UNIVERSIDADE FEDERAL DE JUIZ DE FORA
ICE/ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Welson de Avelar Soares Filho

Aprendizado de máquina com variáveis categóricas para previsão de vazão em bacias hidrográficas de Minas Gerais

Juiz de Fora
2024

Welson de Avelar Soares Filho

Aprendizado de máquina com variáveis categóricas para previsão de vazão em bacias hidrográficas de Minas Gerais

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional

Orientador: Doutor Leonardo Goliatt

Juiz de Fora

2024

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Soares Filho, Welson de Avelar.
Aprendizado de máquina com variáveis categóricas para previsão de vazão em bacias hidrográficas de Minas Gerais / Welson de Avelar Soares Filho. – 2024.
161 f. : il.

Orientador: Leonardo Goliatt
Dissertação (Mestrado Acadêmico) – Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2024.

1. Aprendizado de Máquina. 2. Modelagem chuva-vazão. 3. Previsão de vazão. 4. Gestão de recursos hídricos I. Goliatt, Leonardo, orient. II. Título.

Welson de Avelar Soares Filho

Aprendizado de máquina com variáveis categóricas para previsão de vazão em bacias hidrográficas de Minas Gerais

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional

Aprovada em 12 de dezembro de 2024

BANCA EXAMINADORA

Doutor Leonardo Goliatt - Orientador
Universidade Federal de Juiz de Fora

Doutora Eliane Aparecida Justino
Universidade Federal de Catalão

Doutor Celso Bandeira de Melo Ribeiro
Universidade Federal de Juiz de Fora

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 16/12/2024, às 15:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Celso Bandeira de Melo Ribeiro**, **Professor(a)**, em 17/12/2024, às 12:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **ELIANE APARECIDA JUSTINO**, **Usuário Externo**, em 27/12/2024, às 14:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2145682** e o código CRC **709EAB0D**.

AGRADECIMENTOS

Agradeço aos meus pais, Regina e Welson, por terem se dedicado, desde sempre, para que eu, meu irmão e irmã, buscássemos nos aprimorar enquanto cidadãos através dos estudos. A participação e acompanhamento em cima de nossos estudos fixaram em mim o desejo pela busca do conhecimento. Este trabalho tem parte de vocês. Muito obrigado.

Meu irmão Raphael e irmã Patrícia. Precisei me abster, em muito, de seu convívio, mas eu nunca esqueci de seu companheirismo, de sua amizade e do quanto de apoio tive neste momento da vida.

Minha esposa, Juliana, minha companheira, que tanto precisou abdicar de seu próprio lazer, de feriados e finais de semana, para que eu ficasse em casa totalmente dedicado e focado neste trabalho. Obrigado por tanto e a vida ao seu lado é, certamente, mais saborosa por isso.

Ao Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais, em especial ao campus de Juiz de Fora, por ter me apoiado no meu desejo de qualificar através do mestrado. Eu saí um tipo de servidor público e retorno agora bastante diferente, e desejo, profundamente, contribuir com a instituição e comunidade acadêmica para nosso aprimoramento.

Todos colegas de repartição, que aqui faço questão de nomear: Diego, Matheus, Bruno, Marcus e Jacqueline. Precisaram cobrir minhas atividades, e o fizeram com maestria, enquanto eu me dedicava ao mestrado. Vocês moram em meu coração.

Meus amigos e amigas pessoais que pouco me viram neste período, familiares também. Jamais esqueci de vocês e o quanto torciam por mim e pelo meu sucesso no cumprimento da atividade enquanto estive ausente do nosso viver. Anseio revê-los.

Leonardo, meu caro orientador. Obrigado por partilhar de seu conhecimento comigo. Suas dicas e observações irão comigo onde quer que eu vá, onde quer que eu esteja.

Aos meus colegas do PGMC: meu muito obrigado. Conhecer pessoas tão incríveis com certeza ajuda a compor a relevância deste trabalho e a amizade de vocês tornou a caminhada menos árdua.

Ao professor Celso Bandeira e à colega de projeto Paula Carvalho pelas longas e prazerosas conversas e, claro, debates sobre os trabalhos desenvolvidos, ideias e tudo mais.

E, finalmente, à Rhama Analysis pela gentileza em me conceder acesso ao seu principal sistema de dados hidrológicos quando mais precisei, à Agência Nacional de Águas e Saneamento Básico (ANA), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e ao Instituto Mineiro de Gestão das Águas (IGAM) pelo apoio financeiro concedido, processo APQ-01226-22.

"I propose to consider the question, 'Can machines think?' "
(Alan Turing em *I — COMPUTING MACHINERY AND INTELLIGENCE*, outubro de 1950)

RESUMO

A gestão sustentável dos recursos hídricos é essencial diante das mudanças climáticas e seus impactos nos sistemas humanos e naturais. Neste cenário, os modelos computacionais têm ajudado no processo de tomadas de decisões, entretanto, frente às limitações dos modelos físicos tradicionais, que frequentemente demandam dados extensivos e complexos processos de calibração, torna-se possível questionar: até que ponto modelos baseados em aprendizado de máquina podem oferecer previsões robustas e aplicáveis para a gestão hídrica em cenários caracterizados por alta variabilidade? Este estudo explora a aplicação de modelos de aprendizado de máquina, incluindo Regressão Linear, *CatBoost* e *Random Forest*, para previsão de vazão em bacias hidrográficas de Minas Gerais, Brasil. Por meio da análise das relações complexas entre precipitação e fluxo dos rios, esta pesquisa busca melhorar a precisão das previsões aplicando determinadas variáveis categóricas, abordando desafios como a assimetria dos dados e o efeito de "cauda longa". Os resultados destacam o potencial do aprendizado de máquina como uma solução eficiente e economicamente viável, com especial destaque à Regressão Linear, oferecendo percepções valiosas para a gestão de recursos hídricos e o planejamento de políticas públicas. Além disso, o estudo contribui para a literatura científica ao preencher uma lacuna de conhecimento sobre o uso de técnicas de aprendizado de máquina nesse contexto regional.

Palavras-chave: Aprendizado de Máquina. Modelagem chuva-vazão. Previsão de vazão. Gestão de recursos hídricos.

ABSTRACT

The sustainable management of water resources is essential in the face of climate change and its impacts on human and natural systems. In this context, computational models have supported decision-making processes; however, given the limitations of traditional physical models – which often require extensive data and complex calibration processes – it raises the question: to what extent can machine learning models provide robust and practical forecasts for water management in scenarios characterized by high variability and climate change? This study investigates the application of machine learning models, including Linear Regression, CatBoost, and Random Forest, for runoff forecasting in the hydrographic basins of Minas Gerais, Brazil. By analyzing the complex relationships between precipitation and river flow, this research aims to improve forecasting accuracy by incorporating specific categorical variables and addressing challenges such as data asymmetry and the "long tail" effect. The results highlight the potential of machine learning, particularly Linear Regression, as an efficient and cost-effective solution, offering valuable insights for water resource management and public policy planning. Furthermore, this study contributes to the scientific literature by filling a knowledge gap regarding the application of machine learning techniques in this regional context.

Keywords: Machine Learning. Rainfall-runoff modeling. Runoff forecasting. Water resources management.

RESUMEN

La gestión sostenible de los recursos hídricos es esencial frente al cambio climático y sus impactos en los sistemas humanos y naturales. En este contexto, los modelos computacionales han apoyado los procesos de toma de decisiones; sin embargo, dadas las limitaciones de los modelos físicos tradicionales, que a menudo requieren datos extensivos y procesos de calibración complejos, surge la pregunta: ¿en qué medida los modelos basados en aprendizaje automático pueden ofrecer predicciones robustas y aplicables para la gestión hídrica en escenarios caracterizados por alta variabilidad y cambio climático? Este estudio investiga la aplicación de modelos de aprendizaje automático, incluyendo Regresión Lineal, *CatBoost* y *Random Forest*, para la previsión de caudales en las cuencas hidrográficas de Minas Gerais, Brasil. A través del análisis de las complejas relaciones entre la precipitación y el flujo de los ríos, esta investigación busca mejorar la precisión de las predicciones incorporando variables categóricas específicas y abordando desafíos como la asimetría de los datos y el efecto de "cola larga". Los resultados destacan el potencial del aprendizaje automático, en particular la Regresión Lineal, como una solución eficiente y económicamente viable, ofreciendo perspectivas valiosas para la gestión de recursos hídricos y la planificación de políticas públicas. Además, este estudio contribuye a la literatura científica al llenar un vacío de conocimiento sobre el uso de técnicas de aprendizaje automático en este contexto regional.

Palabras clave: Aprendizaje automático. Modelado lluvia-caudal. Predicción de caudales. Gestión de recursos hídricos.

LISTA DE ILUSTRAÇÕES

Figura 1.1 – Carros revirados após as chuvas fortes em Mimoso do Sul–ES. 18
Figura 1.2 – Enchente devasta o Rio Grande do Sul em 2024. 18
Figura 2.1 – Critérios de inclusão e exclusão. 24
Figura 2.2 – Dados em formato tabular comumente visto em aprendizado de máquina. 36
Figura 2.3 – Dados com a dimensão temporal na coluna *"Time Stamp"*. 36
Figura 3.1 – As bacias estudadas e seus respectivos rios principais. 44
Figura 3.2 – Rio Jequitinhonha e as estações utilizadas. 46
Figura 3.3 – Rio Doce e as estações utilizadas. 47
Figura 3.4 – Rio Grande e as estações utilizadas. 48
Figura 3.5 – Rio São Francisco e as estações utilizadas. 49
Figura 3.6 – Regime médio da precipitação no intervalo 1991 - 2020. 50
Figura 3.7 – Série temporal incompleta da estação t_vz_54790000. 52
Figura 3.8 – Detalhe da série temporal da estação t_vz_54790000, ainda sem dados imputados, de 2013 a 2016. 53
Figura 3.9 – Detalhe da série temporal da estação t_vz_54790000, com dados imputados, de 2013 a 2016. 53
Figura 3.10–Série temporal incompleta da estação t_vz_54790000 no detalhe entre 2021 e 2022. 53
Figura 3.11–Série temporal completa da estação t_vz_54790000 no detalhe entre 2021 e 2022. 54
Figura 3.12–Série temporal completa da estação t_vz_54790000. 54
Figura 3.13–Série temporal incompleta da estação t_cv_54790000. 54
Figura 3.14–Série temporal completa da estação t_cv_54790000. 55
Figura 3.15–Série temporal completa da estação t_cv_01640000. 55
Figura 3.16–Série temporal completa da estação c_vz_56994500. 56
Figura 3.17–Série temporal da estação t_cv_56990850 - não utilizada. 57
Figura 3.18–Série temporal da estação t_cv_56994500 - não utilizada. 57
Figura 3.19–Série temporal completa da estação c_cv_01941010. 57
Figura 3.20–Série temporal completa da estação c_cv_01941004. 58
Figura 3.21–Série temporal completa da estação c_cv_01941006. 58
Figura 3.22–Série temporal completa da estação t_cv_56990005. 58
Figura 3.23–Série temporal incompleta da estação t_vz_62020080. 59
Figura 3.24–Série temporal completa (cortada) da estação t_vz_62020080. 59
Figura 3.25–Série temporal completa (cortada) da estação t_cv_61998080. 60
Figura 3.26–Detalhe do trecho com dados nulos da estação c_vz_44290002. 61
Figura 3.27–Série temporal completa da estação c_vz_44290002. 61
Figura 3.28–Série temporal completa da estação c_cv_01544017. 62

Figura 3.29–Série temporal completa da estação c_cv_01544032. 62
Figura 3.30–Série temporal completa da estação c_cv_01544036. 63
Figura 3.31–Autocorrelação para a vazão do rio Jequitinhonha. 67
Figura 3.32–Componente sazonal da série de vazão do rio Jequitinhonha. 68
Figura 3.33–Autocorrelação para a vazão do rio Doce. 68
Figura 3.34–Componente sazonal da série de vazão do rio Doce. 69
Figura 3.35–Autocorrelação para a vazão do rio Grande. 69
Figura 3.36–Componente sazonal da série de vazão do rio Grande. 70
Figura 3.37–Autocorrelação para a vazão do rio São Francisco. 70
Figura 3.38–Componente sazonal da série de vazão do rio São Francisco. 71
Figura 3.39–Dados originais para o rio Jequitinhonha. 71
Figura 3.40–Dados log-transformados para o rio Jequitinhonha. 72
Figura 3.41–Dados originais para o rio Doce. 72
Figura 3.42–Dados log-transformados para o rio Doce. 73
Figura 3.43–Dados originais para o rio Grande. 73
Figura 3.44–Dados log-transformados para o rio Grande. 74
Figura 3.45–Dados originais para o rio São Francisco. 74
Figura 3.46–Dados log-transformados para o rio São Francisco. 75
Figura 3.47–Fluxo de trabalho. 78
Figura 3.48–Diagrama mostrando a divisão dos dados de treino/teste com *refit*. . . . 79
Figura 3.49–WFV com janela expandida e *refit* - imagem 1. 80
Figura 3.50–WFV com janela expandida e *refit* - imagem 2. 80
Figura 3.51–WFV com janela expandida e *refit* - imagem 3. 80
Figura 3.52–Diagrama de como se calcula o intervalo de previsão. 81
Figura 3.53–Previsão com valor fora dos limites. 82
Figura 3.54–Previsão completamente nos limites. 82
Figura 4.1 – Resultado do *Seasonal Naive* no teste *Walk-Forward Validation*. 85
Figura 4.2 – *Walk-Forward Validation* para o modelo Regressão Linear. 86
Figura 4.3 – Detalhe do trecho final dos dados, em 2022, usados para treinamento. . 86
Figura 4.4 – Resíduos da previsão ao longo do tempo. 88
Figura 4.5 – Histograma dos resíduos. 88
Figura 4.6 – Resíduos da previsão ao longo do tempo. 89
Figura 4.7 – *Walk-Forward Validation* para o modelo *CatBoost*. 91
Figura 4.8 – *Walk-Forward Validation* para o modelo *Random Forest*. 91
Figura 4.9 – Dispersão dos resíduos ao longo do ano. 92
Figura 4.10–Dispersão dos resíduos ao longo do ano. 93
Figura 4.11–Histograma e curva-normal dos resíduos. 93
Figura 4.12–Histograma e curva-normal dos resíduos. 94
Figura 4.13–Gráfico ACF dos resíduos. 95

Figura 4.14–Gráfico ACF dos resíduos. 96
Figura 4.15–*Walk-Forward Validation* para o modelo Regressão Linear. 97
Figura 4.16–Dispersão dos resíduos ao longo do ano. 97
Figura 4.17–Histograma e curva-normal dos resíduos. 98
Figura 4.18–Gráfico ACF dos resíduos. 98
Figura 4.19–*Walk-Forward Validation* para o modelo *CatBoost*. 99
Figura 4.20–*Walk-Forward Validation* para o modelo *Random Forest*. 100
Figura 4.21–Dispersão dos resíduos ao longo do ano. 101
Figura 4.22–Dispersão dos resíduos ao longo do ano. 102
Figura 4.23–Histograma e curva-normal dos resíduos. 103
Figura 4.24–Histograma e curva-normal dos resíduos. 103
Figura 4.25–Gráfico ACF dos resíduos. 105
Figura 4.26–Gráfico ACF dos resíduos. 105
Figura 4.27–*Walk-Forward Validation* para o modelo *CatBoost*. 108
Figura 4.28–*Walk-Forward Validation* para o modelo *Linear Regression*. 108
Figura 4.29–*Walk-Forward Validation* para o modelo *Random Forest*. 109
Figura 4.30–Dispersão dos resíduos ao longo do ano. 109
Figura 4.31–Dispersão dos resíduos ao longo do ano. 110
Figura 4.32–Dispersão dos resíduos ao longo do ano. 110
Figura 4.33–Histograma e curva-normal dos resíduos. 111
Figura 4.34–Histograma e curva-normal dos resíduos. 112
Figura 4.35–Histograma e curva-normal dos resíduos. 112
Figura 4.36–Gráfico ACF dos resíduos. 113
Figura 4.37–Gráfico ACF dos resíduos. 114
Figura 4.38–Gráfico ACF dos resíduos. 114
Figura 4.39–*Walk-Forward Validation* para o modelo *CatBoost*. 116
Figura 4.40–*Walk-Forward Validation* para o modelo *Linear Regression*. 117
Figura 4.41–*Walk-Forward Validation* para o modelo *Random Forest*. 117
Figura 4.42–Dispersão dos resíduos ao longo do ano. 119
Figura 4.43–Dispersão dos resíduos ao longo do ano. 119
Figura 4.44–Dispersão dos resíduos ao longo do ano. 120
Figura 4.45–Histograma e curva-normal dos resíduos. 121
Figura 4.46–Histograma e curva-normal dos resíduos. 122
Figura 4.47–Histograma e curva-normal dos resíduos. 122
Figura 4.48–Gráfico ACF dos resíduos. 123
Figura 4.49–Gráfico ACF dos resíduos. 124
Figura 4.50–Gráfico ACF dos resíduos. 124
Figura 4.51–*Walk-Forward Validation* para o modelo *CatBoost*. 126
Figura 4.52–*Walk-Forward Validation* para o modelo *Linear Regression*. 126

Figura 4.53–*Walk-Forward Validation* para o modelo *Random Forest*. 127
Figura 4.54–Dispersão dos resíduos ao longo do ano. 128
Figura 4.55–Dispersão dos resíduos ao longo do ano. 129
Figura 4.56–Dispersão dos resíduos ao longo do ano. 129
Figura 4.57–Histograma e curva-normal dos resíduos. 130
Figura 4.58–Histograma e curva-normal dos resíduos. 130
Figura 4.59–Histograma e curva-normal dos resíduos. 131
Figura 4.60–Gráfico ACF dos resíduos. 132
Figura 4.61–Gráfico ACF dos resíduos. 133
Figura 4.62–Gráfico ACF dos resíduos. 133
Figura 4.63–Média dos valores SHAP - rio Jequitinhonha. 135
Figura 4.64–Média dos valores SHAP - rio Doce. 135
Figura 4.65–Média dos valores SHAP - rio Grande. 136
Figura 4.66–Média dos valores SHAP - rio São Francisco. 136
Figura 4.67–*Walk-Forward Validation* para o modelo *CatBoost* após remoção de variáveis. 138
Figura 4.68–*Waterfall plot* de uma previsão específica no rio Jequitinhonha. 139
Figura 4.69–Média dos valores SHAP - rio Jequitinhonha. 140
Figura 4.70–Média dos valores SHAP - rio Doce. 140
Figura 4.71–Média dos valores SHAP - rio Grande. 141
Figura 4.72–Média dos valores SHAP - rio São Francisco. 141
Figura 4.73–*Waterfall plot* de uma previsão específica no rio Doce. 142
Figura 4.74–*Walk-Forward Validation* para o modelo *Linear Regression* após remoção de variáveis. 143
Figura 4.75–Média dos valores SHAP - rio Jequitinhonha. 144
Figura 4.76–Média dos valores SHAP - rio Doce. 144
Figura 4.77–Média dos valores SHAP - rio Grande. 145
Figura 4.78–Média dos valores SHAP - rio São Francisco. 145
Figura 4.79–*Waterfall plot* de uma previsão específica no rio São Francisco. 146

