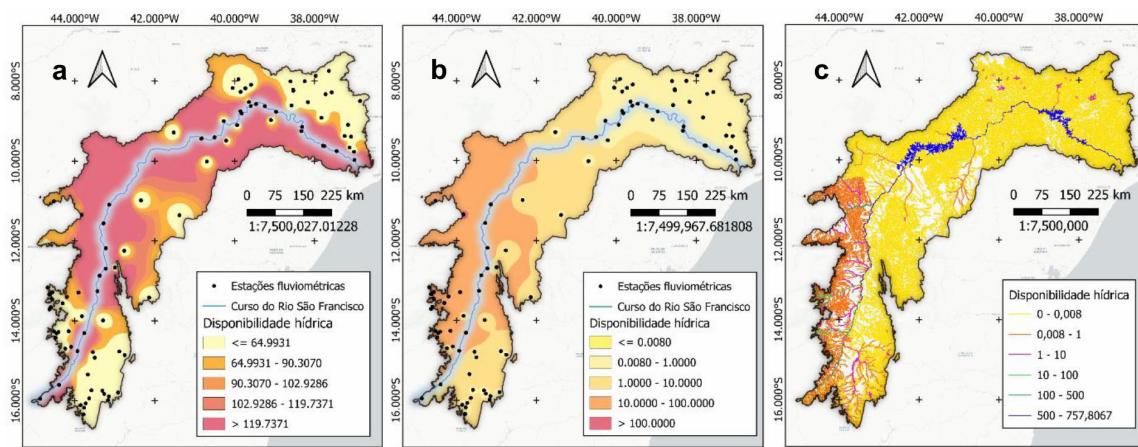


Figura 8. a) Q95 amostra ARSF; b) Q95 amostra AsRSF; c) Q95 dados ANA (2017).

Fonte: Autoria própria (2024)

A discrepância entre as vazões do Rio São Francisco e dos corpos d'água em seu entorno pode ser verificada na Tabela 2, mostrando que a mediana referente aos demais cursos d'água corresponde a apenas 0,036% da mediana do rio principal.

Tabela 2. Comparação das vazões Q95 no Rio São Francisco e demais cursos de água (valores diários, sem outliers)

	Q95 diária no Rio São Francisco (m ³ /s)	Q95 diária nos demais cursos d'água (m ³ /s)
Mínimo	0	0
1º Quartil	712.84	0
Mediana	942.15	0,34
Média	1024.48	18,63
3º Quartil	1240.03	14,54
Máximo	2307.61	282,62
Desvio padrão	425.06	40,46

Fonte: Autoria própria (2024)

Portanto, visando garantir a melhor representatividade da variável disponibilidade hídrica (vazão Q95) para a região, foram considerados os dados da amostra AsRSF no desenvolvimento dos modelos de predição.

3.3 Descrição dos dados usados na modelagem

As medidas-resumo para o conjunto de dados espacializados no período de 2012 a 2022, obtido a partir dos mapas interpolados, está apresentado na Tabela 3. A Figura 9 apresenta as correlações de Pearson para esses dados, variando entre -1 (alta correlação negativa), 0 (ausência de correlação) e +1 (alta correlação positiva).

Tabela 3. Dados após regionalização.

	Prec (mm)	Rad (kJ/m ²)	Temp (°C)	Umid (%)	Vento (m/s)	Q95 (m ³ /s)
Mínimo	0,07	538,25	20,52	22,73	0,52	0
1º Quar.	0,67	1557,38	25,34	56,26	2,01	0,96
Mediana	0,83	1618,17	26,15	59,38	2,35	6,45
Média	0,86	1605,53	26,03	59,49	2,37	14,30
3º Quar.	1,03	1674,06	26,76	62,43	2,72	19,94
Máximo	3,39	2593,21	29,88	79,09	4,64	216,63
Desvio padrão	0,28	130,77	0,99	4,69	0,48	19,16

Prec.: Precipitação diária média; Rad.: Radiação diária média; Temp.: Temperatura diária média; Umid.: Umidade do ar diária média; Vento: Velocidade do vento diária média; Q95: vazão Q95 diária média.