LISTA DE TABELAS

Tabela 2.1 – Tabela de Definições de PICO(t). 22
Tabela 2.2 – Tabela com artigos removidos na fase de extração. 25
Tabela 2.3 – Hiperparâmetros para o modelo *CatBoost*. 39
Tabela 2.4 – Hiperparâmetros para o modelo *Random Forest* 40
Tabela 3.1 – Taxonomia do trabalho. 41
Tabela 3.2 – Estações usadas no rio Jequitinhonha. 45
Tabela 3.3 – Estações usadas no rio Doce. 46
Tabela 3.4 – Estações usadas no rio Grande. 47
Tabela 3.5 – Estações usadas no rio São Francisco. 48
Tabela 3.6 – Variáveis utilizadas - rio Jequitinhonha. 64
Tabela 3.7 – Variáveis utilizadas - rio Doce. 64
Tabela 3.8 – Variáveis utilizadas - rio Grande. 64
Tabela 3.9 – Variáveis utilizadas - rio São Francisco. 65
Tabela 3.10–Variáveis acumuladas para as precipitações. 65
Tabela 3.11–Configurações do computador usado nos experimentos. 83
Tabela 4.1 – Resultados antes e depois da remoção das variáveis - rio São Francisco. 137
Tabela 4.2 – Resultados antes e depois da remoção das variáveis - rio Doce. 142
Tabela 4.3 – Resultados dos modelos. 147

LISTA DE ABREVIATURAS E SIGLAS

MG	Minas Gerais
ANA	Agência Nacional de Águas e Saneamento Básico
IGAM	Instituto Mineiro de Gestão das Águas
IBGE	Instituto Brasileiro de Geografia e Estatística
CBH	Comitê de Bacia Hidrográfica
CB	Modelo *CatBoost*
LR	Modelo *Linear Regression* – Regressão Linear
RF	Modelo *Random Forest* – Floresta Aleatória
SN	Modelo *Seasonal Naive*
SVM	Modelo *Support Vector Machine* – Máquina de Vetores de Suporte
SVR	Modelo *Support Vector Regression* – Regressão por Vetores de Suporte
CART	Modelo *Classification and Regression Trees* – Árvores de Classificação e Regressão
ELM	Modelo *Extreme Learning Machine* – Máquina de Aprendizado Extremo
KRR	Modelo *Kernel Ridge Regression* – Regressão de Ridge com Kernel
ARIMA	Modelo *AutoRegressive Integrated Moving Average* – Modelo Auto-Regressivo Integrado de Média Móvel
SARIMA	Modelo *Seasonal ARIMA* – ARIMA Sazonal
SARIMAX	Modelo *Seasonal ARIMA with Exogenous* – ARIMA Sazonal com variáveis exógenas
kNN	Modelo *k-Nearest Neighbors* – k-Vizinhos mais próximos
RNA(s)	Rede(s) Neural(is) Artificial(is)
RNN	Modelo *Recurrent Neural Network* – Rede Neural Recorrente
LSTM	Modelo *Long Short-Term Memory* – Memória de Longo e Curto Prazo
FFNN	Modelo *Feedforward Neural Network* – Rede Neural *Feedforward*
MLP	Modelo *Multilayer Perceptron* – *Perceptron* Multicamadas
TFT	Modelo *Temporal Fusion Transformer* – Transformador de Fusão Temporal
ST(s)	Série(s) Temporal(is)
KGE	Métrica *Kling-Gupta Efficiency* – Eficiência de Kling-Gupta
MAPE	Métrica *Mean Absolute Percentage Error* – Erro Médio Absoluto Percentual
PBIAS	Métrica *Percent Bias* – Viés Percentual
RMSE	Métrica *Root Mean Square Error* – Raiz do Erro Quadrático Médio
NSE	Métrica *Nash-Sutcliffe Efficiency* – Eficiência de Nash-Sutcliffe
SHAP	*Shapley Additive Explanations* – Explicações Aditivas de Shapley
RFE	*Recursive Feature Elimination* – Eliminação Recursiva de Atributos
FFS	*Forward Feature Selection* – Seleção de Características Avançada

SUMÁRIO

1 INTRODUÇÃO 17
1.1 CONTEXTUALIZAÇÃO 17
1.2 JUSTIFICATIVA 19
1.3 PROBLEMA DE PESQUISA 19
1.4 OBJETIVOS 21
1.4.1 Objetivos gerais 21
1.4.2 Objetivos específicos 21
2 REVISÃO SISTEMÁTICA DA LITERATURA 22
2.1 QUESTÃO DE PESQUISA 22
2.2 BASES UTILIZADAS 23
2.3 FASE DE SELEÇÃO 23
2.4 FASE DE EXTRAÇÃO DE DADOS 25
2.5 SÍNTESE DOS TRABALHOS DE REFERÊNCIA 28
2.6 CONCEITOS FUNDAMENTAIS 35
2.6.1 Série temporal 35
2.6.2 Modelos de aprendizado de máquina 37
2.6.2.1 Seasonal Naive 37
2.6.2.2 Regressão linear - LR 38
2.6.2.3 CatBoost - CB 38
2.6.2.4 Floresta Aleatória - RF 39
3 PROCEDIMENTOS METODOLÓGICOS 41
3.1 TAXONOMIA DA SÉRIE TEMPORAL 41
3.2 DESCRIÇÃO DAS ÁREAS DE ESTUDO 42
3.3 DADOS UTILIZADOS 43
3.4 PRÉ-PROCESSAMENTO DOS DADOS 50
3.4.1 Rio Jequitinhonha 52
3.4.2 Rio Doce 56
3.4.3 Rio Grande 59
3.4.4 Rio São Francisco 61
3.5 VARIÁVEIS UTILIZADAS 64
3.6 ANÁLISE EXPLORATÓRIA DOS DADOS 66
3.7 MÉTRICAS DE AVALIAÇÃO 76
3.8 MODELO PROPOSTO 77
4 RESULTADOS E DISCUSSÃO 84
4.1 DESEMPENHO DOS MODELOS 84
4.1.1 Rio Jequitinhonha 84
4.1.2 Rio Doce 106

4.1.3 **Rio Grande** . . . 115
4.1.4 **Rio São Francisco** . . . 125
4.2 IMPORTÂNCIA DAS VARIÁVEIS . . . 134
4.2.1 **Catboost** . . . 135
4.2.2 **Regressão linear** . . . 138
4.2.3 **Random forest** . . . 143
4.3 ANÁLISE DOS RESULTADOS . . . 146
5 **CONCLUSÃO E PERSPECTIVAS** . . . **150**
5.1 CONCLUSÃO . . . 150
5.2 CONTRIBUIÇÕES PARA A ÁREA . . . 150
5.3 RECOMENDAÇÕES PARA TRABALHOS FUTUROS . . . 151
REFERÊNCIAS . . . **152**

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

A gestão dos recursos hídricos desempenha um papel crucial nas políticas públicas. Nos âmbitos socioeconômico, cultural e de saúde pública, em um contexto de mudanças climáticas e seus impactos nos sistemas humano e natural (55), conhecer a dinâmica dos recursos hídricos e entender como fatores externos impactam seu comportamento é de grande importância para os gestores públicos. A compreensão desses aspectos permite uma melhor tomada de decisões, garantindo a sustentabilidade dos recursos, a segurança hídrica e o bem-estar da população.

Neste sentido, prever a vazão de rios é um componente essencial na gestão de recursos hídricos, operação de reservatórios e mitigação de desastres naturais, e tem um caráter especial em regiões onde a hidroeletricidade desempenha um papel crucial na matriz energética, como o Brasil. Segundo o Balanço Energético Nacional de 2023, ano-base 2022, divulgado pelo Ministério de Minas e Energia, esta matriz energética representa cerca de 64% da oferta interna total de geração de energia elétrica (32). Desta forma, a previsão da vazão de rios que abastecem os reservatórios das hidrelétricas tem importância no impacto econômico que uma usina em baixa capacidade de geração pode causar.

Em uma perspectiva direcionada à população, os rios abastecem represas e açudes que fornecem água potável para consumo humano e animal, além de irrigar a lavoura – não apenas os rios, evidentemente, mas também a chuva, tem impacto significativo nesse cenário. Uma análise criteriosa da previsibilidade da vazão dos rios dias à frente permite ao poder público, por exemplo, reduzir ou até mesmo suspender outorgas para retirada de água destinadas a outros tipos de uso, visando o bem-estar populacional. Além disso, essa análise pode auxiliar no planejamento de regimes de racionamento. Basta lembrarmos do ano de 2015, quando se noticiava o "uso do volume morto" na Cantareira, no estado de São Paulo, pois a estiagem fora além do previsto e o abastecimento de cidades – da cidade de São Paulo propriamente – foram severamente afetados (37).

E quando se fala em bem-estar populacional, não podemos deixar de considerar os eventos climáticos extremos. Recordemos dos últimos desastres ocorridos nos estados de Minas Gerais (8), Rio de Janeiro (38), Espírito Santo (41), São Paulo (39), Bahia (7) e mais recentemente, em maio de 2024, a tragédia no Rio Grande do Sul (40) em que choveu $43mm$ em 6 horas – $43mm$ de chuva correspondem a 43 litros de água por metro quadrado (l/m^2) – (5), que trouxeram não apenas perda econômica como também perda de vidas.

Há que se pontuar ainda que populações carentes sofrem mais as consequências de tais eventos extremos, ao ponto de cerca de $3,6$ bilhões de pessoas estarem em condições de alta vulnerabilidade às mudanças climáticas, conforme descrito no relatório IPCC (54).

Figura 1.1 – Carros revirados após as chuvas fortes em Mimoso do Sul–ES.

Fonte: Foto por Fernando Madeira. (41)

Figura 1.2 – Enchente devasta o Rio Grande do Sul em 2024.

Fonte: Foto por Reuters/Amanda Perobelli. (6)

1.2 JUSTIFICATIVA

Este estudo contribui para a compreensão das dinâmicas de chuva-vazão nas bacias hidrográficas analisadas em Minas Gerais, especialmente em um contexto climático mundial que tornarão tais eventos ainda piores.(17) Ao aplicar modelos de aprendizado de máquina, como Regressão Linear (LR), *CatBoost* (CB) e *Random Forest* (RF), foi possível identificar padrões e relações complexas entre precipitação e vazão, proporcionando compreensão sobre o comportamento hidrológico dessas regiões. Esses avanços são particularmente relevantes, uma vez que as alterações climáticas impõem desafios adicionais na previsão e gestão dos recursos hídricos, exigindo abordagens mais sofisticadas e adaptáveis.

O conhecimento adquirido a partir desta pesquisa possui uma aplicação prática significativa, podendo ser diretamente empregado pelas autoridades competentes na tomada de decisões estratégicas para a gestão de recursos hídricos e mitigação de riscos de inundações. A capacidade de prever com boa precisão as vazões das bacias permite a implementação de medidas preventivas mais eficazes, otimização de infraestrutura hídrica e planejamento urbano mais resiliente. Além disso, as previsões aprimoradas podem informar políticas e estratégias públicas de adaptação às mudanças climáticas, promovendo uma gestão sustentável e eficiente dos recursos naturais.

Cabe destacar ainda que a revisão bibliográfica, considerando as bases consultadas, revelou uma lacuna na literatura referente à aplicação de técnicas de aprendizado de máquina nas bacias hidrográficas de Minas Gerais. Poucos estudos existentes exploraram essa interseção, indicando uma área de conhecimento com vastas possibilidades de expansão e desenvolvimento. Este fato sublinha a importância e a originalidade do trabalho, que não apenas preenche uma lacuna existente, mas também estabelece uma base para futuras pesquisas que possam explorar e aprimorar ainda mais a utilização de metodologias avançadas de aprendizado de máquina na análise hidrológica regional.

1.3 PROBLEMA DE PESQUISA

No campo dos recursos hídricos, a utilização de modelos físicos para estudar fenômenos hidrológicos é uma prática amplamente estabelecida. Modelos e ferramentas como o SMAP (*Soil Moisture Account Procedure*) (60), SWAT (*Soil and Water Assessment Tool*) (69) e HEC-HMS (*Hydrologic Engineering Center - Hydrologic Modeling System*) (46) são comumente empregados para simular e prever comportamentos de vazão em bacias hidrográficas. No entanto, esses modelos e ferramenta apresentam uma série de desafios que impõem limites à sua aplicabilidade.

Primeiramente, os modelos físicos são intrinsecamente complexos, exigindo uma vasta quantidade de dados de entrada para sua correta operação. Entre estes dados pode-se citar vazão, precipitação, evapotranspiração, uso e ocupação do solo, tipo de

solo, entre outros. Os dados também possuem diferentes escalas temporais e espaciais e sua obtenção pode ser dificultosa, resultando em lacunas significativas que podem comprometer a precisão das simulações. A coleta e integração desses dados demandam recursos substanciais, tanto em termos de tempo quanto de custo de manutenção, tornando a aplicação prática dos modelos físicos um empreendimento oneroso.

Além das dificuldades relacionadas à obtenção dos dados, os modelos físicos também enfrentam desafios no que concerne à calibração de parâmetros. A regulagem desses modelos depende diretamente da qualidade e abrangência dos dados de entrada, o que adiciona uma camada adicional de complexidade. A necessidade de otimização dos parâmetros de ajuste, frequentemente realizada mediante métodos computacionalmente intensivos, eleva ainda mais os custos associados à utilização desses modelos. Esse processo de otimização não apenas consome tempo, mas também requer infraestrutura computacional potente, limitando sua viabilidade em cenários onde recursos computacionais são restritos.

Em contraste, os modelos de aprendizado de máquina emergem como uma alternativa promissora frente aos desafios apresentados pelos modelos físicos. Esses modelos são orientados por dados e, em geral, possuem um número reduzido de parâmetros a serem otimizados, o que simplifica seu processo de implementação e calibração. A capacidade de lidar com grandes volumes de dados de forma eficiente e a flexibilidade para modelar relações complexas e não-lineares tornam o aprendizado de máquina altamente relevante para a previsão de vazões. Além disso, modelos de aprendizado de máquina podem gerar resultados de forma rápida, consumindo menos recursos computacionais em comparação com os modelos físicos, o que os torna particularmente atrativos para aplicações em larga escala e em tempo real.

No entanto, é crucial reconhecer que os modelos orientados por dados também possuem limitações. A qualidade das previsões geradas por esses modelos também está relacionada à qualidade dos dados de entrada utilizados no treinamento. Dados incompletos, ruidosos ou enviesados podem comprometer o desempenho dos modelos de aprendizado de máquina. Além disso, alguns tipos de modelos podem também requerer uma quantidade substancial de dados de treinamento para alcançar altos níveis de precisão, o que pode ser um obstáculo em regiões onde a coleta de dados é limitada. Apesar dessas limitações, a capacidade de gerar previsões rapidamente e a menor demanda por recursos computacionais fazem dos modelos de aprendizado de máquina uma ferramenta valiosa para a gestão de recursos hídricos.

Como ilustração prática, consideramos a aplicação do modelo de Regressão Linear para o rio Doce, onde foram obtidos os maiores conjuntos de dados de precipitação e vazão para treinamento. Utilizando um computador equipado com um processador Intel Core i5 8500 (oitava geração) e 8 *gigabytes* de RAM, o modelo de Regressão Linear levou aproximadamente 14 segundos para gerar os resultados. Embora a eficiência computacional

observada seja um indicativo positivo, a avaliação da qualidade dos resultados obtidos será aprofundada ao longo deste trabalho. Este exemplo serve para ilustrar que modelos de aprendizado de máquina, mesmo com configurações computacionais modestas, podem produzir resultados de maneira eficiente, destacando sua viabilidade para aplicações práticas em gestão de recursos hídricos.

Em suma, esta pesquisa visa resolver alguns dos problemas inerentes à utilização de modelos físicos na previsão de vazão, propondo a adoção de modelos de aprendizado de máquina como uma alternativa mais flexível e menos dispendiosa. Esta transição para abordagens orientadas por dados pode melhorar a capacidade de prever vazões em cenários caracterizados por alta variabilidade. A combinação dessas vantagens posiciona os modelos de aprendizado de máquina como ferramentas essenciais para a modernização e aprimoramento da gestão de recursos hídricos, promovendo uma abordagem mais adaptável e eficiente diante dos desafios climáticos contemporâneos.

1.4 OBJETIVOS

1.4.1 Objetivos gerais

Desenvolver e avaliar modelos de aprendizado de máquina para previsão de vazão em bacias hidrográficas de Minas Gerais, oferecendo uma alternativa eficiente e menos onerosa em comparação aos modelos físicos tradicionais.

1.4.2 Objetivos específicos

- Identificar e mitigar problemas de assimetria na distribuição dos dados das séries temporais de vazão e precipitação – efeito "cauda longa" –, objetivando aprimorar a precisão das previsões dos modelos de aprendizado de máquina;
- Realizar análise das variáveis de entrada aplicando o método *SHapley Additive exPlanations* (SHAP) e identificar quais destas tiveram maior impacto na previsão final dos modelos;
- Contribuir para a literatura científica ao preencher a lacuna existente na aplicação de técnicas de aprendizado de máquina nas bacias hidrográficas de Minas Gerais.

2 REVISÃO SISTEMÁTICA DA LITERATURA

Como forma de auxiliar o desenvolvimento do trabalho, uma revisão sistemática foi conduzida. Buscou-se por trabalhos que analisassem os rios ou bacias hidrográficas de Minas Gerais aplicando aprendizado de máquina em modelagem chuva-vazão, com intenção de realizar previsão de vazão na finalidade de antever possível situação de enchente.

Como visto em Donato and Donato (27), depreende-se que uma revisão sistemática é uma metodologia de investigação estruturada e replicável que visa compilar e analisar evidências de diversos estudos sobre uma questão de pesquisa específica. A relevância e importância da revisão sistemática, obedecendo a protocolos de seleção, inclusão e exclusão de trabalhos referentes à área, é o que garante a esta ser "abrangente, transparente e replicável", nas palavras das autoras Donato and Donato (27).

2.1 QUESTÃO DE PESQUISA

Para que se realize uma revisão sistemática é preciso definir uma questão de pesquisa. Para este trabalho, a questão principal foi obtida aplicando o critério de elegibilidade denominado **PICO** - sigla para **P**roblema ou **p**opulação, **I**ntervenção, **C**omparação e ***O****utcome*/resultado (87). Sobre a delimitação dessa questão de pesquisa, Galvão and Ricarte (42) nos diz que

> deve contemplar a especificação da população, ou do problema ou da condição que será estudada, o tipo de intervenção que será analisado, se haverá comparação entre intervenções e o desfecho que se pretende estudar (42)

Para este trabalho foi utilizada a extensão temporal do critério de elegibilidade, tornando-se, desta forma, **PICO(t)**. O 't' é o tempo. Veja a tabela PICO(t) (2.1) como ficou para este trabalho.

Tabela 2.1 – Tabela de Definições de PICO(t).

Categoria	**Definição**	**Descrição**
P	Problema	Enchentes urbanas em Minas Gerais cada vez mais intensas e frequentes
I	Intervenção	Aprendizado de máquina com variáveis categóricas para realizar a previsão de vazão em rios
C	Comparação/Controle	Comparar com estudos que realizaram a mesma tarefa utilizando aprendizado de máquina
O	*Outcome*	Avaliar a robustez dos resultados dos modelos selecionados para o trabalho
t	Tempo	Prever a vazão diária ao longo de 365 dias – 1 ano

Fonte: O autor

Tendo esta tabela como base, foi possível formular a seguinte pergunta: É viável empregar modelos de aprendizado de máquina na modelagem com variáveis categóricas que apresentem resultados robustos em simular a vazão ao longo de 365 dias adiante de determinados rios de Minas Gerais?

Com esta questão de pesquisa pronta, foi possível determinar as palavras-chave do trabalho: aprendizado de máquina, previsão de vazão, modelagem com variáveis categóricas e Minas Gerais. As pesquisas foram realizadas em bases internacionais, por isso os termos foram traduzidos para o inglês, ficando *machine learning*, *runoff forecasting* e *categorical features modeling*. Estas palavras foram combinadas utilizando os operadores booleanos *AND* (E) e *OR* (OU) para gerar a *string* de busca nas bases. Mas houve problemas com a especificidade do alvo do trabalho, buscando por rios ou bacias de Minas Gerais e também de emprego de modelagem com variáveis categóricas. Quando estes termos foram utilizados, houve poucas respostas nas bases e nenhuma que relacionasse, precisamente, com as palavras-chave almejadas. Foi preciso ampliar a pesquisa e abriu-se a possibilidade para rios do Brasil, cuja consulta retornou 30 resultados. Ainda ficou insatisfatório, pois não havia sido realizado nenhum procedimento de remoção neste ponto e esta quantidade poderia se reduzir drasticamente – em verdade, ficariam apenas 8 artigos após os cortes. O que permaneceu, enfim, foi o emprego de trabalhos restritos à América do Sul, em inglês, *South America*. Observou-se também que os termos *"machine learning"* e *"deep learning"* (aprendizado profundo) muitas vezes foram empregados como sinônimos e por isso acrescentou-se à pesquisa o termo *"deep learning"*. Apenas artigos dos últimos 5 anos foram utilizados, perfazendo os anos de 2020 a 2024 até a data de início de escrita deste trabalho, em março.

2.2 BASES UTILIZADAS

As bases utilizadas nesta revisão estão disponíveis na internet e foram acessadas via plataforma Periódicos da Capes (Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (1)). Foram utilizadas 3 bases: *Scopus*, *Science Direct*, *Web of Science*. O processo de busca nas bases resultou em 355 trabalhos para a fase de seleção.

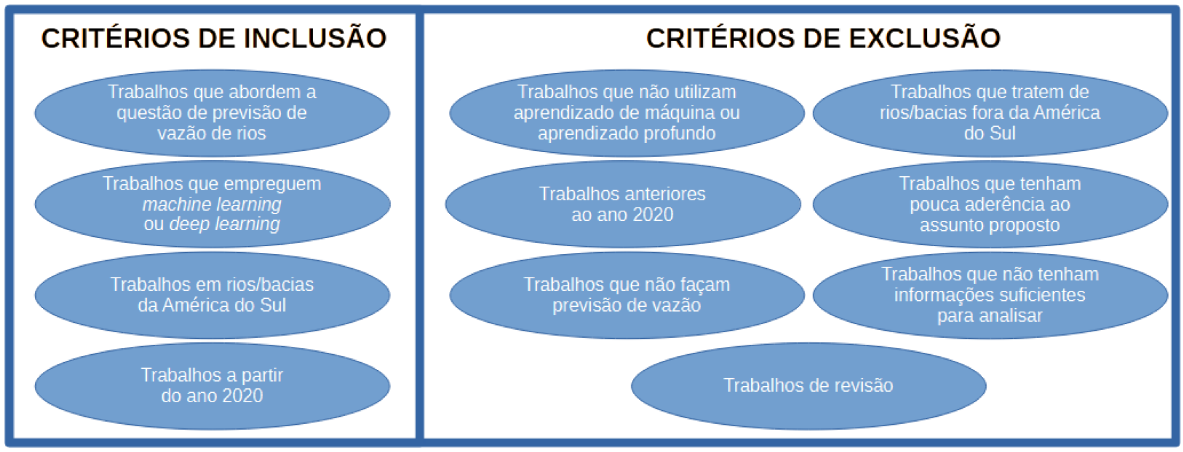
2.3 FASE DE SELEÇÃO

Bases consultadas, resultados prontos, partiu-se para a seleção dos trabalhos que comporiam o arcabouço de conhecimento deste trabalho. Os seguintes tipos de trabalho foram aceitos: artigos publicados em jornais e revistas; dissertações e teses. Não foram aceitos trabalhos dos seguintes tipos para a composição da revisão sistemática: artigos de conferência, livros ou capítulos de livros. Após essa análise inicial, foram identificados 21 trabalhos, de um total de 355, fora dos tipos aceitos, sendo então removidos.

Os 334 trabalhos restantes passaram pela etapa de seleção, em que se verificou as possíveis duplicatas. Ao todo 82 duplicatas foram identificadas e devidamente removidas. Após remoção de duplicatas, avaliou-se o ano de publicação dos artigos, em que 74 foram removidos por serem anteriores ao ano de 2020, o ano-base adotado nesta revisão. Em seguida, foram lidos os seguintes campos: título, resumo e palavras-chave. A partir de então, era verificado se o trabalho atendia aos critérios de inclusão ou exclusão definidos. A figura 2.1 mostra os critérios adotados para a inclusão ou a exclusão. Foi estipulado que um trabalho só seria incluído se atendesse aos três critérios simultaneamente. Porém, para ser excluído, bastava atender apenas a um dos critérios de exclusão.

Neste ponto de leitura, após toda resolução, restaram 178 artigos para análise. Deste montante, 34 trabalhos corresponderam ao assunto tratado na pesquisa, atendendo aos critérios de seleção estipulados, e passaram para a fase de extração.

Figura 2.1 – Critérios de inclusão e exclusão.



Fonte: O autor

2.4 FASE DE EXTRAÇÃO DE DADOS

A etapa de extração de dados compreendeu a leitura integral dos artigos e a extração das informações pertinentes. As questões listadas a seguir foram formuladas para auxiliar no processo de extração dos dados provenientes dos artigos.

- Quais os objetivos do trabalho?
- Os rios ou bacias analisadas situam-se na América do Sul?
- É um trabalho que realiza previsão (*forecasting*) de vazão?
- O trabalho está se valendo de modelos de aprendizado de máquina (*machine learning*) ou de aprendizado profundo (*deep learning*)?
- Quais são os resultados do trabalho?

Durante a fase de extração, outros 15 artigos foram desconsiderados para o trabalho, restando 18 trabalhos totalmente aderentes ao mote desta pesquisa. Foi uma redução significativa, mas após análise o que se concluiu foi que o foco desta pesquisa se insere em um nicho muito específico dentro do que se realiza em trabalhos na área. O simples fato de, necessariamente, se buscar trabalhos que empreguem *machine learning* ou *deep learning*, fez com que 9 artigos precisassem ser retirados da revisão sistemática. Estes 15 artigos removidos estão na tabela 2.2, com observações que se fizeram relevantes e justificativas sobre a não consideração dos mesmos.

Tabela 2.2 – Tabela com artigos removidos na fase de extração.

Referência	**Observações e Justificativas**
Kolling Neto et al. (56)	Removido, pois aplica um modelo hidrológico, ou seja, modelo físico, de escala continental.
Páliz Larrea et al. (76)	Removido, pois faz previsão do nível da água em um reservatório e não de vazão de rio.
Singh and Goyal (88)	Trabalha por uma compreensão do risco financeiro relacionado ao clima. Empreende esforço para incluir informações críticas sobre os riscos prováveis relacionados às mudanças climáticas na tomada de decisões financeiras.
Yan et al. (95)	Aplica o modelo hidrológico – físico – *Dominant River Routing Integrated with VIC Environment*, DRIVE.

Referência	**Observações e Justificativas**
Stefenon et al. (91)	É um trabalho muito recente, de fevereiro de 2024, aplicando uma técnica bastante interessante na área de previsão de séries temporais, o modelo de aprendizado profundo *Temporal Fusion Transformer*, TFT – uma versão com automatização para busca de hiperparâmetros AutoTFT –, contudo o direcionamento do artigo é todo para a previsão de nível de água em barragens em eventos de inundação.
Bordin et al. (4)	Este estudo explorou o papel atual e futuro do aprendizado de máquina, *machine learning*, no setor hidrelétrico, com ênfase no agendamento de operação de usinas. O objetivo principal foi revisar as aplicações existentes de *machine learning* na literatura e propor direções futuras de pesquisa para abordar lacunas ainda não exploradas.
Fernandez et al. (34)	Este trabalho propõe a criação de um *framework*, baseado em um modelo hidrológico simplificado, para realizar o balanceamento hídrico em cabeceira de bacias em regiões com escassez de dados. O *framework* apresentado é bastante interessante e pode, com adaptações, ser possivelmente aplicado com aprendizado de máquina. Ele apresenta um fluxo para identificação e seleção de estações relevantes de precipitação, vazão e nível de rio, posterior conversão de nível de rio para vazão, aplicando modelos matemáticos, e final agregação dos dados que irá alimentar o modelo hidrológico simplificado proposto.
Paez-Trujilo et al. (71)	Calibra e valida o modelo SWAT para realizar simulações do balanço hídrico no rio Cesar, Colômbia. Com estes dados simulados os autores realizam uma análise de seca, o objeto de estudo da pesquisa, e só então aplicam *machine learning* visando explicabilidade na relação entre as características hidroclimáticas e severidade de secas.

Referência	**Observações e Justificativas**
Liu et al. (59)	Aplica um modelo hidrológico combinado com *machine learning* – em verdade, apesar de os autores usarem a expressão *"machine learning"* na realidade aplicam um modelo de *"deep learning"*, rede *Long Short-Term Memory*, LSTM – para gerar uma análise em escala global sobre a magnitude de inundações, ou seja, a América do Sul é apenas tangencialmente tratada, o que foge do objetivo desta pesquisa.
Siqueira et al. (89)	O trabalho aplica um modelo hidrológico-hidrodinâmico para realizar previsão de vazão, em escala continental, em determinados rios e bacias do continente.
Amorim et al. (2)	Calibra o modelo SWAT, um modelo físico, para a bacia hidrográfica do rio Tocantins-Araguaia
de Jong et al. (24)	Aqui é aplicado três diferentes modelos climáticos globais para realizar simulações de longo prazo, para os anos 2030 e 2080, em cenário de elevada emissão de carbono na atmosfera. O objetivo é avaliar o impacto da mudança climática no potencial hidrelétrico em várias das bacias da América do Sul.
Carvalho et al. (11)	A pesquisa aplica modelos de aprendizado de máquina e de aprendizado profundo para estudo de padrões de demanda de água na cidade de Fortaleza, Ceará. Não é um estudo de previsão de vazão.
Pereira et al. (73)	O trabalho combina a aplicação de modelo hidrológico e de clima para dar suporte à operação de reservatórios com vista à gestão e alocação de água em região semiárida.
Rodríguez et al. (82)	Aplica aprendizado de máquina e aprendizado profundo exclusivamente direcionado à análise de qualidade de água – preenchimento de dados faltantes. É, no entanto, um estudo bastante interessante, pois visa regiões com dados escassos para qualidade de água.

Fonte: O autor

2.5 SÍNTESE DOS TRABALHOS DE REFERÊNCIA

Esta seção sintetiza os principais aspectos de cada artigo incluído nesta revisão sistemática, proporcionando uma visão abrangente do conteúdo abordado em cada estudo. Cada resumo delineia de maneira concisa os objetivos específicos das pesquisas e destaca os resultados mais relevantes alcançados, facilitando a compreensão das contribuições individuais para o campo de estudo. Ao apresentar essas sínteses, a seção permite uma análise comparativa e crítica dos trabalhos revisados, contribuindo para uma melhor interpretação das tendências, metodologias adotadas e descobertas significativas no contexto investigado.

O estudo Meis et al. (65) aborda a aplicação de uma ferramenta estatística para previsão hidrometeorológica na bacia inferior do rio La Plata. O objetivo principal foi melhorar a capacidade de prever eventos extremos de vazão, como cheias severas associadas ao fenômeno El Niño, cujo impacto socioeconômico é bastante relevante na região. Séries temporais de vazão foram analisadas utilizando modelos estatísticos ARIMA Sazonal (SARIMA) e ARIMA Sazonal com variáveis exógenas (SARIMAX), com a incorporação de variáveis exógenas derivadas do índice NIÑO 3.4 para melhorar a precisão das previsões, especialmente durante eventos extremos. Os resultados mostraram que o modelo SARIMAX apresentou maior eficácia na previsão de eventos extremos de vazão, em comparação ao SARIMA, durante períodos influenciados pelo El Niño. Em condições gerais, o desempenho dos dois modelos foi semelhante, mas o SARIMAX destacou-se na previsão de cenários críticos.

Vilanova et al. (93) realiza um trabalho de calibração e transferência de modelos baseados em *deep learning*, mais exatamente Redes Neurais Artificiais (RNAs), para a simulação de vazões em sub-bacias da Mata Atlântica brasileira, sendo o foco a bacia do rio Itapemirim. As redes neurais empregadas foram do tipo MLP. Os objetivos principais foram (I) calibrar RNA(s) para estimar vazões diárias específicas em três sub-bacias de tamanhos variados e (II) avaliar a viabilidade da transferência dessas RNA(s) para outras sub-bacias da mesma bacia hidrográfica. O estudo utilizou dados diários de 32 anos de precipitação e vazão consistidos em 12 sub-bacias. As RNA(s) foram eficazes para simular vazões nas sub-bacias para as quais foram calibradas, com desempenhos satisfatórios ao serem transferidas para outras sub-bacias. A estimativa de vazões máximas apresentou limitações, atribuídas à escala temporal diária dos dados de precipitação e à variabilidade hidrológica e as diferenças na resposta hidrológica entre sub-bacias grandes e pequenas destacaram a necessidade de modelos ajustados às condições específicas de cada bacia.

O trabalho conduzido por Korsic et al. (57) foi realizado na bacia do rio Tupungato, nos Andes Centrais da Argentina, e avaliou a eficácia da técnica *Support Vector Regression* (SVR) para prever a vazão mensal com antecedência de um mês. O foco foi na gestão dos recursos hídricos em uma região que é dependente do derretimento de neve, afetada pelas

mudanças climáticas e escassez de dados hidrológicos. O estudo comparou a eficiência do SVR com outros modelos, como o *Classification and Regression Trees* (CART) e o tradicional modelo ARIMA. O SVR alcançou coeficientes de determinação (R^2) entre $0,75$ e $0,89$. Simulações com entradas mais completas – incluindo defasagens e média sazonal de vazão – mostraram os melhores desempenhos. O modelo SVR superou significativamente o modelo CART em todos os cenários e também apresentou melhores resultados do que o ARIMA. A área de cobertura de neve, obtida a partir do satélite MODIS, com 5 meses de defasagem – *lags* – foi a variável mais correlacionada à vazão, ao passo que a precipitação foi considerada irrelevante para o modelo.

O estudo (94) teve como objetivo principal avaliar a importância de variáveis de chuva-vazão no desempenho de modelos de previsão de escoamento diário utilizando o modelo *Random Forest* (RF). Este trabalho foi realizado na bacia hidrográfica Santa Lucía Chico, no Uruguai, que possui relevância estratégica devido ao abastecimento de água para grande parte da população do país. O trabalho objetivou compreender o impacto dos parâmetros físicos de entrada no desempenho do modelo RF, determinar quais variáveis possuem maior importância para simular a vazão diária e, finalmente, avaliar os pontos fortes e fracos dos métodos adotados no estudo. Os dados de entrada foram séries temporais de vazão, precipitação, temperaturas mínima e máxima, radiação solar, velocidade do vento, umidade relativa e evapotranspiração potencial, o que se assemelha, em parte, ao trabalho realizado nesta pesquisa. O trabalho dos autores também conduziu uma análise com o método SHAP para realizar interpretação global e local da influência de cada variável. Os principais resultados foram:

- Q_{t-1}, *lag* de 1 dia da vazão, e Q_{t-2} foram as variáveis mais importantes
- A precipitação acumulada nos últimos 7 dias (P_{accum}) foi variável eficiente para as condições de umidade antecedente do solo
- Apenas o uso da temperatura do ar (T_{max} e T_{min}) foi suficiente para substituir todas as variáveis climáticas, sem perda de desempenho

Este foi o trabalho mais aderente ao tema da pesquisa realizada nesta dissertação e as conclusões vistas aqui neste estudo se assemelham em alguns pontos às conclusões da dissertação.

Gomaa et al. (43) analisa o uso de algoritmos híbridos de aprendizado de máquina para prever a vazão diária no reservatório de Três Marias, localizado na bacia do Rio São Francisco, no Brasil. A pesquisa utilizou dados de chuva obtidos pela missão *Tropical Rainfall Measuring Mission* (TRMM), e dados de vazão observados, abrangendo um período de 22 anos, de 1998 a 2019. O trabalho avalia a relação entre chuva e escoamento utilizando modelos de aprendizado de máquina, como Rede Neural de Função de Base

Radial Gaussiana (GRNN), Regressão por Processo Gaussiano (GPR), e *Multilayer Perceptron* (MLP) otimizado por Otimização por Enxame de Partículas (MLP-PSO), todas as siglas no original em inglês. Outra tarefa desempenhada foi investigar a eficácia da integração da Decomposição Empírica de Modo, EMD em inglês, com a Transformada de Hilbert-Huang, HHT também em inglês, buscando melhorar a previsão de vazões. As conclusões que os autores chegaram foi que o modelo MLP-PSO obteve o melhor desempenho em termos de precisão, com o menor *Root Mean Square Error* (RMSE), superando os modelos GRNN e GPR em diferentes combinações de entradas. A integração de EMD-HHT ao MLP-PSO, denominado MLP-PSO-EMD, aumentou significativamente a precisão na previsão de vazões, particularmente durante a fase de testes. O modelo MLP-PSO-EMD atingiu um RMSE de $60,03 m^3/s$ e um *Nash-Sutcliffe Efficiency* (NSE) de $0,96$.

O estudo desenvolvido por Fernando De Toledo et al. (35) explora o impacto de índices climáticos na previsão de séries de vazão mensal em reservatórios de usinas hidrelétricas no Brasil. O objetivo principal foi investigar se a inclusão de índices climáticos, mais especificamente 27 índices, melhora a precisão dos modelos de previsão, especialmente no contexto de mudanças climáticas que afetam os padrões hidrológicos. Comparou-se o desempenho de 3 modelos de aprendizado de máquina – os modelos SVR, *Extreme Learning Machine* (ELM) e *Kernel Ridge Regression* (KRR) – e um modelo estatístico – SARIMAX – na previsão de vazões, com e sem o uso de índices climáticos. Os resultados mostraram que os modelos que incorporaram índices climáticos apresentaram uma redução significativa nos erros de previsão, em especial em regiões equatoriais – norte e nordeste do Brasil –, onde a correlação entre índices climáticos e vazões foi mais forte. Em média, $69,5\%$ das previsões feitas com emprego de modelos de aprendizado de máquina tiveram melhorias quando índices climáticos foram incluídos, com ganhos superiores a 10% em 48% das previsões realizadas, quase metade dos testes analisados.

Saraiva et al. (84) desenvolveram um trabalho em que aborda a previsão de vazão diária no reservatório de Sobradinho, localizado no nordeste do Brasil, utilizando modelos de aprendizado de máquina acoplados a transformadas *wavelet* e técnicas de *bootstrapping*. Os autores analisaram dois modelos de aprendizado de máquina, RNA e SVM, combinados com técnicas de pré-processamento de dados ora mencionadas. Aferiu-se a eficácia destes métodos em conjunto com as transformações para prever vazões de rios e identificaram combinações que aumentassem a precisão e robustez destas previsões. O modelo híbrido *Boostrap-Wavelet-Neural-Network* (BWNN) foi identificado como a abordagem mais promissora para previsões de vazões diárias em séries temporais hidrológicas. Os resultados obtidos a partir desta abordagem proporcionaram avanços na precisão e robustez das previsões, com potencial para aplicação em outros estudos hidrológicos e séries temporais ambientais.