Fonte: Autoria própria (2024)

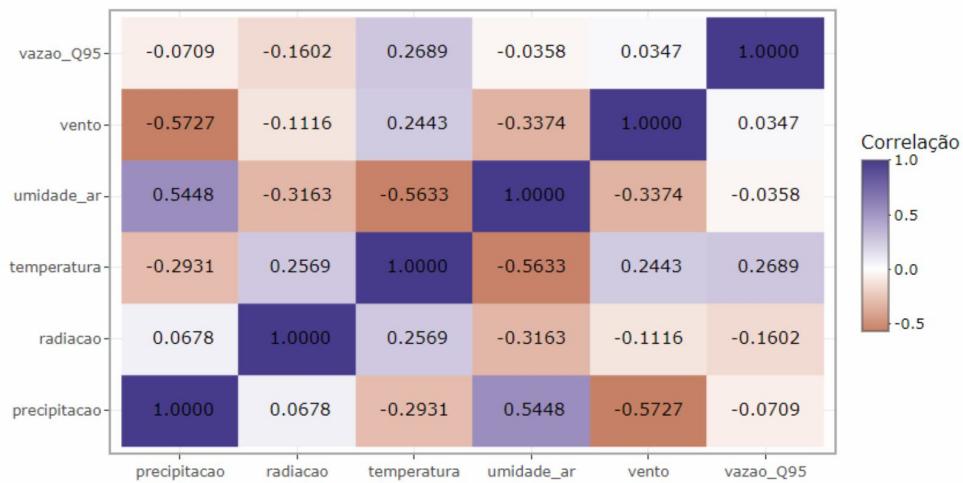


Figura 9. Mapa de calor da correlação de Pearson para as variáveis quantitativas.

Fonte: Autoria própria (2024)

A velocidade do vento e a precipitação apresentaram o maior grau de correlação (-0,57), seguido da temperatura e umidade do ar (-0,56) e da precipitação e umidade do ar (0,54). Ou seja, altos índices de precipitação foram relacionados a velocidades de vento inferiores e umidade do ar elevada. Já as altas temperaturas podem indicar a ocorrência de baixos níveis de umidade do ar. A vazão Q95 apresentou baixa correlação para todas as variáveis explicativas. Um melhor entendimento dessas correlações pode ser observado por meio da análise de componentes principais discutida no tópico seguinte.

Quanto à variável quantitativa, a Tabela 4 apresenta a distribuição de frequências para as classes de uso e cobertura do solo, com predominância de cobertura vegetal nativa (51,8%) e da Agropecuária (41,2%).

Tabela 4. Frequência das categorias de uso e cobertura do solo.

	Agropecuária	Área não vegetada	Corpo d'água	Floresta	Formação natural não florestal
Frequência absoluta	12131	355	403	15282	1287
Frequência relativa	41,2%	1,2%	1,4%	51,8%	4,4%

Fonte: Autoria própria (2024)

3.4 Análise por componentes principais

Na PCA, os 2 primeiros fatores explicaram 72,5% da variância dos dados originais (Figura 10). No caso de PC1, este componente parece estar relacionado principalmente com a temperatura (-0,704) e a umidade do ar (0,848). Isso sugere que o PC1 pode representar condições climáticas mais úmidas quando ocorrem temperaturas mais baixas.

O componente PC2 apresenta maior relação principalmente com a radiação solar (0,834) e a velocidade do vento (-0,516), inferindo que o PC2 pode representar condições climáticas com alta radiação solar e baixa velocidade do vento.