Em Nogueira Filho et al. (70) foi abordado o desafio de modelar vazão em bacias

não monitoradas – *ungauged basins*, no original –, utilizando *Recurrent Neural Network* (RNN) do tipo LSTM para regionalização de vazões em regiões semiáridas, especificamente no estado do Ceará, Brasil. O objetivo principal foi avaliar a eficácia das redes LSTM em regiões com dados limitados em comparação com redes *Feedforward Neural Network* (FFNN) tradicionais e o modelo hidrológico conceitual *Soil Moisture Accounting Procedure* (SMAP). Um procedimento adotado foi investigar se alterações simples nas FFNN, como a inclusão de *lags* de chuva e vazão, podem, eventualmente, igualar ou até mesmo superar o desempenho das redes LSTM. Os principais resultados foram:

- As redes LSTM demonstraram desempenho superior ao modelo conceitual SMAP nos experimentos;
- Uma FFNN com *lags* de vazão, mais especificamente a FFNN-2 com 3 meses de *lags* de precipitação e 2 meses de *lags* de vazão observada, apresentou o melhor desempenho geral, superando a LSTM em termos de eficiência, apresentando um valor NSE de $0,866$ contra $0,808$ da LSTM;
- A inclusão de memória de curto prazo mostrou ser altamente eficiente em regiões semiáridas, devido à curta memória hidrológica nessas áreas.

Continuando, o estudo apresentado por Reis et al. (78) explorou a aplicação de aprendizado de máquina para prever o fluxo diário de rios, focando nos métodos de seleção de covariáveis ambientais para melhorar a precisão e a estabilidade dos modelos. A pesquisa foi conduzida na bacia do rio Piranga, Minas Gerais, usando dados históricos de 18 anos, de 2000 a 2017. Três modelos de aprendizado de máquina foram aplicados no estudo: *Multivariate Adaptive Regression Splines*, variante EARTH (MARS-EARTH), Regressão Linear Múltipla e RF. E também aplicou 2 métodos de seleção de covariáveis: *Recursive Feature Elimination* (RFE), que elimina recursivamente variáveis menos relevantes e avalia o resultado do modelo, se melhora ou não, e *Forward Feature Selection* (FFS), cujo procedimento é adicionar variáveis e avaliar o comportamento do modelo a cada etapa. Todos os modelos apresentaram bom desempenho geral – coeficiente NSE acima de $0,75$ e coeficiente de determinação, R^2, acima de $0,80$ –, mas demonstraram limitações na previsão de eventos extremos de vazão máxima. As covariáveis mais importantes destacadas neste trabalho foram a vazão de rios afluentes, precipitação acumulada, porcentagem de cobertura florestal e plantação, índice EVI (*Enhanced Vegetation Index*) e o mês representando a sazonalidade dos eventos. O método FFS mostrou-se mais vantajoso, reduzindo o número de covariáveis necessárias para alguns modelos e melhorando o desempenho em comparação ao RFE. Já entre os modelos avaliados, o RF foi o mais robusto, enquanto o MARS-EARTH apresentou maior instabilidade. O trabalho concluiu, por fim, que modelos baseados em aprendizado de máquina são eficazes para prever vazões diárias de rios e oferecem melhor desempenho em comparação com metodologias tradicionais e que o método FFS deve ser

mais considerado em estudos da área devido à sua capacidade de melhorar o desempenho dos modelos com menos covariáveis, reduzindo custos computacionais e riscos de sobreajuste – *overfitting*.

No estudo realizado por Lujano et al. (62) foi abordada uma modelagem hidrológica utilizando o modelo kNN para prever as vazões médias diárias do rio Ramis, no Peru. O estudo destacou a importância do trabalho, pois considerou que essa previsão realizada seria essencial para sistemas de alerta e gestão de inundações. Dados hidrometeorológicos, incluindo precipitação média da bacia e vazões médias diárias, foram utilizados como entradas para o modelo, considerando certas defasagens temporais – *lags*. Os resultados sugeriram que o modelo *k-Nearest Neighbors* (kNN) é robusto para previsões de curto prazo – granularidade diária no estudo –, especialmente ao usar dados de vazões com *lags* de 1 e 2 dias e precipitação com *lag* de 3 dias. Algumas combinações de modelagem foram realizadas e a melhor, a que obteve o modelo mais eficaz, incluiu como entradas a *lag* de vazão de 1 dia (Q_{t-1}) e a precipitação média com *lag* de 3 dias (M_{t-3}), obtendo um valor NSE de $0,979$. Por fim, a pesquisa destacou que, apesar de simples, o modelo kNN é eficaz para fortalecer sistemas de previsão hidrológica e pode complementar modelos baseados em processos físicos, com menor custo e complexidade de implementação.

Rodrigues et al. (80) realizou a modelagem hidrológica da bacia do rio Manuel Alves da Natividade, localizada no bioma Cerrado brasileiro, para auxiliar a gestão dos recursos hídricos na região. Duas abordagens foram comparadas:

- Modelo hidrológico baseado em processos: SWAT;
- Modelo orientado por dados: RNA.

Como dados de entrada, o artigo utilizou precipitação, temperaturas máxima e mínima, radiação solar, umidade relativa do ar, velocidade do vento e dados históricos de vazão. O período de treinamento foi o intervalo de dados correspondente entre 1986 e 1995 e validação foi entre 1996 e 2005. Para alimentar a RNA também empregou-se o uso de *lags* de 1 dia de vazão – Q_{t-1}. Os resultados apontaram que ambos os modelos são adequados para simular vazões diárias, tendo a RNA o melhor desempenho na métrica NSE de $0,91$, superando o SWAT. Os autores terminam por concluir que, apesar de as análises terem indicado que a RNA é mais precisa, o modelo SWAT ofereceu maior detalhamento dos processos hidrológicos, sendo mais indicado para estudos envolvendo mudanças de uso do solo ou cenários climáticos.

Torres et al. (92) introduz um *framework* baseado em Modelagem por Média Bayesiana – *Bayesian Model Averaging*, BMA – para previsão de vazões em sistemas dominados pela geração hídrica, utilizando uma combinação de modelos hidrológicos e de aprendizado de máquina. A pesquisa abordou a importância de considerar as incertezas nas previsões de vazões e seu consequente impacto direto no planejamento e na operação

de sistemas hidrelétricos. Foram utilizados dados de entrada como precipitação e medições históricas de vazão para calibrar os modelos Modelo de Grandes Bacias, – sigla MGB, este sendo um modelo hidrológico –, SMAP/ONS e RNA(s) do tipo MLP. Os resultados mostraram que o *framework* proposto pelos autores melhorou a precisão das previsões de vazões em comparação aos modelos individualmente, com destaque para as regiões Sul e Sudeste do Brasil. As atualizações semanais dos pesos com base nos últimos 120 dias de dados melhoraram a acurácia e foi o grande destaque da abordagem BMA.

Em de Sousa et al. (25) o estudo explorou a previsão de vazões diárias, variando de Q_{t+1} até Q_{t+3}, em uma região de transição entre biomas de Savana e Amazônia, no Brasil, utilizando modelos de aprendizado de máquina. Foram empregados 4 modelos: kNN, SVM, RF, e RNA do tipo MLP. A pesquisa focou na previsão de vazões diárias com base em variáveis explicativas, como chuvas médias e *lags* de vazões, extraídas tanto de dados observados em estações pluviométricas quanto de produtos de sensoriamento remoto, incluindo PDIR-Now, PERSIANN e PERSIANN-CCS. Os resultados indicaram que o modelo RF foi o mais eficaz, apresentando menor erro na previsão. A precisão das previsões foi consistente para horizontes de até 3 dias, mesmo em bacias com dados hidrológicos limitados. Os autores destacaram, desta forma, o potencial do aprendizado de máquina para apoiar o planejamento e a gestão de recursos hídricos, com aplicações específicas em usinas hidrelétricas e para mitigação de enchentes.

A abordagem em Duarte et al. (29) foi a previsão de vazão diária nas bacias dos rios Sono, Manuel Alves da Natividade e Palma, localizadas na região hidrográfica Tocantins-Araguaia, Brasil. O objetivo principal foi selecionar as melhores variáveis de entrada para modelos de aprendizado de máquina aplicando algoritmos de seleção – genético, RFE e VSURF – e avaliar o desempenho de quatro modelos: RF, RNA do tipo MLP, SVR e *M5 Model Tree*, M5T. As variáveis utilizadas incluíam *lags* de vazão, precipitação média e evapotranspiração. Os resultados demonstraram que o modelo M5T apresentou o melhor desempenho geral, particularmente com variáveis selecionadas pelo método RFE nas bacias do Sono e Palma. O estudo também destacou limitações, como a baixa densidade de estações pluviométricas para capturar a variabilidade espacial da precipitação.

O estudo realizado por Essenfelder et al. (33) propôs o desenvolvimento de uma ferramenta de nome *Smart Climate Hydropower Tool* (SCHT) uma solução de previsão sazonal híbrida para auxiliar na gestão de reservatórios de hidrelétricas. A ferramenta proposta utiliza dados climáticos sazonais do *Copernicus Climate Data Store* (CCDS) e combina diferentes algoritmos de aprendizado de máquina para prever o volume acumulado de descarga de afluentes. Os horizontes de previsão considerados variaram de 1 a 6 meses, com resultados avaliados para diferentes períodos. O estudo utilizou lags de 1 a 6 meses como variáveis de entrada para os modelos de previsão e estas foram incorporadas para capturar a memória temporal do sistema hidrológico e melhorar a precisão dos resultados. Os algoritmos de aprendizado de máquina, incluindo rede LSTM, e NARX, demonstraram

alta precisão ao prever a dinâmica temporal de vazão acumulada. A ferramenta SCHT foi eficaz para gerenciar recursos hídricos e planejamento financeiro, com estimativas econômicas mostrando ganhos financeiros potenciais volumosos em cenários simulados. O trabalhou destacou ainda que os modelos com arquiteturas complexas ou recorrentes apresentaram os melhores desempenhos em previsões de longo prazo, superando médias climatológicas e regressões lineares.

Filho et al. (36) apresentou no trabalho o desenvolvimento de modelo do tipo MLP para previsão de enchentes no Rio Branco, situado na bacia Amazônica. Foram utilizados dados de níveis fluviais e precipitação média acumulada em 48 horas, derivados do produto de sensoriamento remoto PDIR-Now. A pesquisa empregou RNA(s) cuja arquitetura foi otimizada com o algoritmo *Shuffled Complex Evolution - University of Arizona*, SCE-UA, resultando em previsões para horizontes de 6, 12 e 24 horas. As RNA(s) demonstraram alta eficácia, com índices NSE superiores a $0,9$ geralmente, especialmente quando a precipitação acumulada foi incorporada como variável de entrada para previsões de 24 horas. O estudo concluiu que os modelos são adequados para sistemas de alerta contra enchentes em regiões com baixa densidade de estações pluviométricas, como o caso da região estudada.

Chegando no trabalho de Pastrana-Cortés et al. (72), este abordou a previsão de recursos hidrológicos, como volume útil e contribuições de vazão, com foco em 23 reservatórios na Colômbia. Utilizando um modelo probabilístico denominado *Multi-Output Variational Gaussian Process* (MOVGP) o estudo propôs uma metodologia que explorou relações entre séries temporais para capturar as incertezas inerentes aos dados hidrológicos. O MOVGP foi comparado com modelos clássicos, como RNA do tipo LSTM e *Linear AutoRegression* (LAR) demonstrando superioridade, especialmente para horizontes de previsão mais longos. Os principais resultados apresentados no trabalho incluem:

- MOVGP apresentou menor erro médio quadrático, MSE, em previsões de médio prazo, mostrando-se mais flexível na captura de interações não-lineares;
- O modelo forneceu intervalos de confiança úteis para detecção de anomalias;
- MOVGP adaptou-se melhor a diferentes escalas de variabilidade nos dados hidrológicos.

Por fim, em Ribeiro et al. (79) houve a aplicação de ELM e *Echo State Networks* (ESN) para prever séries temporais de vazão. Um novo método de treinamento foi proposto para ESN, permitindo o uso de *bagging – bootstrap aggregation*. Utilizou-se, ainda, otimização multiobjetivo para ajustar os pesos dos modelos-base no *ensemble*, buscando equilíbrio entre viés e variância. O trabalho utilizou séries históricas de vazão de 5 usinas hidrelétricas brasileiras, Sobradinho, Serra da Mesa, Jiraú, Furnas e Água

Vermelha, de 1935 a 2010. Compararam-se os modelos ELM e ESN individualmente, como *ensembles* e *ensembles* otimizados via otimização multiobjetivo em horizontes de previsão de 1, 3, 6 e 12 meses. Os resultados ressaltaram que os *ensembles* otimizados, ESN-MOB e ELM-MOB, apresentaram melhor desempenho global segundo as métricas do trabalho, principalmente para previsões de longo prazo, de 6 e 12 meses. Os *ensembles* com ESN otimizados superaram modelos tradicionais como ARIMA e SARIMA.

2.6 CONCEITOS FUNDAMENTAIS

Nesta seção serão apresentados alguns dos conceitos mais importantes que estão presentes neste trabalho. Não se pretende cobrir toda a área do conhecimento aqui utilizado, mas fornecer um embasamento mínimo que poderá facilitar a compreensão do trabalho.

2.6.1 Série temporal

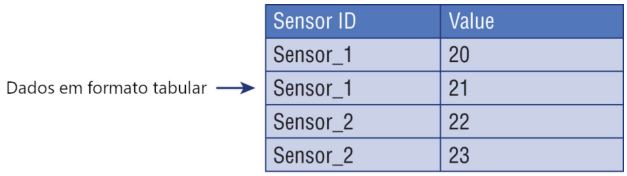
Série Temporal é um tipo de dado bastante específico, pois a informação que ela mensura ocorre ao longo de um período e registra como determinado fenômeno se altera. Apesar de sua especificidade, as Séries Temporais estão em várias áreas do conhecimento. As séries podem ser regulares, como, por exemplo, a chuva acumulada por dia nos últimos anos – algo feito neste trabalho –, total de manchas solares anuais dos últimos séculos (81), entre muitas outras aplicações. E as Séries Temporais podem ser irregulares também. Uma aplicação de série temporal irregular, ou seja, que ocorre em intervalos de tempo variados, não fixados no mesmo ritmo, seria a mensuração da ocorrência de falhas em equipamentos. (61)

Um importante aspecto da Série Temporal é que ela se estrutura de maneira hierárquica, ou seja, o evento registrado no momento anterior $(t-1)$ tem precedência sobre o momento atual (t) e assim por diante $(t+1,\ t+2...)$ até o momento final $(t+n)$. Por isso a análise de séries temporais demanda um ferramental específico e conceituação distinta de outras áreas da Ciência de Dados, pois uma série temporal pode apresentar diversas características que não se encontram em dados tabulares, tais como Sazonalidade, Ciclo, Tendência e Autocorrelação, por possuírem a dimensão extra do tempo. (58)

Considere as imagens 2.2 e 2.3 e veja como a adição da coluna com a data passa uma compreensão nova ao problema. A coluna com o valor aferido está crescendo, denotando tendência com variação de um dia para cada um dos sensores.

Para compreender o problema que se deseja estudar e ter uma visão ampla de como o(s) modelo(s) podem solucioná-lo é que se emprega o conceito de Taxonomia de Séries Temporais. (10) Esta é uma forma estruturada de analisar a Série Temporal e responder às questões pode ser valioso para o desenvolvimento e aprimoramento do trabalho. Uma

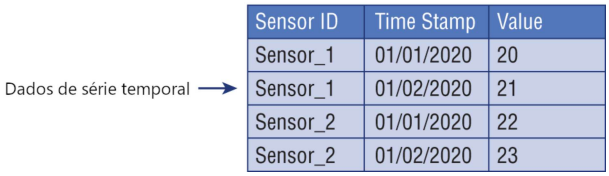
Figura 2.2 – Dados em formato tabular comumente visto em aprendizado de máquina.



Sensor ID	Value
Sensor_1	20
Sensor_1	21
Sensor_2	22
Sensor_2	23

Fonte: Lazzeri (58)

Figura 2.3 – Dados com a dimensão temporal na coluna *"Time Stamp"*.



Sensor ID	Time Stamp	Value
Sensor_1	01/01/2020	20
Sensor_1	01/02/2020	21
Sensor_2	01/01/2020	22
Sensor_2	01/02/2020	23

Fonte: Lazzeri (58))

versão mais elaborada da taxonomia é apresentada em Lazzeri (58), aqui preferiu-se por uma versão simplificada, mas que será suficiente para o que se pretende.

- *Input* x *Output*: Esta é uma etapa de avaliação do que será utilizado como entrada aos modelos e o que se desejará prever. Em geral, um problema de predição envolve ter observações do passado (*inputs*) e, a partir de tais dados observados, antever um ou mais futuros possíveis (*outputs*).

- Endógeno x Exógeno: Uma variável de entrada é considerada endógena quando existem outras variáveis no sistema que influenciam ela – incluindo ela própria – e o resultado (*output*) depende desta variável. O outro caso, quando a variável de entrada é exógena, é quando ela independe de outras variáveis do sistema – inclusive de si mesma –, apesar de o resultado depender desta variável exógena. Para facilitar a compreensão, a variável endógena para este trabalho é a vazão e as variáveis exógenas são os dados de precipitação, variáveis categóricas e os valores defasados (*lags*) da própria vazão.

- Não-estruturado x Estruturado: Série Temporal estruturada é o caso de haver, sistematicamente, padrões de dependência temporal na série, tais como tendência, sazonalidade, ciclo. E estes padrões podem acontecer juntos ou não, não existe regra. Pode haver apenas um ou dois ao mesmo tempo. Ou até mesmo nenhum deles, o que nos leva à Série Temporal não-estruturada, que é, por óbvio, quando não está presente nenhum padrão de dependência temporal entre os dados da série.

- Regressão x Classificação: Regressão é quando se deseja prever um valor, uma quantidade. A classificação é rotular os dados da série com um ou mais marcadores,

rótulos (*labels*).

- Univariado x Multivariado: Uma previsão é univariada quando uma única variável é mensurada ao longo do tempo. Por sua vez, o problema é tido como multivariado quando o trabalho aferir mais de uma variável ao longo do tempo.

- *Single-step* x *Multi-step*: O problema é do tipo *single-step* quando a previsão for apenas do próximo passo de tempo da série. E *multi-step* em caso contrário, quando se busca prever mais de um passo no futuro.

- Estático x Dinâmico: Um modelo de predição é estático quando, uma vez treinado e seus parâmetros ajustados, ele é aplicado diretamente nos novos dados que surgirem. O modelo dinâmico é retreinado e reajustado quando surgem novos dados (treino e teste) antes de cada previsão.

- Contínuo x Descontínuo: A Série Temporal é contínua quando existe uniformidade nas observações, por exemplo, observação a cada hora, minuto, dia, semana, mês e daí em diante. A descontinuidade numa Série Temporal pode ser por falta de informação – não foi coletada – ou corrupção de dados. Contudo, pode ser uma característica do problema em questão também. Tome como exemplo uma série temporal que registra oscilações elétricas anômalas em uma máquina industrial, com vias a detectar uma possível falha. As oscilações são registradas no tempo, mas podem ocorrer em uma frequência não uniforme. Ela pode acontecer agora, a próxima daqui a 5 minutos, depois disso pode levar 4 horas para uma nova oscilação anômala.

Este trabalho não se presta a exaurir toda conceituação acerca de Séries Temporais. O que foi mostrado aqui é o que de mais importante foi empregado para a compreensão do problema e como desenvolver modelos que pudessem resolvê-lo.

2.6.2 Modelos de aprendizado de máquina

As subseções a seguir descrevem os diversos modelos de regressão utilizados para previsão de vazão. Primeiramente, é apresentado brevemente o modelo de referência, seguido pelos modelos mais complexos e robustos que são alvo de estudo e análise deste trabalho.

2.6.2.1 Seasonal Naive

O *Seasonal Naive* é um método similar ao *Naive* – que meramente aplica a média histórica passada –, contudo, usa a última observação conhecida do mesmo período – por exemplo, mesmo dia do mês anterior, mesma semana do ano anterior – numa tentativa de capturar o comportamento sazonal.