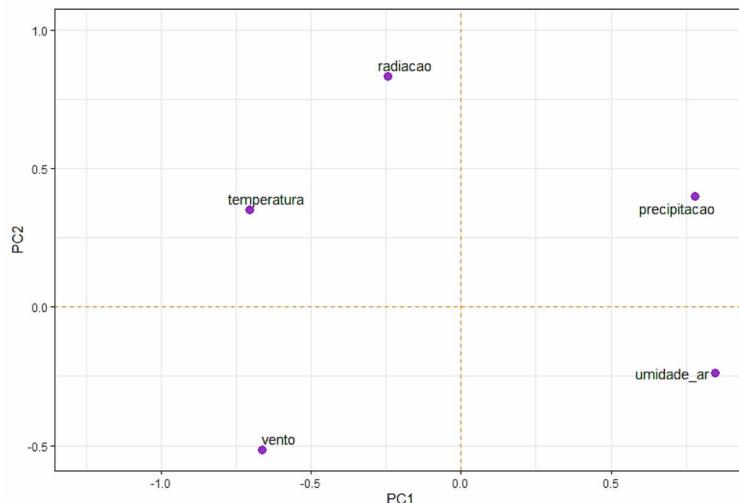


Figura 10. Análise por Componentes Principais

Fonte: Autoria própria (2024)

A precipitação contribui positivamente para PC1 (0,780) e PC2 (0,398), sugerindo que está relacionada tanto com condições de temperatura e umidade (PC1) quanto com radiação solar e velocidade do vento (PC2).

As communalidades indicam a proporção da variância de cada variável que é explicada pelos componentes principais. Valores mais altos de communalidade (próximos a 1) sugerem que a variável está bem representada pelos componentes principais. Todas as variáveis

apresentaram comunalidades razoavelmente altas (precipitação = 0,767; radiação = 0,755; temperatura = 0,619; umidade do ar = 0,776; velocidade do vento = 0,706), indicando que são bem representadas pelos componentes principais extraídos.

Portanto, decidiu-se manter todas as variáveis no modelo, uma vez que cada uma delas é considerada relevante para a análise abrangente do sistema climático na Região Hidrográfica do São Francisco. Dado que o clima na região é complexo e caracterizado por múltiplos fatores interrelacionados, a inclusão de todas as variáveis pode permitir uma representação mais completa e precisa do sistema climático, garantindo que nenhuma informação importante seja negligenciada durante a análise dimensional.

3.5 Avaliação dos modelos supervisionados

3.5.1 Modelo de Regressão Linear Múltipla

Os resultados do Modelo de Regressão Linear (Tabela 5) indicam que todas as variáveis independentes possuem coeficientes significativos, como evidenciado pelos *p-values* iguais a zero em todos os casos, exceto para a variável "Corpo d'água", que tem um valor-p de 0,03. Isso sugere que todas as variáveis independentes, com exceção de "Corpo d'água", têm um impacto estatisticamente significativo na variável dependente.

O teste F também se apresentou estatisticamente significativo, indicando que o modelo como um todo fornece um ajuste melhor do que um modelo sem variáveis independentes.

No entanto, o teste Shapiro-Francia e o teste Breusch-Pagan foram ambos significativos, com valores-p menores que 0,05, indicando possíveis violações das suposições de normalidade dos resíduos e homocedasticidade, respectivamente. Isso sugere que os resíduos do modelo podem não seguir uma distribuição normal e a variância dos resíduos pode não ser constante em relação às variáveis independentes.

O coeficiente de determinação (R^2) igual a 0,33 indica que aproximadamente 33% da variabilidade na variável dependente é explicada pelas variáveis independentes incluídas no modelo. Embora não seja um valor muito alto, mostra que o modelo tem algum poder preditivo.

Tabela 5. Parâmetros do Modelo de Regressão Linear Múltipla

Modelo	Parâmetros	R^2
Regressão Linear Múltipla	<u>Coeficientes significativos (teste t):</u> Intercepto (p-value: 0) Precipitação (p-value: 0) Radiação (p-value: 0) Temperatura (p-value: 0) Umidade do ar (p-value: 0) Velocidade do vento (p-value: 0)	0,33

Agropecuária (p-value: 0)
Área não vegetada (p-value: 0)
Corpo d'água (p-value: 0,03)
Formação natural não florestal (p-value: 0)

Teste F: significativo (p-value: 0)
Teste Shapiro-Francia: significativo (p-value: 0)
Teste Breusch-Pagan: significativo (p-value: 0)

Fonte: Autoria própria (2024)

3.5.2 Modelo de Regressão Não Linear

Os resultados do Modelo de Regressão Não Linear (Tabela 6) mostram que a transformação de Box-Cox foi aplicada aos dados, com um parâmetro encontrado de 0,1257. Esta transformação é comumente utilizada para estabilizar a variância ou tornar os resíduos mais normalmente distribuídos, o que ajuda a atender às suposições do modelo.