Este modelo não pode ser considerado um método de previsão sofisticado, pois não captura as tendências e as variações complexas. Em vez disso, ele funciona como um modelo de referência, *baseline*, servindo como ponto de partida para avaliar o desempenho de outros modelos de previsão.(51)

Para este modelo ser executado foi preciso ajustar o hiperparâmetro de sazonalidade, que é de 365 dias.

2.6.2.2 Regressão linear - LR

O modelo de Regressão Linear, *Linear Regression*, é uma abordagem estatística simples, porém poderosa, que busca modelar o relacionamento entre uma variável dependente e uma ou mais variáveis independentes por meio de uma linha reta.

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t \tag{2.1}$$

O funcionamento da Regressão Linear envolve o cálculo de coeficientes para as variáveis independentes, que determinam o peso de cada uma destas variáveis na previsão da variável dependente. O objetivo do modelo é minimizar o somatório do erro quadrático, ou seja, a soma dos quadrados das diferenças entre os valores previstos e os valores reais.(50)

$$\sum_{t=1}^{n} \varepsilon_t^2 = \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 x_{1,t} - \beta_2 x_{2,t} - \cdots - \beta_k x_{k,t})^2. \tag{2.2}$$

- y_t valores observados
- $\beta_0, \beta_1, \cdots, \beta_k$ coeficientes para as variáveis independentes
- $x_{1,t}, x_{2,t}, \cdots, x_{k,t}$ variáveis independentes
- n número de observações

Os resultados obtidos com este modelo mostraram-se ótimos, e, na verdade, ele se destacou como um modelo-base robusto para os outros modelos mais complexos. Sua simplicidade e eficácia tornam-no uma escolha bastante sólida. Não houve nenhuma seleção de hiperparâmetro para executar este modelo, foi utilizado conforme o padrão da biblioteca.

2.6.2.3 CatBoost - CB

CatBoost (*Categorical Boosting*) tem o funcionamento baseado no princípio da construção de modelos fracos de árvores de decisão (*decision tree*) de forma sequencial, onde cada árvore sucessiva é treinada para corrigir os erros da árvore anterior. A estratégia

do *CatBoost*, no entanto, tem um nome: "*Ordered Boosting*". A construção das árvores se dá de maneira sequencial, porém não se usa todos os dados disponíveis para isso. Os dados de treinamento são ordenados de maneira aleatória e apenas partições destes dados são utilizados no processo. Por trabalhar sempre com uma amostra dos dados de treinamento, e a apresentação aleatória destes dados ao modelo, o CatBoost tem resiliência ao *overfitting*, mas parâmetros que realizam ajustes nas árvores de decisão também estão presentes e o utilizador tem controle sobre eles.(96)(28)(75)

Os hiperparâmetros para execução do modelo foram escolhidos para que o modelo executasse sem escrever ('*allow_writing_files*') cada passo de treinamento em disco durante a execução, para salvar espaço em disco e evitar *overhead* dessa escrita desnecessária. As variáveis categóricas foram passadas através do hiperparâmetro '*cat_features*' e '*thread_count*' para executar em paralelo, para acelerar o treinamento do modelo. O hiperparâmetro 'verbose' é para não escrever cada passo de treinamento na saída de texto do sistema e '*random_seed*' é para garantir reprodutibilidade dos resultados em todas as execuções do modelo. Apenas estes hiperparâmetros foram selecionados, tudo mais ficou segundo o padrão da biblioteca.

Tabela 2.3 – Hiperparâmetros para o modelo *CatBoost*.

Hiperparâmetro	**Valor**
random_seed	1984
cat_features	['estacao', 'month']
verbose	False
allow_writing_files	False
thread_count	8

Fonte: O autor

2.6.2.4 Floresta Aleatória - RF

Floresta Aleatória (*Random Forest*) é um modelo que também é baseado em árvores de decisão e é amplamente utilizado em tarefas de regressão e classificação. O modelo combina diversas árvores de decisão para aprimorar a precisão e a estabilidade dos resultados.

O modelo cria diversas árvores de decisão independentemente, e cada destas árvores é treinada com partições aleatórias dos dados de treinamento, garantindo que as árvores sejam distintas entre si. Na seleção dos dados pode haver reposição (*bootstrapping*), ou seja, os mesmos dados podem ser utilizados para treinar árvores diferentes ou até mesmo dados podem nunca ser utilizados em nenhuma árvore. Isso é para garantir diversidade entre as árvores. Durante a criação de cada nó da árvore, um subconjunto aleatório dos dados de treinamento é utilizado e isso vai até não ser mais possível criar novas divisões

(*Depth-wise growth*). Ao fim, quando cada árvore foi treinada, a agregação dos resultados se dá calculando a média entre todas elas (*bagging*) (83)(85).

Seguindo a mesma intenção de antes, '*random_state*' está aqui para garantir reprodutibilidade dos resultados entre diversas execuções. Por fim, '*verbose*' serve para não ficar escrevendo na saída de texto do sistema e '*n_jobs*' para executar com paralelismo o treinamento do modelo. O modelo *Random Forest* não trata variáveis categóricas diretamente, por isso não tem nenhum hiperparâmetro a respeito disso.

Tabela 2.4 – Hiperparâmetros para o modelo *Random Forest*

Hiperparâmetro	**Valor**
random_state	1984
verbose	0
n_jobs	8

Fonte: O autor

A escolha do modelo certo para um problema de previsão desta natureza depende das características dos dados e dos objetivos do estudo. O *Seasonal Naive*, embora simples, é um importante ponto de referência inicial. Contudo, foi o modelo Regressão Linear que, de fato, serviu como uma *baseline* confiável devido à sua eficácia e simplicidade. Já os modelos *CatBoost* e *Random Forest* foram opções mais avançadas, capazes de lidar com a complexidade dos dados e suas relações não-lineares.

3 PROCEDIMENTOS METODOLÓGICOS

3.1 TAXONOMIA DA SÉRIE TEMPORAL

Apresentada a taxonomia de série temporal anteriormente, agora é aplicar o conceito para o problema proposto neste trabalho.

Tabela 3.1 – Taxonomia do trabalho.

Procedimento/Classificação	**Justificativa**
Inputs x *Outputs*	Os dados de entrada (*inputs*) serão vazões e precipitações diárias registradas no período de 01/01/2013 a 31/12/2022 (para o rio Grande foi de setembro de 2020 a dezembro de 2022), variáveis categóricas extraídas da coluna "data" e valores de precipitações acumuladas. Os valores de saída (*outputs*) dos modelos serão previsões de vazão diária para 365 dias, neste caso, para o ano de 2023 inteiro.
Endógeno x Exógeno	Variável endógena é o dado de vazão para cada rio estudado e exógenas são os dados de precipitação, as variáveis de precipitação acumuladas, as *lags* de vazão e as variáveis categóricas.
Não-estruturado x Estruturado	As séries temporais de vazão e precipitação apresentam sazonalidade e tendência, ou seja, são séries estruturadas.
Regressão x Classificação	Previsão de série temporal pode ser convertido para um problema do tipo regressão, o que será realizado neste trabalho.
Univariado x Multivariado	Este problema é do tipo univariado com uso de variáveis exógenas. Se assemelha aos problemas multivariados, mas apenas uma variável será o alvo neste caso, que é a vazão.
Single-step x *Multi-step*	A previsão realizada será do tipo *single-step*, granularidade de um dia à frente.
Estático x Dinâmico	O trabalho é do tipo dinâmico. Antes de cada previsão (*predict*) será executado o passo de ajuste dos modelos aos dados (*fit*).
Contínuo x Descontínuo	As séries temporais utilizadas são descontínuas, seja por falha de sensor, de medição manual ou manutenção de equipamento. Para o preenchimento das lacunas, procedimentos específicos serão empregados.

Fonte: O autor

3.2 DESCRIÇÃO DAS ÁREAS DE ESTUDO

Os rios estudados foram os rios Jequitinhonha, Doce, Grande e São Francisco, e nestes rios, buscou-se por dados e informações que estivessem estritamente dentro do estado de Minas Gerais. Onde esta regra foi violada, destacou-se e explicou-se a motivação.

O rio Jequitinhonha, com sua bacia hidrográfica abrangendo cerca de 65 mil quilômetros quadrados, é uma peça fundamental na paisagem e na vida socioeconômica de Minas Gerais. Seu curso percorre 82 municípios mineiros, abrigando uma população de aproximadamente 939 mil habitantes que dependem diretamente de suas águas (19) (21) (22) (23).

A Mata Atlântica é o bioma predominante e a atividade econômica principal na região é a agropecuária, ocupando uma área considerável de $2,6$ milhões de hectares, em contraste com os $3,8$ milhões de hectares de floresta. Essa dinâmica evidencia a pressão exercida sobre o ecossistema, demandando um olhar atento para a gestão sustentável dos recursos naturais (9).

Com informações do Comitê da Bacia Hidrográfica do Rio Doce (14), o rio Doce, com sua nascente na Serra da Mantiqueira, em Minas Gerais, percorre 853 quilômetros até desaguar no Oceano Atlântico, no Espírito Santo, delineando uma bacia hidrográfica muito importante para o Sudeste brasileiro. Com uma área de drenagem de 86 mil quilômetros quadrados, abriga 200 municípios mineiros, onde residem cerca de $3,5$ milhões de pessoas, e desempenha um papel crucial na dinâmica socioeconômica e ambiental da região.

Continuando sobre a bacia do rio Doce, há a predominância da Mata Atlântica, com uma pequena fração de Cerrado, conferindo à bacia uma vasta biodiversidade, porém, ao mesmo tempo, a bacia encontra-se em posição de vulnerabilidade frente às pressões antrópicas. As principais atividades econômicas são a agropecuária, a agroindústria e o setor de mineração/siderurgia (Quadrilátero Ferrífero) e impulsionam a economia regional. A cobertura do solo, com $2,4$ milhões de hectares de floresta e $5,5$ milhões de hectares dedicados à agropecuária, reflete essa dualidade entre desenvolvimento e preservação (14) (9).

O rio Grande é um dos cursos d'água mais importantes do Brasil, escoando por Minas Gerais e São Paulo até se unir ao rio Paranaíba e formar o rio Paraná. Com uma área de drenagem de 143 mil quilômetros quadrados, sua bacia abrange 393 municípios, impactando a vida de cerca de 9 milhões de pessoas. Nascido na Serra da Mantiqueira, o rio Grande é um rio de planalto, percorrendo $1286km$ até seu encontro com o rio Paranaíba. O Cerrado é o bioma predominante em sua bacia, com trechos que se limitam à Mata Atlântica, criando uma rica diversidade de ecossistemas. (15)

A agroindústria é a principal atividade econômica na região, impulsionada pela fertilidade do solo e pela disponibilidade de água. No entanto, o rio Grande também

é fundamental para a geração de energia elétrica, abrigando 13 barragens, incluindo a importante usina de Furnas. Essa combinação de agropecuária e geração de energia molda a paisagem da bacia, com 11,1 milhões de hectares dedicados à agricultura e 2,1 milhões de hectares cobertos por florestas (15) (9).

Por fim, mas certamente, não menos importante, rio São Francisco. Carinhosamente apelidado de "Velho Chico", conecta regiões, culturas e economias ao longo de seus quase 3 mil quilômetros de extensão. Com uma bacia hidrográfica que abrange 639 mil quilômetros quadrados, a bacia do São Francisco corresponde a 8% do território nacional, impactando diretamente a vida de cerca de 15 milhões de pessoas. Sua grandiosidade é tamanha que sua bacia abrange seis estados brasileiros – Minas Gerais, Goiás, Bahia, Pernambuco, Alagoas e Sergipe – e devido à tanta complexidade, a bacia é dividida em quatro regiões fisiográficas distintas: Alto, Médio, Submédio e Baixo São Francisco, cada qual com particularidades geográficas, sociais e econômicas (16).

A diversidade de biomas que o Velho Chico atravessa é outro ponto marcante: do Cerrado à Caatinga, passando por fragmentos da Mata Atlântica, essa variedade se reflete também na cobertura do solo de sua bacia, com predominância de áreas destinadas à agropecuária (12,4 milhões de hectares) e florestas (15,3 milhões de hectares), indicando a importância tanto da produção agrícola quanto da preservação ambiental. Suas águas impulsionam atividades como a siderurgia, mineração, indústria química e têxtil, além de pesca e agropecuária (16) (9).

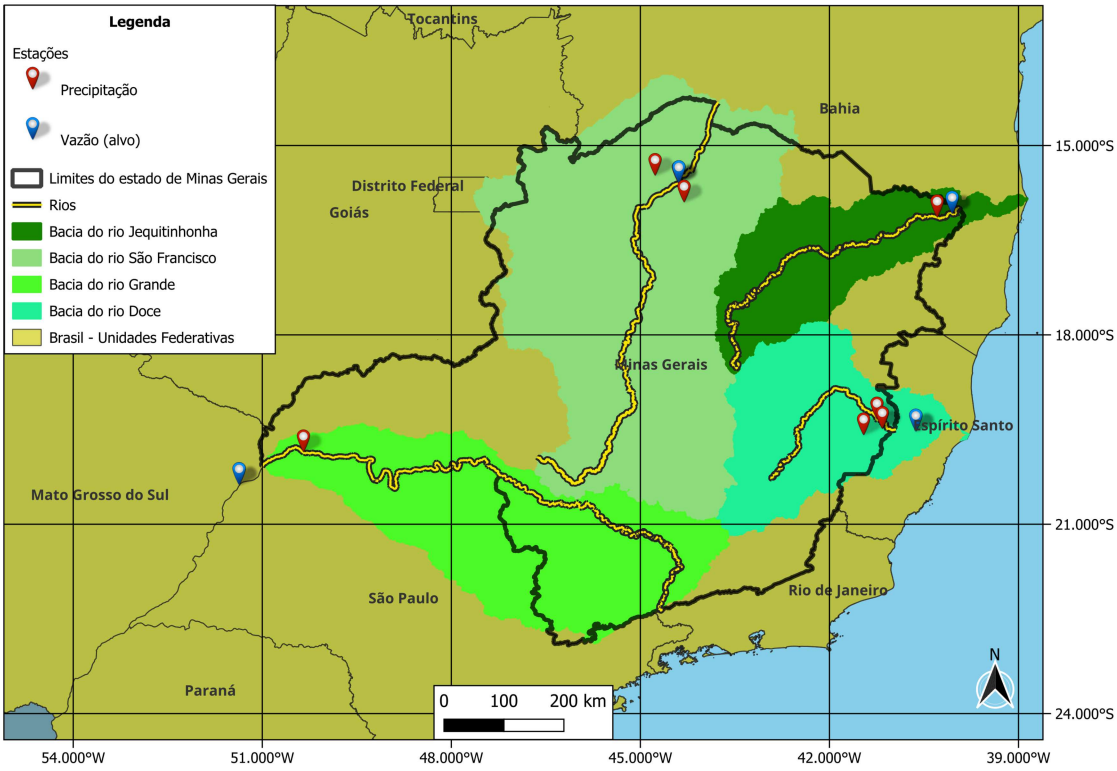
A figura 3.1 apresenta uma visão ampla das bacias estudadas, seus respectivos rios principais e dimensões comparativamente ao estado de Minas Gerais. As bacias quase cobrem o estado inteiro.

Todos os mapas criados para esta dissertação fizeram amplo uso dos repositórios públicos de arquivos *shapefile* fornecidos por IBGE (20) e ANA (26). A ferramenta QGis (77) foi empregada para agregar os arquivos *shapefile* e gerar os mapas, tarefa esta executada pelo autor.

3.3 DADOS UTILIZADOS

A coleta dos dados foi realizada com uma combinação de uso da biblioteca HydroBR (18), funções próprias de extração de dados e consulta ao sistema on-line da Rhama Analysis. A biblioteca permitiu a listagem de todas as estações hidrométricas disponíveis, de vazão e precipitação, telemétrica ou convencional. Após a identificação e seleção das estações de interesse, cujos códigos estavam disponíveis na base de dados da ANA, desenvolveu-se um conjunto de funções para automatizar o processo de extração. Essas funções permitiram o *download* dos dados referentes ao período especificado diretamente do *webservice* fornecido pela ANA.

Figura 3.1 – As bacias estudadas e seus respectivos rios principais.



Fonte: O autor

A variável correspondente à data foi formatada como "*datetime*", enquanto os dados de precipitação e vazão, bem como as variáveis acumuladas, foram representados como valores de ponto flutuante ("*float*"). Finalmente, as variáveis categóricas utilizadas foram representadas por números inteiros ("*int*"). Embora a frequência diária tenha sido adotada, é importante destacar que nem todas as séries temporais estavam originalmente nesse formato. Foi necessário lidar com quebra na continuidade das datas e com dados ausentes.

Os dados de precipitação e vazão obtidos do site da ANA já estavam ajustados nas escalas padrão utilizadas em estudos hidrológicos. A precipitação foi fornecida em milímetros por dia (mm/dia), refletindo a quantidade de chuva sobre uma unidade de área (m^2) em um período de 24 horas e as vazões, por sua vez, foram disponibilizadas em metros cúbicos por segundo (m^3/s), indicando o volume de água que passa por uma seção transversal do rio a cada segundo. Em algumas estações foram observados valores extremamente elevados para determinados dias, tanto nas séries de precipitação quanto nas de vazão, os quais podem ser considerados *outliers*. Em relação aos dados de vazão, verificou-se a ocorrência de valores nulos – vazão igual a 0 –, o que indicaria a interrupção

completa do fluxo do rio. Esse fenômeno, no entanto, não faz sentido, considerando que não há registro de eventos de seca tão severos nos rios analisados. Apesar destas anomalias, os dados não foram descartados, pois tanto os registros de vazão quanto os de precipitação utilizados nesta pesquisa foram considerados consistidos pela ANA, ou seja, foram validados pela agência. O presente trabalho não questionou a veracidade dos dados; eles foram utilizados conforme disponibilizados pela agência e pelo sistema da Rhama Analysis.

Em cada rio analisado, a estação alvo, com a vazão que se pretendia prever, foram destacadas em itálico para ficar claro ao leitor como identificá-las.

A distinção entre estação convencional e telemétrica deve-se a esta ter informações a cada quinze minutos, a cada trinta minutos ou ser do tipo horária. Onde ocorreu de ter informações tão granuladas assim, para a precipitação foi feito o somatório para um dia e a vazão foi a média de um dia.

Cabe destacar algumas observações sobre as estações do rio Grande. Durante o período pesquisado, apenas foram encontrados dados de precipitação e vazão em estações localizadas no estado de São Paulo. As estações utilizadas para o rio Grande, as mais próximas da foz do rio e próximas à divisa com o estado de Minas Gerais, são aquelas listadas na tabela.

Uma situação semelhante ocorreu com o rio Doce. Não foram encontradas estações com dados disponíveis na foz do rio Doce, localizada no estado de Minas Gerais. Portanto, foi necessário utilizar a estação 56994500, situada no estado do Espírito Santo.

Estas são as únicas observações relevantes sobre as estações utilizadas.