Tabela 6. Parâmetros do Modelo de Regressão Não Linear

Modelo	Parâmetros	R ²
Regressão Não Linear (Transformação de Box-Cox)	<u>Parâmetro de Box-Cox: 0,1257</u>	0,35
<u>Coeficientes significativos (teste t):</u>		
Intercepto (p-value: 0) Precipitação (p-value: 0) Radiação (p-value: 0,03) Temperatura (p-value: 0) Umidade do ar (p-value: 0) Velocidade do vento (p-value: 0) Agropecuária (p-value: 0) Área não vegetada (p-value: 0,003) Corpo d'água (p-value: 0) Formação natural não florestal (p-value: 0)		
<u>Teste F:</u> significativo (p-value: 0) <u>Teste Shapiro-Francia:</u> significativo (p-value: 0) <u>Teste Breusch-Pagan:</u> significativo (p-value: 0)		

Fonte: Autoria própria (2024)

Os coeficientes estimados para todas as variáveis independentes, exceto "Radiação" e "Área não vegetada", apresentaram valores-p muito baixos, indicando que são estatisticamente significativos para prever a variável dependente. Isso sugere que essas variáveis possuem um impacto significativo no resultado do modelo. Para "Radiação", o valor-p foi 0,03, o que ainda indica significância, mas com uma certa ressalva. Já para "Área não vegetada", o valor-p é ainda menor, 0,003, reforçando sua relevância. O teste F também foi

estatisticamente significativo, indicando que o modelo como um todo fornece um ajuste adequado para as variáveis explicativas.

No entanto, assim como ocorreu para o modelo linear, tanto o teste Shapiro-Francia quanto o teste Breusch-Pagan foram significativos, com valores-p maiores que 0,05. Isso mostra que provavelmente ocorreu a não aderência à normalidade dos termos de erro, indicando que o modelo foi especificado incorretamente quanto à forma funcional e que houve a omissão de variáveis explicativas relevantes. Além disso, ressalta-se o problema da heterocedasticidade, com a não constância da variância dos resíduos ao longo da variável explicativa.

Por fim, o coeficiente de determinação (R^2) do modelo foi igual a 0,35, indicando um baixo poder preditivo. Nesse caso, apenas 35% da variabilidade na variável resposta é explicada pelas variáveis explicativas incluídas no modelo.

3.5.3 Árvore de Regressão

O modelo de árvore de regressão (Tabela 7) foi configurado com uma profundidade máxima de 30, indicando que a árvore pode ter até 30 níveis de divisão para capturar a complexidade dos dados. No entanto, o parâmetro de complexidade (cp) encontrado foi bastante baixo, aproximadamente 5.105288e-08, o que sugere que o modelo pode ter sido podado de forma agressiva durante o treinamento para evitar o *overfitting*, ou seja, para garantir que não se adaptasse em excesso aos dados de treinamento e, assim, mantivesse sua capacidade de generalização para novos dados.

Tabela 7. Parâmetros do modelo de Árvore de Regressão

Modelo	Parâmetros	R^2
Árvore de Regressão	maxdepth = 30 cp = 5.105288e-08 k-fold = 10	0,68

Fonte: Autoria própria (2024)

A validação cruzada foi realizada com 10 *folds*, o que significa que os dados foram divididos em 10 partes iguais, e o modelo foi treinado em 9 partes e avaliado na parte restante, repetindo esse processo 10 vezes. Esse método é útil para avaliar o desempenho do modelo e garantir que ele generalize bem para dados não vistos.