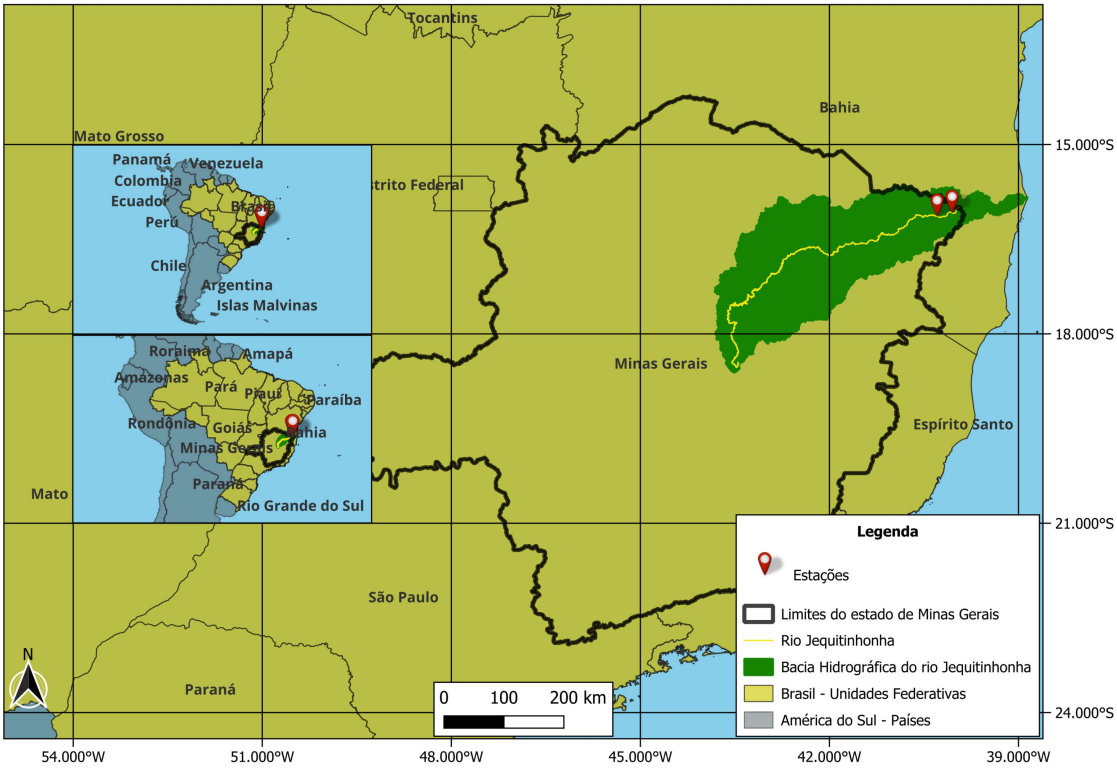
Para facilitar a visualização, as estações de vazão e precipitação utilizadas no trabalho são apresentadas abaixo – tabelas 3.2, 3.3, 3.4 e 3.5 e figuras 3.2, 3.3, 3.4 e 3.5.

Tabela 3.2 – Estações usadas no rio Jequitinhonha.

Telemétricas				
Pluviométricas/Fluviométricas				
Código	**Nome**	**Município**	**Latitude**	**Longitude**
54790000	*UHE Itapebi montante 1*	*Salto da Divisa*	*-16,08*	*-40,0521*
01640000	Jacinto	Jacinto	-16,1386	-40,2903

Fonte: O autor

Figura 3.2 – Rio Jequitinhonha e as estações utilizadas.



Fonte: O autor

Tabela 3.3 – Estações usadas no rio Doce.

Convencionais				
Pluviométricas				
Código	**Nome**	**Município**	**Latitude**	**Longitude**
01941010	São Sebastião da Encruzilhada	Aimorés	-19,4925	-41,1617
01941004	Resplendor - jusante	Resplendor	-19,3431	-41,2461
01941006	Assarai - montante	Pocrane	-19,5947	-41,4581
Telemétricas				
Pluviométricas/Fluviométricas				
Código	**Nome**	**Município**	**Latitude**	**Longitude**
56990005	UHE Aimorés rio Manhuaçu	Aimorés	-19,4917	-41,1614
56994500	*Colatina ponte*	*Colatina*	*-19,5333*	*-40,6297*

Fonte: O autor

Figura 3.3 – Rio Doce e as estações utilizadas.

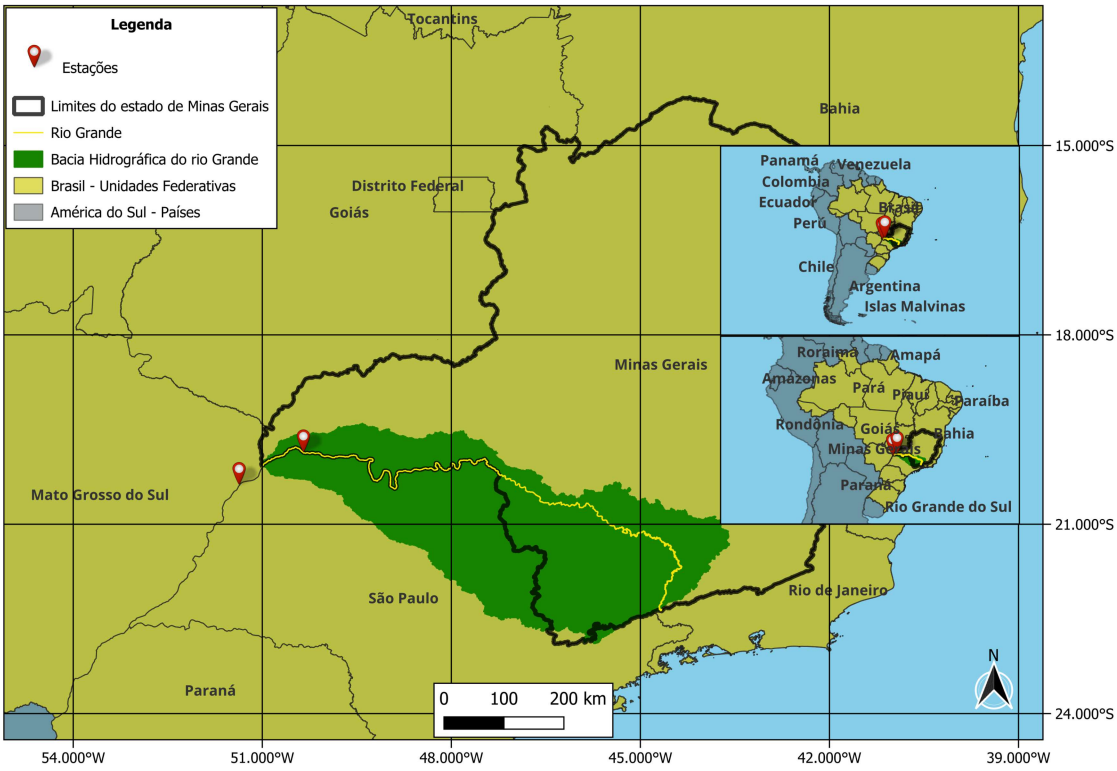
Fonte: O autor

Tabela 3.4 – Estações usadas no rio Grande.

Telemétricas				
Fluviométricas				
Código	**Nome**	**Município**	**Latitude**	**Longitude**
62020080	*UHE Ilha Solteira barramento*	*Ilha Solteira*	*-20,3797*	*-51,3686*
Pluviométricas/Fluviométricas				
Código	**Nome**	**Município**	**Latitude**	**Longitude**
61998080	UHE Água Vermelha barramento	Ouroeste	-19,8628	-50,3475

Fonte: O autor

Figura 3.4 – Rio Grande e as estações utilizadas.



Fonte: O autor

Tabela 3.5 – Estações usadas no rio São Francisco.

Convencionais				
Fluviométricas				
Código	**Nome**	**Município**	**Latitude**	**Longitude**
44290002	*Pedras de Maria da Cruz*	*Pedras de Maria da Cruz*	*-15,6011*	*-44,3967*
Pluviométricas/Fluviométricas				
Código	**Nome**	**Município**	**Latitude**	**Longitude**
01544017	Pedras de Maria da Cruz	Januária	-15,5978	-44,3903
01544032	Usina do Pandeiros montante	Januária	-15,4831	-44,7672
01544036	Lontra	Lontra	-15,9056	-44,3072

Fonte: O autor

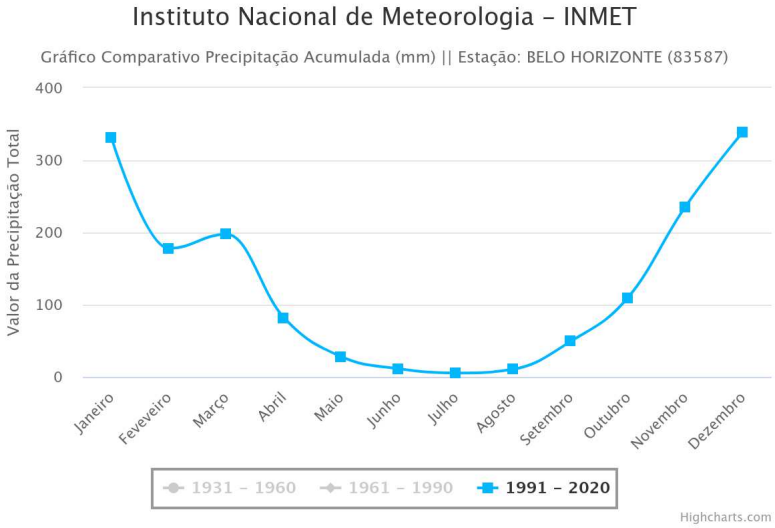
Figura 3.5 – Rio São Francisco e as estações utilizadas.

Fonte: O autor

3.4 PRÉ-PROCESSAMENTO DOS DADOS

A sazonalidade da precipitação é um fenômeno bem estabelecido na análise hidrológica para as bacias de Minas Gerais. O aumento da precipitação começa na primavera, em setembro, e atinge seus picos nos meses de dezembro e janeiro, durante o verão. Consequentemente, as vazões dos rios aumentam. Com a chegada do outono e, posteriormente, do inverno, os índices pluviométricos diminuem, assim como as vazões dos rios. Na figura 3.6 é mostrado o regime médio da precipitação no estado de Minas Gerais entre os anos de 1991 e 2020, tomando como base a estação de Belo Horizonte do INMET (53).

Figura 3.6 – Regime médio da precipitação no intervalo 1991 - 2020.



Fonte: INMET (53), estação de Belo Horizonte

Considerando esse fenômeno, o preenchimento dos dados faltantes foi realizado replicando o padrão sazonal. Para preencher um dia faltante em julho, por exemplo, foi utilizado o valor correspondente ao mesmo dia nos anos anteriores. Para evitar a repetição exata do ano anterior, utilizou-se a média dos últimos três anos. As funções desenvolvidas para essa finalidade são personalizáveis, permitindo que se opte por repetir exatamente o ano anterior ou considerar mais de três anos, dependendo das necessidades do estudo.

Note que a estratégia de realizar a média, para o dia, dos anos anteriores nem sempre preenchia exatamente as lacunas. Quando havia muitos dados faltantes no início da série, por exemplo, isso causava problema e a inserção de dados falhava. O que é o comportamento normal.

Foi então que se realizou uma nova contagem dos dados que ainda permaneciam faltantes. Para esses casos nulos foi aplicada a imputação de dados utilizando o modelo kNN, visando garantir uma melhor dispersão dos valores imputados. O modelo kNN operou calculando a distância euclidiana dos pontos nulos utilizando os sete vizinhos mais próximos, atribuindo maior peso aos vizinhos mais próximos no cálculo. Vizinhos, neste contexto,

cabe destacar, são os valores de vazão observados mais próximos ao dia faltante. Continuando. Esse método de imputação visou preservar a tendência local e o comportamento da série temporal dentro da semana em que o dado faltante estava. Após esta nova fase de imputação dos dados as séries ficaram completamente preenchidas. O motivo de utilizar especificamente este modelo foi inspirado no trabalho desenvolvido por Portuguez-Maurtua et al. (74).

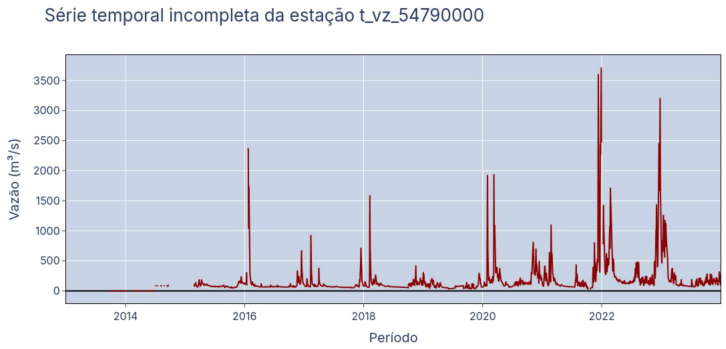
É muito importante o destaque para esta fase de preenchimento de dados faltantes, e os desafios que isso apresentou ao trabalho, porque a escassez de informação foi um problema. Quando o período faltante era curto, o comportamento da série temporal preservou coerentemente os padrões sazonais e de tendência. Contudo, mais especificamente para o rio Grande, isso tudo ainda não foi suficiente. A série temporal de vazão não preservou o comportamento sazonal esperado, ficando com muitos ruídos, a ponto de ser preciso remover grande trecho dos dados, ficando apenas com dados de 2020 até o ano de 2023.

Neste momento cabe explicar uma nomenclatura utilizada no trabalho para rapidamente identificar o tipo de estação, se convencional ou telemétrica, de que dado ela trata, chuva ou vazão, e o código da estação. Tomemos dois exemplos que serão vistos nesta seção. Esta é a estação 'c_cv_01941010', utilizada na análise do rio Doce. A letra 'c' designa 'convencional' e as letras 'cv' significam 'chuva', consequentemente, a sequência numérica é o código da estação em questão. A mesma analogia serve para as estações telemétricas. O nome 't_vz_54790000' significa 'estação telemétrica (t) de vazão (vz), código 54790000'.

3.4.1 Rio Jequitinhonha

A estação de vazão utilizada no rio Jequitinhonha apresentou uma quantidade significativa de dados faltantes, especialmente no início da série temporal. Observa-se na figura 3.7 uma clara sazonalidade na série, com picos de vazão ocorrendo predominantemente no final e início de cada ano. As figuras 3.7, 3.12, 3.8, 3.9, 3.10 e 3.11 apresentam gráficos comparativos entre a série original, sem dados imputados, incompleta, e a série após a imputação de dados, completa. Essa comparação, contrapondo a série fornecida pela ANA e os resultados após a inserção de dados, permitirá uma visualização mais clara do impacto das técnicas de preenchimento. Esta estação, identificada como t_vz_54790000, apresentou 532 dias de dados nulos, o que corresponde a aproximadamente $13,24\%$ do total. Essa é a estação alvo para a previsão das vazões.

Figura 3.7 – Série temporal incompleta da estação t_vz_54790000.



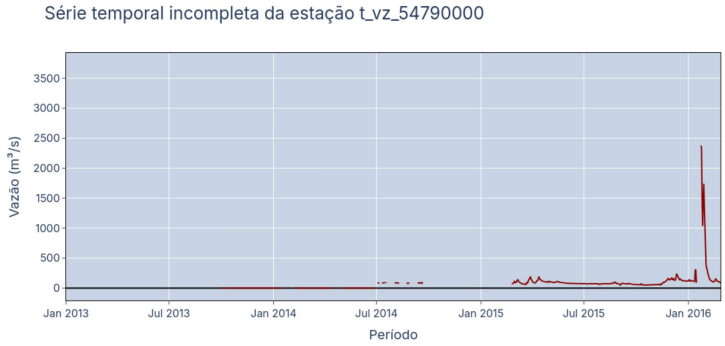
Fonte: O autor

A seguir, destaca-se o trecho da série com a maior quantidade de dados faltantes – figura 3.8 –, que abrange o período de janeiro de 2013 a janeiro de 2016. Em sequência, é apresentada a série após a imputação dos dados – figura 3.9. Notavelmente, essa seção não apresentou resultados ideais, uma vez que a imputação atribuiu vazões zero em vários dias, o que não é realista, pois isso indicaria a secagem completa do rio, o que é improvável. No entanto, esses valores zero não impactaram significativamente os resultados da análise, já que se referem a um período distante do foco principal deste estudo. Uma alternativa seria excluir todo o trecho anterior ao ano de 2016, mas optou-se por manter a uniformidade nos critérios de aproveitamento dos dados ao longo do trabalho, dado que outros rios também foram analisados, e buscava-se assegurar consistência entre os resultados.

Observe também o trecho de dados faltantes mais próximo ao final dos anos analisados, em 2021 e 2022, na figura 3.10. Esta porção da série ficou boa por haver informação prévia suficiente e a inserção de dados capturou a média da sazonalidade passada – figura 3.11.

Por fim, uma visão ampla de como ficou a série temporal após os procedimentos de imputar os dados – figura 3.12.

Figura 3.8 – Detalhe da série temporal da estação t_vz_54790000, ainda sem dados imputados, de 2013 a 2016.



Fonte: O autor

Figura 3.9 – Detalhe da série temporal da estação t_vz_54790000, com dados imputados, de 2013 a 2016.

Fonte: O autor

Figura 3.10 – Série temporal incompleta da estação t_vz_54790000 no detalhe entre 2021 e 2022.

Fonte: O autor

A mesma análise foi realizada para as estações de chuva. Na estação t_cv_54790000 (figura 3.13) faltavam 273 dias de dados (6, 79%), no início do ano de 2013, em sua vasta maioria. Já a estação t_cv_01640000 estava totalmente preenchida, sem valores nulos –

Figura 3.11 – Série temporal completa da estação t_vz_54790000 no detalhe entre 2021 e 2022.

Fonte: O autor

Figura 3.12 – Série temporal completa da estação t_vz_54790000.

Fonte: O autor

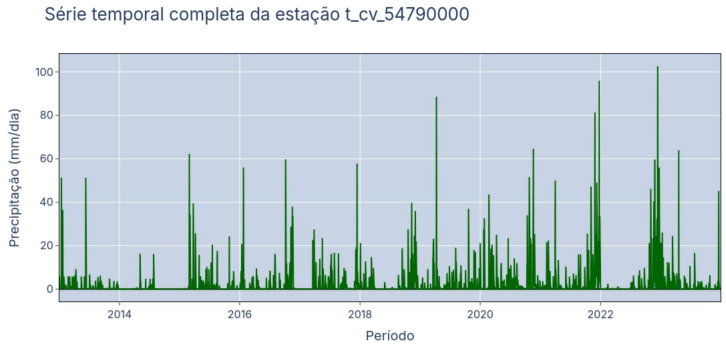
figura 3.15.

Figura 3.13 – Série temporal incompleta da estação t_cv_54790000.

Fonte: O autor

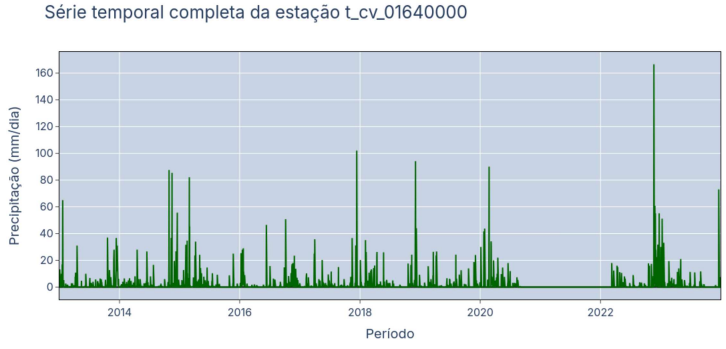
As séries de chuva completas ficaram desta forma (figuras 3.14 e 3.15)

Figura 3.14 – Série temporal completa da estação t_cv_54790000.



Fonte: O autor

Figura 3.15 – Série temporal completa da estação t_cv_01640000.