Os resultados revelaram um coeficiente de determinação (R^2) de 0,86 na base de treinamento. Por outro lado, na base de teste, o modelo alcançou um R^2 de 0,68, o que significa que aproximadamente 68% da variabilidade na variável dependente foi explicada

pelas variáveis explicativas. Esses resultados sugerem que o modelo de árvore de regressão tem um bom desempenho na previsão da variável dependente, embora seja importante notar uma redução no R^2 ao aplicar o modelo em dados não utilizados durante o treinamento. Isso pode indicar uma capacidade ligeiramente inferior de generalização em comparação com a performance observada nos dados de treinamento.

3.5.4 Random Forest

O modelo Random Forest (Tabela 8) foi treinado com número de árvores igual a 50. Com uma divisão dos dados em 80% para treinamento e 20% para teste, observou-se um coeficiente de determinação (R^2) de 0,94 na base de treinamento. Este valor sugere que aproximadamente 94% da variabilidade na variável dependente foi adequadamente capturada pelo modelo, demonstrando um ajuste altamente satisfatório aos dados de treinamento.

Tabela 8. Parâmetros do modelo de Random Forest

Modelo	Parâmetros:	R^2
Random Forest	ntree = 50	0,80

Fonte: Autoria própria (2024)

Ao avaliar o modelo na base de teste, constatou-se um R^2 de 0,80. Esse resultado indica que cerca de 80% da variabilidade na variável resposta foi explicada pelo modelo quando aplicado a dados não vistos. Embora ligeiramente inferior ao R^2 da base de treinamento, essa diferença não é excessivamente grande, sugerindo que o modelo conseguiu generalizar adequadamente para novos dados. A abordagem de avaliação do modelo na base de teste é fundamental para validar sua capacidade de generalização e prevenir o *overfitting*, garantindo que as previsões sejam precisas em situações do mundo real.

Em resumo, o modelo de regressão linear demonstrou coeficientes altamente significativos para todas as variáveis independentes, porém, evidenciou possíveis violações das suposições de normalidade dos resíduos e homocedasticidade. Já o modelo de regressão não linear com transformação de Box-Cox apresentou uma melhoria no R^2 em relação ao modelo linear, embora ainda tenha indicado possíveis desafios em relação à normalidade dos resíduos e homocedasticidade. A árvore de regressão, por sua vez, ofereceu uma abordagem não paramétrica, capturando relações não lineares entre as variáveis, enquanto o modelo Random Forest demonstrou um desempenho superior na previsão da variável vazão Q95, tanto nos dados de treinamento quanto nos de teste, destacando sua capacidade de

generalização e robustez. Essa análise comparativa ressalta a importância de considerar diferentes abordagens de modelagem para compreender e prever com precisão os fenômenos estudados.

3.6 Incertezas das estimativas e prováveis fontes de erros

Os modelos avaliados apresentaram ajustes insatisfatórios, com um R^2 máximo de 80%. Esse resultado pode estar associado às incertezas nas estimativas de vazão de referência, que ocorrem especialmente quando dados são limitados. Uma das fontes de erro foi a falta de postos fluviométricos na Região Hidrográfica do São Francisco, resultando em uma distribuição de vazões inadequada. Além disso, os dados de vazões coletados podem conter erros ou terem sido baseados em modelos matemáticos inadequados (Collischonn et al., 2023).

Outra possível causa para a ineficiência dos modelos foi a análise considerando apenas variáveis climáticas e tipo de uso e cobertura do solo como entradas, excluindo outras variáveis relevantes como tipo de solo, evapotranspiração, relevo, águas subterrâneas e vazões outorgadas de água. Portanto, os modelos não são infalíveis e devem ser utilizados com cautela na tomada de decisões para garantir o desenvolvimento sustentável dos recursos hídricos na região estudada.

4 Conclusões

Este estudo focalizou a área correspondente ao bioma Caatinga na Bacia Hidrográfica do São Francisco. Apesar da predominância do clima semiárido, com escassos eventos de precipitação, especialmente nas regiões do Baixo e do Submédio, observam-se maiores índices pluviométricos na região do Alto São Francisco, o que pode exercer um impacto significativo na disponibilidade hídrica superficial do rio. O Rio São Francisco, como principal curso d'água da região, apresenta vazões consideravelmente superiores aos demais corpos hídricos locais, destacando assim sua importância como fonte primordial de recursos hídricos e seu papel crucial para a preservação do bioma Caatinga.