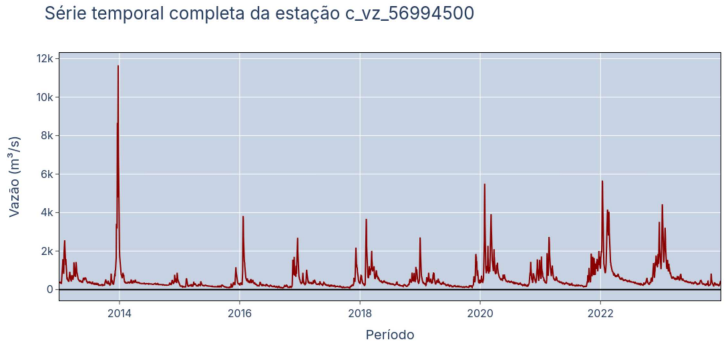


Fonte: O autor

3.4.2 Rio Doce

A estação alvo para o rio Doce é a estação c_vz_56994500. Sua série temporal foi a que apresentou melhor qualidade no que diz respeito à frequência de medições realizadas. Faltavam apenas 3 dias dos 4017 dias do período inteiro. Apenas o preenchimento sazonal bastou para completar a série e não foi preciso mais que isso. Cabe destacar a sazonalidade da série. Ficou bastante evidente este comportamento – figura 3.16.

Figura 3.16 – Série temporal completa da estação c_vz_56994500.



Fonte: O autor

Se para os dados de vazão no rio Doce a série foi, digamos, mais comportada, o mesmo não se pode dizer exatamente das estações de chuva. Ao menos, não para duas delas. Estas estações tiveram os dados desconsiderados e foram removidos das análises. Primeiro foi a estação t_cv_56990850 que possuía valores discrepantes demais para serem considerados. Valores da ordem de 7000 mm/dia, 8500 mm/dia. Além deste problema, havia ainda 3134 dias com dados nulos, o que representava 78% do total – figura 3.17.

A outra estação removida foi a t_cv_56994500. Conforme pode ser observado na figura 3.18, nela havia um longo hiato de dados zerados, voltando à normalidade apenas mais recentemente. Como as informações de precipitação que deveria haver para a estação no período do hiato podem ser retiradas de outras estações usadas na modelagem, optou-se por remover esta estação completamente do trabalho.

As estações que, enfim, foram empregadas na modelagem são as que estão nas figuras 3.19, 3.20, 3.21 e 3.22.

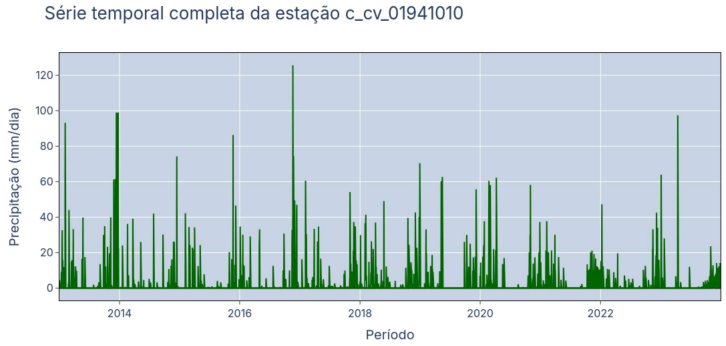
Figura 3.17 – Série temporal da estação t_cv_56990850 - não utilizada.

Fonte: O autor

Figura 3.18 – Série temporal da estação t_cv_56994500 - não utilizada.

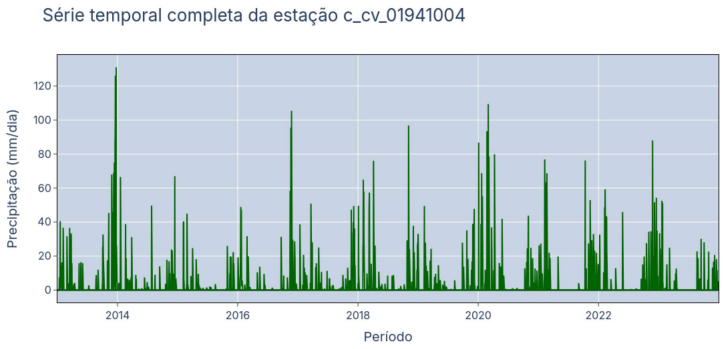
Fonte: O autor

Figura 3.19 – Série temporal completa da estação c_cv_01941010.



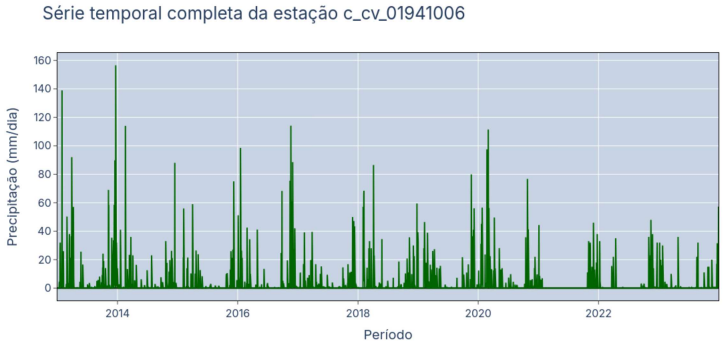
Fonte: O autor

Figura 3.20 – Série temporal completa da estação c_cv_01941004.



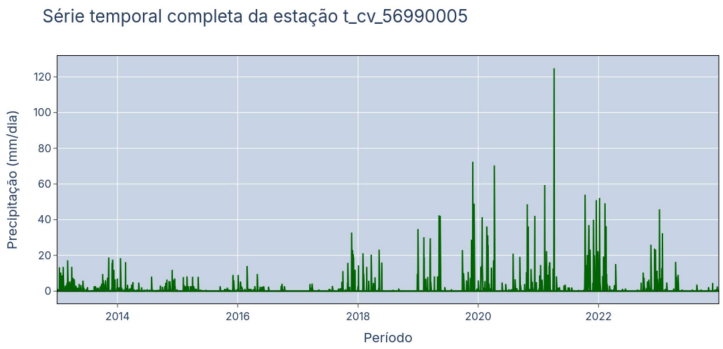
Fonte: O autor

Figura 3.21 – Série temporal completa da estação c_cv_01941006.



Fonte: O autor

Figura 3.22 – Série temporal completa da estação t_cv_56990005.



Fonte: O autor

3.4.3 Rio Grande

O rio Grande apresentou desafios significativos ao longo de todo o desenvolvimento deste trabalho. A dificuldade inicial surgiu na ausência de dados disponíveis em estações dentro do estado de Minas Gerais para o período de análise estipulado. Foi necessário buscar uma estação o mais próxima possível da divisa com Minas Gerais, localizada no estado de São Paulo, especificamente no município de Ilha Solteira. Entretanto, os desafios não se limitaram a essa questão geográfica.

A série temporal de vazão da estação selecionada, denominada t_vz_62020080, estava incompleta e não abrangia todo o período de 11 anos estipulado para a análise. (figura 3.23) Os dados disponíveis mais antigos datavam de setembro de 2020. Desta forma, para evitar problemas com a análise deste trabalho, optou-se por utilizar apenas os dados a partir de 2020 mesmo, sem realizar preenchimento anterior à data de setembro de 2020. É importante ter isso em mente em todo trabalho desenvolvido para o rio Grande.

Figura 3.23 – Série temporal incompleta da estação t_vz_62020080.

Fonte: O autor

Após o corte a série ficou com 1217 registros totais – figura 3.24.

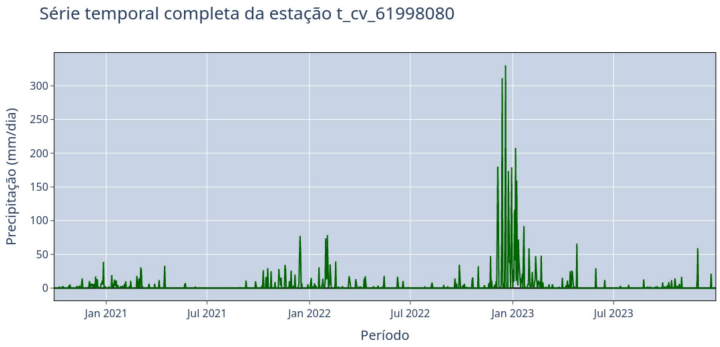
Figura 3.24 – Série temporal completa (cortada) da estação t_vz_62020080.

Fonte: O autor

A estação de precipitação utilizada, a única neste caso, foi a estação t_cv_61998080, pois foi a única que apresentou dados válidos. Curiosamente, outra estação de precipitação disponível também apresentou dados para o período analisado, mas a base de dados consistia totalmente em valores 0. Por essa razão, a estação t_cv_62020080 foi completamente excluída do estudo.

Em relação à estação t_cv_61998080, houve necessidade de preencher apenas um número reduzido de dados ausentes, totalizando 169 registros, o que correspondia a $5,14\%$ do total. Trata-se de uma série com uma quantidade expressiva de dados, que efetivamente pôde contribuir de maneira significativa para as análises realizadas. Seguindo o disposto para a série de vazões, os dados de precipitação foram removidos no período pré-setembro de 2020. A figura 3.25 já consta o período ajustado.

Figura 3.25 – Série temporal completa (cortada) da estação t_cv_61998080.



Fonte: O autor

3.4.4 Rio São Francisco

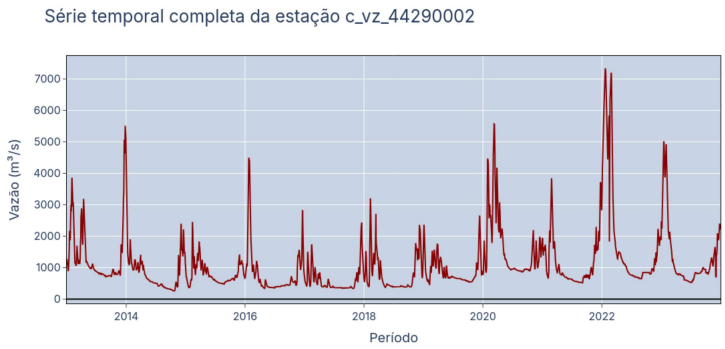
Por fim, foi realizado o procedimento de preenchimento dos dados nulos para o rio São Francisco. A estação alvo c_vz_44290002 apresentou uma série bastante completa ao longo do período de análise, com apenas 120 dias nulos em um total de 4017 dias. O trecho com dados faltantes pode ser observado em detalhe na figura 3.26.

Para esta estação, o preenchimento sazonal foi suficiente para suprir as lacunas existentes, não sendo necessário aplicar procedimentos adicionais de imputação de dados – figura 3.27.

Figura 3.26 – Detalhe do trecho com dados nulos da estação c_vz_44290002.

Fonte: O autor

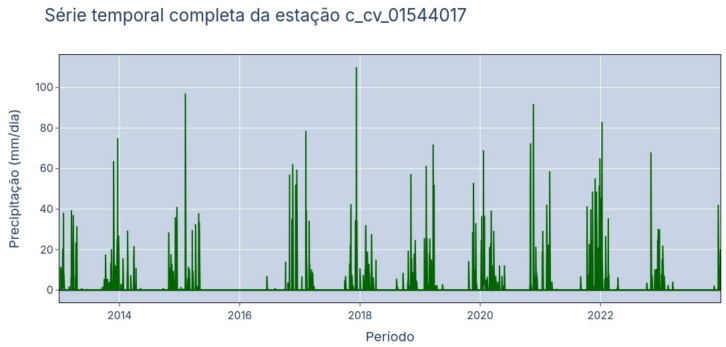
Figura 3.27 – Série temporal completa da estação c_vz_44290002.



Fonte: O autor

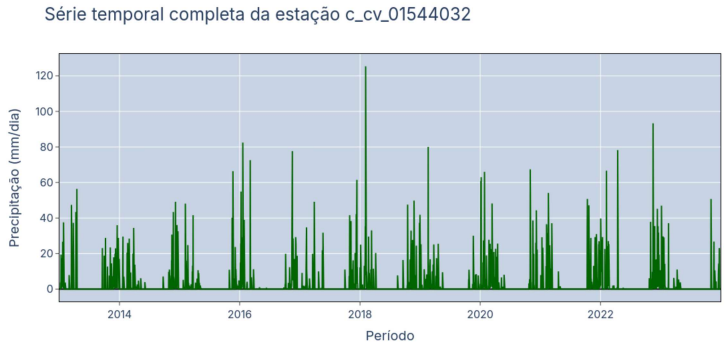
No que se refere às estações de precipitação selecionadas para a análise no rio São Francisco, não foi necessário realizar nenhuma inserção de dados, uma vez que todas as séries estavam completas, abrangendo a totalidade dos 4017 dias de registro. As séries temporais correspondentes podem ser visualizadas nos gráficos apresentados a seguir – figuras 3.28, 3.29, 3.30.

Figura 3.28 – Série temporal completa da estação c_cv_01544017.



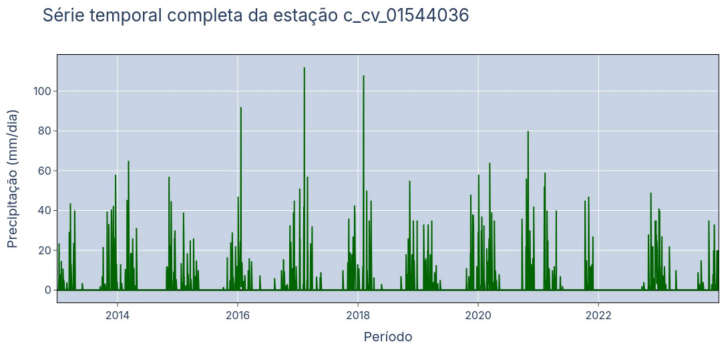
Fonte: O autor

Figura 3.29 – Série temporal completa da estação c_cv_01544032.



Fonte: O autor

Figura 3.30 – Série temporal completa da estação c_cv_01544036.



Fonte: O autor

3.5 VARIÁVEIS UTILIZADAS

As variáveis do trabalho, exceto as categóricas, obviamente, são todas contínuas e de ponto flutuante como dito anteriormente. Todas as séries temporais foram ajustadas para estarem contínuas dentro do período trabalhado para o que se obteve de dado em cada rio.

As *lags*, ou seja, os valores defasados de vazão da estação alvo foram utilizadas como variáveis preditoras. Utilizou-se 7 dias passados, o valor de uma semana passada para prever o comportamento de um dia à frente.

Para que se tenha uma noção melhor, abaixo seguem alguns dados estatísticos relevantes que informam sobre os dados de vazão e precipitação utilizados. Conste-se que as unidades de precipitação estão em mm/dia e vazão em m^3/s.

Tabela 3.6 – Variáveis utilizadas - rio Jequitinhonha.

Variável	**Amostras**	**Média**	**Mediana**	**Desvio-padrão**	**Mín.**	**Máx.**
t_cv_01640000	4017	1,58	0,00	6,85	0,00	166,60
t_cv_54790000	4017	1,68	0,00	6,31	0,00	102,60
t_vz_54790000 (y)	4017	160,95	95,95	267,90	0,00	3716,65

Fonte: O autor

Tabela 3.7 – Variáveis utilizadas - rio Doce.

Variável	**Amostras**	**Média**	**Mediana**	**Desvio-padrão**	**Mín.**	**Máx.**
c_cv_01941010	4017	2,21	0,00	8,41	0,00	125,60
c_cv_01941004	4017	2,59	0,00	9,83	0,00	131,00
c_cv_01941006	4017	2,44	0,00	9,76	0,00	156,60
t_cv_56990005	4017	1,15	0,00	5,19	0,00	124,80
c_vz_56994500 (y)	4017	542,05	341,26	656,99	75,15	11655,20

Fonte: O autor

Tabela 3.8 – Variáveis utilizadas - rio Grande.

Variável	**Amostras**	**Média**	**Mediana**	**Desvio-padrão**	**Mín.**	**Máx.**
t_cv_61998080	3289	3,13	0,00	13,92	0,00	330,40
t_vz_62020080 (y)	3289	3405,27	3170,08	873,88	1603,58	11939,49

Fonte: O autor

É possível identificar algumas questões importantes sobre a massa de dados a partir destas tabelas. Observe que para o rio Jequitinhonha, tabela 3.6, a vazão mínima foi $0,00m^3/s$, o que denotaria que o rio passou por um período de seca. Porém, não foi encontrado, seja em artigos científicos sobre o rio, quanto em matérias de jornais, que o rio Jequitinhonha tenha passado por isso no período analisado. Não é de se surpreender,

Tabela 3.9 – Variáveis utilizadas - rio São Francisco.

Variável	**Amostras**	**Média**	**Mediana**	**Desvio-padrão**	**Mín.**	**Máx.**
c_cv_01544017	4017	1,60	0,00	7,39	0,00	110,00
c_cv_01544032	4017	2,30	0,00	8,13	0,00	125,40
c_cv_01544036	4017	1,99	0,00	7,59	0,00	112,00
c_vz_44290002 (y)	4017	1115,88	812,26	998,40	254,75	7338,65

Fonte: O autor

contudo, que estes valores zero tenham sido inseridos quando da imputação dos dados, visto que este trecho da série temporal era onde estava a maior lacuna. No entanto, não foi feita substituição dos valores zero por, digamos, a média de vazão. Estas e outras incertezas que permearam todas as análises foram, onde puder e couber, discutidas, mas manteve-se o trabalho mesmo com estas questões levantadas, sem fazer um tratamento específico. Uma observação geral sobre os dados de vazão é que existe uma amplitude elevada entre o mínimo e o máximo, em todas as estações utilizadas, com uma pequena variação para o rio Grande. Mas o rio Doce é realmente considerável – tabela 3.7. Vale destacar, no entanto, que esta amplitude não especifica se foi em um ano. É ao longo de toda série temporal, ou seja, ao longo dos 11 anos de dados considerados.

As variáveis de precipitação, mesmo considerando o somatório diário de precipitação, tiveram muitos dados zero. Nota-se isso a partir da análise da coluna "Mediana", que significa metade de toda massa de dados de precipitação estavam abaixo deste valor, que é o valor central dos dados, ou seja, metade de todos os dados de precipitação estavam em $0,00 mm/dia$. Isso reflete a dificuldade em se obter dados de precipitação. Essa enorme quantidade de valores zerados poderiam prejudicar as análises. Foi então que algumas variáveis adicionais foram inseridas para todos os rios deste trabalho em vias de aflorar a importância das precipitações nas análises, tentar dar um peso significativo a elas no treinamento dos modelos. As novas variáveis foram:

Tabela 3.10 – Variáveis acumuladas para as precipitações.

Nome da variável	**Descrição**
<ESTAÇÃO>_soma_3_dias	Soma das precipitações dos últimos 3 dias
<ESTAÇÃO>_soma_7_dias	Soma das precipitações dos últimos 7 dias
<ESTAÇÃO>_media_30_dias	Média das precipitações dos últimos 30 dias

Fonte: O autor

Quanto aos dados categóricos utilizados neste estudo, foi inicialmente pensado utilizar as informações de dia do ano ('*dayofyear*'), semana do ano ('*week*'), mês ('*month*') e estação do ano. Com exceção da variável de estação do ano ('estacao'), para a qual foi desenvolvido um algoritmo específico, as demais informações foram extraídas utilizando a biblioteca Pandas (64). Porém, as variáveis '*week*' e '*dayofyear*' possuem uma elevada

cardinalidade, que vão de 1 a 53 e de 1 a 365 (366 em ano bissexto), respectivamente. Isso poderia representar um problema para o modelo de regressão linear, causar um enviesamento do modelo devido a estes valores ficarem numa escala superior às demais variáveis, fosse com os dados log-transformados, fosse com os dados normalizados de 0 a 1 (MinMax). E estas variáveis não devem ser transformadas, pois isso tiraria o caráter categórico delas. Como prática comum, aplicam-se transformações apenas nas variáveis contínuas e para as variáveis categóricas, codificações. Foi realizada Codificação Ordinal ('*Ordinal Encoding*')[1] e em seguida, Codificação *One-Hot* (*One Hot Encoding*)[2] nas variáveis categóricas. Mas a codificação '*one-hot*' cria uma dificuldade que é a elevação da dimensionalidade, ou seja, várias novas colunas são inseridas nos dados de treinamento.