Todas as variáveis explicativas, incluindo fatores climáticos e características de uso e cobertura do solo, mostraram-se significantes no desenvolvimento dos modelos. Embora o modelo Random Forest tenha demonstrado ser o mais eficaz, seu ajuste ainda poderia ser aprimorado. Este cenário se deve à presença de diversas fontes potenciais de erros e incertezas relacionadas aos dados utilizados e ao método de estimação da dispersão espacial IDW.

Propostas para estudos futuros abrangem a análise utilizando dados de satélite, a incorporação de novas variáveis, a adoção da interpolação geoestatística por meio do método de Top-Kriging e o uso de modelos de redes neurais. Essas abordagens podem enriquecer a compreensão dos processos hidrológicos na região, possibilitando uma gestão mais precisa e sustentável dos recursos hídricos no contexto do bioma Caatinga e da Bacia Hidrográfica do São Francisco.

Referências

- Ahmed, A.A., Sayed, S., Abdoulhalik, A., Moutari, S., Oyedele, L. 2024. Applications of machine learning to water resources management: A review of present status and future opportunities. *J Clean Prod.* <https://doi.org/10.1016/j.jclepro.2024.140715>
- Amorim, E.L.C., Peiter, F.S., Soares, E.C., Silva, J. V. 2022. Fossas agroecológicas para o tratamento de efluentes sanitários em escolas municipais do Baixo São Francisco, in: O Baixo São Francisco: Características Ambientais e Sociais. EDUFAL, Maceió, pp. 434–446.
- Agência Nacional de Águas e Saneamento Básico [ANA]. 2024. Impacto da Mudança Climática nos Recursos Hídricos no Brasil. Agência Nacional de Águas e Saneamento Básico, Brasília.
- Agência Nacional de Águas e Saneamento Básico [ANA]. 2021. Disponibilidade Hídrica Superficial (BHO 2017 5K) [WWW Document]. Catálogo de Metadados da Agência Nacional de Águas e Saneamento Básico. URL <https://metadados.snirh.gov.br/geonetwork/srv/api/records/7ac42372-3605-44a4-bae4-4dee7af1a2f8> (accessed 1.9.24).
- Agência Nacional de Águas e Saneamento Básico [ANA]. 2020a. Regiões Hidrográficas [WWW Document]. Catálogo de Metadados da Agência Nacional de Águas e Saneamento Básico. URL <https://metadados.snirh.gov.br/geonetwork/srv/api/records/0574947a-2c5b-48d2-96a4-b07c4702bbab> (accessed 1.9.24).
- Agência Nacional de Águas e Saneamento Básico [ANA]. 2020b. Atualização da Base de Disponibilidade Hídrica Superficial da ANA. Brasil.
- Azevêdo, E. de L., Alves, R.R.N., Dias, T.L.P., Molozzi, J. 2017. How do people gain access to water resources in the Brazilian semiarid (Caatinga) in times of climate change? *Environ Monit Assess* 189. <https://doi.org/10.1007/s10661-017-6087-z>
- Comitê da Bacia Hidrográfica do São Francisco [CBH SAO FRANCISCO]. 2015. SF Map [WWW Document]. Comitê da Bacia Hidrográfica do São Francisco. URL <https://sigar.cbhsaofrancisco.org.br/sfmap/#> (accessed 1.9.24).
- Centro de Estudos da Metrópole [CEM]. 2021. Zonas Climáticas do Brasil, conforme Köppen [WWW Document]. Centro de Estudos da Metrópole. URL <https://centrodametropole.fflch.usp.br/pt-br/node/9973> (accessed 1.10.24).
- Collischonn, W., Sorribas, M., Paiva, R., Araujo, A., Souza, S.A. de. 2023. Métodos simples para estimar vazões de referência e sua incerteza. ABRHidro, Porto Alegre, RS.

Correia, R.C., Kiill, L.H.P., de Moura, M.S.B., Cunha, T.J.F., de Jesus Júnior, L.A., de Araújo, J.L.P. 2011. A região semiárida brasileira, in: Voltolini, T.V. (Ed.), Produção de Caprinos e Ovinos No Semiárido. Embrapa Semiárido, Petrolina.

Drumond, M.A. 2004. Recomendações para o uso sustentável da biodiversidade no Bioma Caatinga, in: Silva, J.M.C. da, Tabarelli, M., Fonseca, M.T. da, Lins, L. V. (Eds.), Biodiversidade Da Caatinga: Áreas e Ações Prioritárias Para a Conservação. Ministério do Meio Ambiente, Brasília, pp. 341–346.

Drumond, M.A., Kiill, L.H.P., Ribaski, J., Aidar, S.T. 2016. Caracterização e usos das espécies da Caatinga: subsídio para programas de restauração florestal nas Unidades de Conservação da Caatinga (UCCAs). Embrapa Semiárido, Petrolina.

Fávero, L.P., Belfiore, P., 2017. Manual de análise de dados. Elsevier, Rio de Janeiro.

Freire, N.C.F., Moura, D.C., Silva, J.B. da, Moura, A.S.S. de, Melo, J.I.M. de, Pacheco, A. da P. 2018. Atlas das caatingas: o único bioma exclusivamente brasileiro. Fundação Joaquim Nabuco, Editora Massangana, Recife.

Hastie, T., Tibshirani, R., Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. ed. Springer, California.

Instituto Brasileiro de Geografia e Estatística [IBGE]. 2019. Biomass | IBGE [WWW Document]. Instituto Brasileiro de Geografia e Estatística. URL <https://www.ibge.gov.br/geociencias/cartas-e-mapas/informacoes-ambientais/15842-biomass.html?edicao=25799&t=acesso-ao-produto> (accessed 1.9.24).

Instituto Nacional de Meteorologia [INMET]. 2024. Dados históricos anuais [WWW Document]. Instituto Nacional de Meteorologia. URL <https://portal.inmet.gov.br/dadoshistoricos> (accessed 4.11.24).

MapBiomass. 2024. MapBiomass Brasil [WWW Document]. URL <https://brasil.mapbiomas.org/> (accessed 1.10.24).

Mapulanga, A.M., Naito, H. 2019. Effect of deforestation on access to clean drinking water. Proc Natl Acad Sci U S A 116, 8249–8254. <https://doi.org/10.1073/pnas.1814970116>

McClave, J.T., Sincich, T. 2018. Statistics, 13 edition. ed. Pearson, New York.

Ministério do Meio Ambiente e Mudança do Clima [MMA]. 2022. Caatinga [WWW Document]. Ministério do Meio Ambiente e Mudança do Clima. URL <https://www.gov.br/mma/pt-br/assuntos/ecossistemas-1/biomass/caatinga> (accessed 10.5.23).

Silva, B.F. da, dos Santos Rodrigues, R.Z., Heiskanen, J., Abera, T.A., Gasparetto, S.C., Biase, A.G., Ballester, M.V.R., de Moura, Y.M., de Stefano Piedade, S.M., de Oliveira Silva, A.K., de Camargo, P.B. 2023. Evaluating the temporal patterns of land use and precipitation under desertification in the semi-arid region of Brazil. Ecol Inform 77, 102192. <https://doi.org/10.1016/J.ECOINF.2023.102192>

Silva, J.L.B. da, Moura, G.B. de A., Silva, M.V. da, Lopes, P.M.O., Guedes, R.V. de S., Silva, È.F. de F. e., Ortiz, P.F.S., Rodrigues, J.A. de M. 2020. Changes in the water resources, soil use and spatial dynamics of Caatinga vegetation cover over semiarid region of the Brazilian Northeast. Remote Sens Appl 20. <https://doi.org/10.1016/j.rsase.2020.100372>

Zhang, M., Wei, X. 2021. Deforestation, forestation, and water supply. *Science* (1979).
<https://doi.org/10.1126/science.abe7821>