Pois bem, havia uma dificuldade aqui, uma vez que essas variáveis categóricas seriam incorporadas visando capturar o comportamento sazonal da série temporal, mas o incremento na dimensionalidade imporia um maior esforço computacional. Numa análise mais aprofundada, optou-se por retirar ambas as variáveis '*week*' e '*dayofyear*', pois as informações que elas guardam poderiam estar contidas nas variáveis de 'estacao' e '*month*'. Há uma correlação elevada entre o dia do ano, a semana do ano e o mês do ano e como o problema era a cardinalidade, e consequentemente a dimensionalidade, daquelas variáveis, mês do ano ('*month*') passou, desta forma, a representar também a informação de que se precisava.

Observa-se que os regimes de precipitação e vazão tendem a se repetir nas estações de primavera e verão, com uma redução significativa durante o outono e inverno. Por isso 'estacao' foi utilizada. Todas as variáveis foram codificadas primeiro no esquema "*ordinal encoding*", ficando assim as faixas de valores foram 1 a 12 para '*month*' e de 1 a 4 para 'estacao'. Depois deste passo, foi aplicado '*One-Hot Encoding*', criando 12 colunas para '*month*' e 4 colunas para 'estacao' totalizando 16 colunas novas. Observe: se as variáveis para '*week*' e '*dayofyear*' fossem utilizadas, além das 16 novas colunas mencionadas antes, mais 418 novas colunas (419 em ano bissexto) seriam criadas. Isso colocaria dificuldades não apenas no pré-processamento dos dados e na execução dos modelos, mas também na própria interpretabilidade dos resultados e na análise SHAP das variáveis dos modelos.

3.6 ANÁLISE EXPLORATÓRIA DOS DADOS

Com os dados ajustados, algumas variáveis removidas e as séries temporais contínuas, deu-se início à análise exploratória dos dados. Esta etapa é fundamental para compreender o comportamento das séries temporais.

[1] Codifica a variável categórica como uma sequência de números inteiros, ou seja, 1 para verão, 2 para outono, 3 para inverno e 4 para primavera, por exemplo.

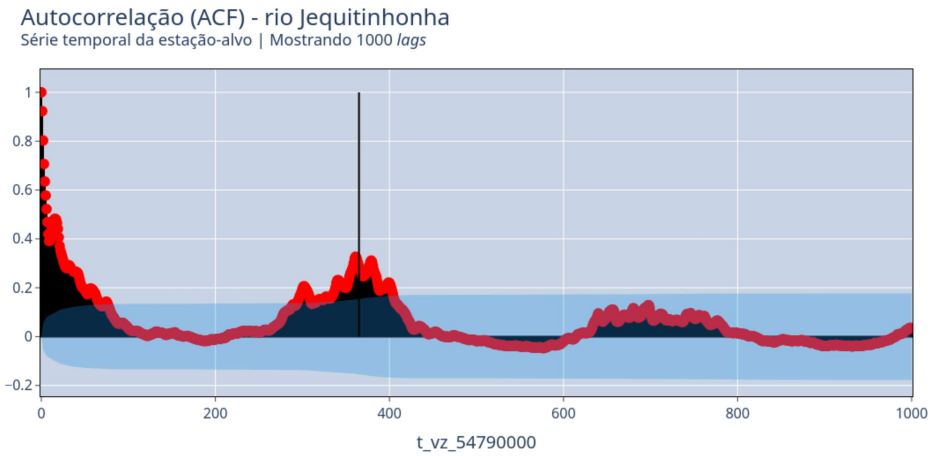
[2] Cria-se uma coluna para cada valor distinto de uma variável categórica da seguinte maneira: 'estacao_verao', 'estacao_outono' e assim por diante. Nas colunas criadas terão apenas valores 0 e 1, sendo 1 para quando a data corresponder ao período de verão e 0 no caso contrário

As análises realizadas foram idênticas para todos os rios estudados, de modo que a descrição desta fase será apresentada de forma geral, sem a necessidade de subdivisão por rio.

O primeiro passo foi verificar a sazonalidade dos dados. Foram avaliadas apenas as variáveis endógenas, ou seja, as vazões. Um teste de autocorrelação foi suficiente para identificar a presença de sazonalidade. Além disso, realizou-se a decomposição das séries temporais em suas componentes sazonais para uma análise mais detalhada. A função de autocorrelação (ACF) (52) é uma ferramenta essencial para identificar como os valores passados influenciam os valores futuros em uma série temporal, permitindo a detecção de padrões sazonais, ciclos e tendências.

A decomposição das séries temporais foi realizada utilizando a biblioteca StatsModels, aplicando o modelo aditivo.(86) A série temporal do rio Grande apresentou o pior desempenho em termos de autocorrelação – figura 3.35. A decomposição da série também revelou um comportamento mais ruidoso – figura 3.36. Nos gráficos de autocorrelação, a *lag* de 365 dias – que corresponde a um ano – foi destacada com uma linha preta vertical.

Figura 3.31 – Autocorrelação para a vazão do rio Jequitinhonha.



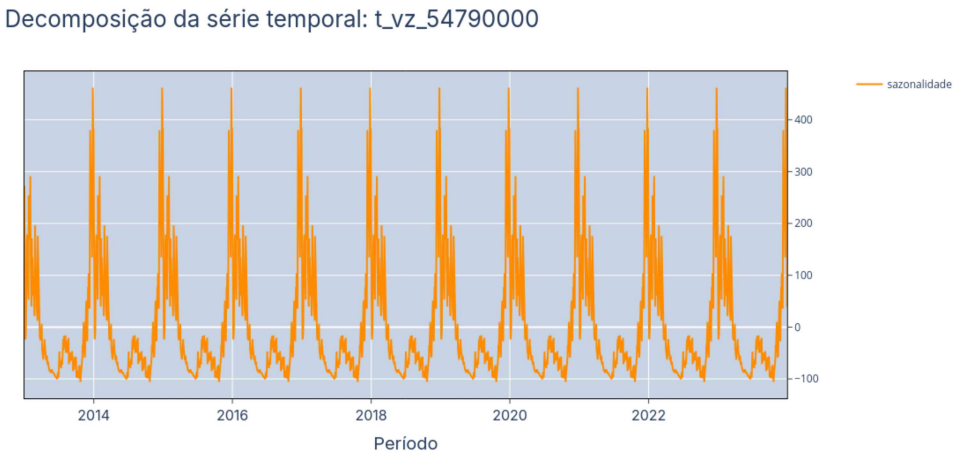
Fonte: O autor

As séries temporais consideradas, digamos, mais bem comportadas foram as dos rios Doce e São Francisco – figuras 3.33 e 3.37. A decomposição sazonal da série do rio Jequitinhonha apresentou algum nível de ruído, embora a sazonalidade tenha sido caracterizada de forma clara.

Outra característica investigada neste estudo foi a presença de 'cauda longa' nos dados de precipitação e vazão. Esse comportamento é comumente observado em dados ambientais dessa natureza.(63)

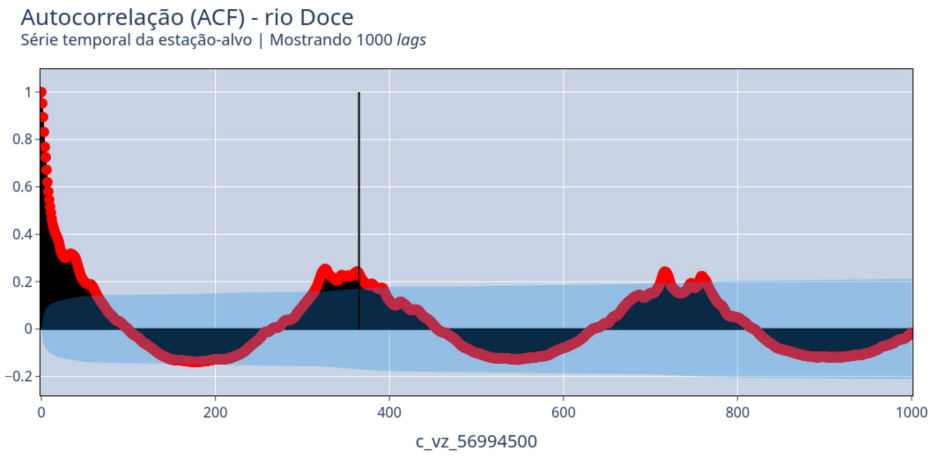
A 'cauda longa' refere-se a uma distribuição de frequência na qual uma proporção significativa dos eventos ocorre em uma região distante do centro ou da média da

Figura 3.32 – Componente sazonal da série de vazão do rio Jequitinhonha.



Fonte: O autor

Figura 3.33 – Autocorrelação para a vazão do rio Doce.

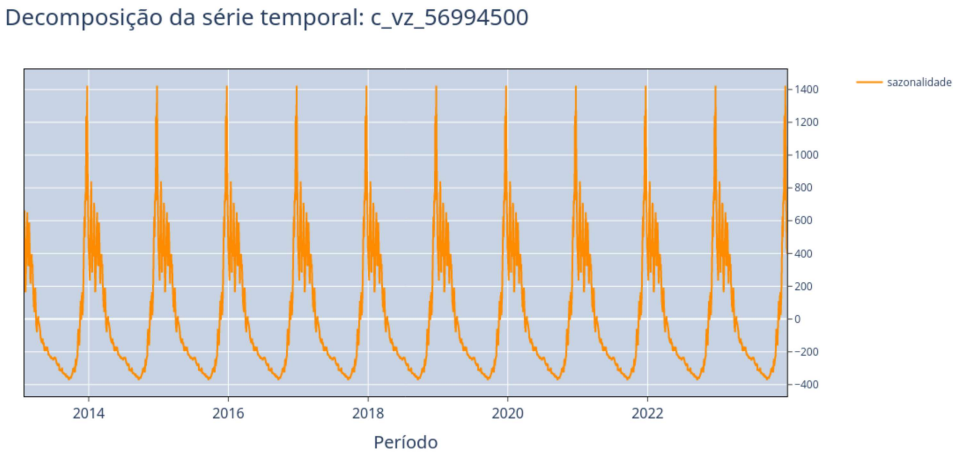


Fonte: O autor

distribuição, causando assimetria em tal distribuição. Quando uma distribuição segue a normal Gaussiana, ou se apróxima desta, a maioria dos eventos se concentra em torno da média, com poucas ocorrências nas extremidades, as caudas. No entanto, na distribuição com cauda longa, essas extremidades contêm uma quantidade substancial de eventos, que, somados, podem representar uma fração importante do total. A análise de cauda longa é um campo específico da estatística, desenvolvido para lidar com eventos de baixa frequência, mas de alta magnitude. No entanto, este trabalho não se aprofundou nas técnicas avançadas de análise de cauda longa; o foco aqui foi identificar a presença desse fenômeno e determinar um tratamento adequado para os dados.

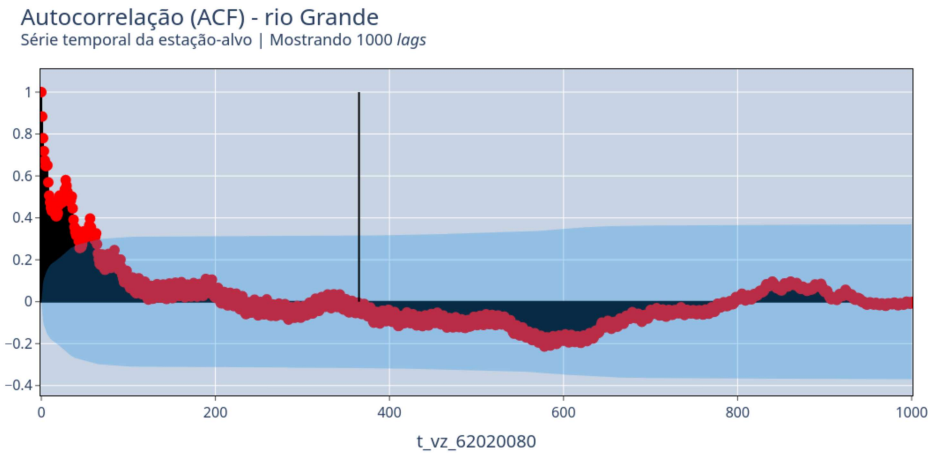
A mitigação do efeito de cauda longa é particularmente relevante para modelos como a Regressão Linear, que pressupõe uma distribuição normal dos dados. Uma distribuição

Figura 3.34 – Componente sazonal da série de vazão do rio Doce.



Fonte: O autor

Figura 3.35 – Autocorrelação para a vazão do rio Grande.



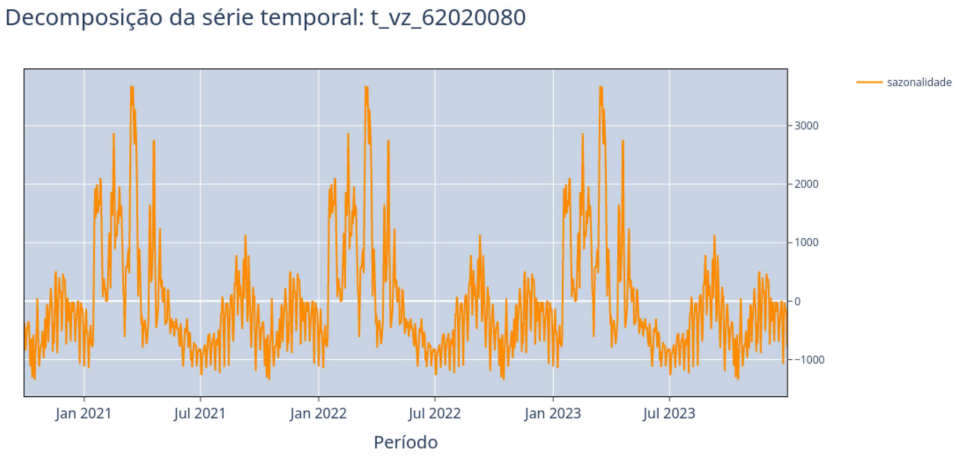
Fonte: O autor

assimétrica pode comprometer a convergência do modelo. Os dados transformados foram utilizados para todos os modelos, indistintamente, visando garantir uma padronização do estudo.

Posto que os dados contêm valores iguais a zero, a transformação pelo logaritmo natural (ln(.)) não foi aplicada, pois o cálculo de logaritmo não é definido para valores zero e negativos. Ao invés disso, foi utilizada a transformação 'log1p(.)' da biblioteca NumPy (45), que adiciona 1 ao valor antes da transformação, evitando erros relacionados ao logaritmo de zero.

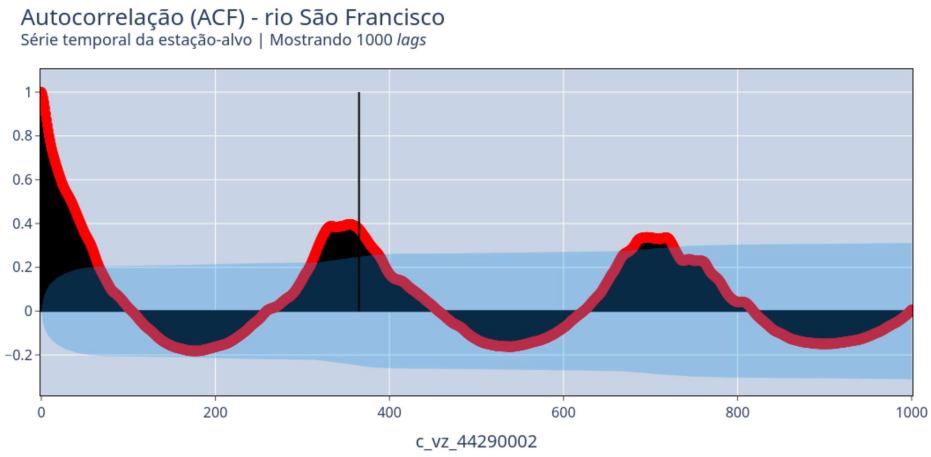
Após a transformação a distribuição dos eventos ficou menos assimétrica, como pode ser visto nas figuras 3.39, 3.40, 3.41, 3.42, 3.43, 3.44, 3.45, 3.46. Para o rio Grande, visualmente, parece não ter havido tanta diferença, mas quando se analisa os valores

Figura 3.36 – Componente sazonal da série de vazão do rio Grande.



Fonte: O autor

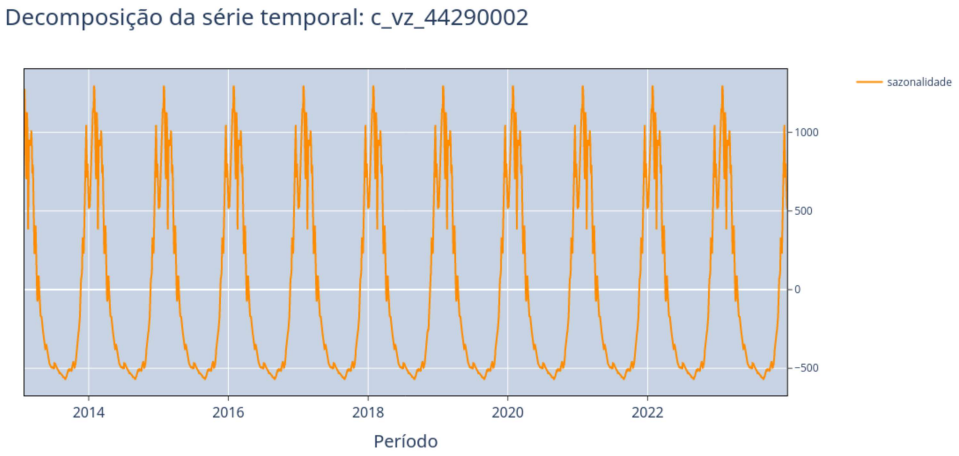
Figura 3.37 – Autocorrelação para a vazão do rio São Francisco.



Fonte: O autor

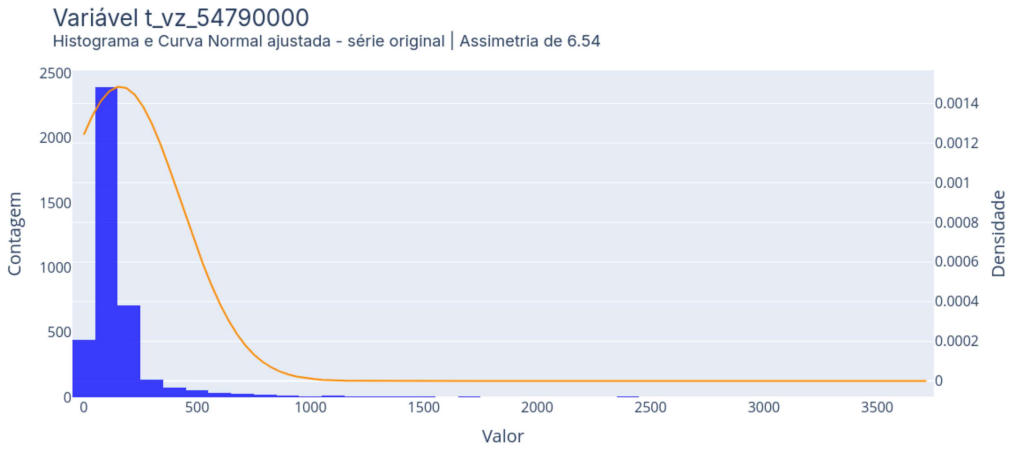
originais e transformados, houve um achatamento na distância entre os valores máximo e o mínimo da série.

Figura 3.38 – Componente sazonal da série de vazão do rio São Francisco.



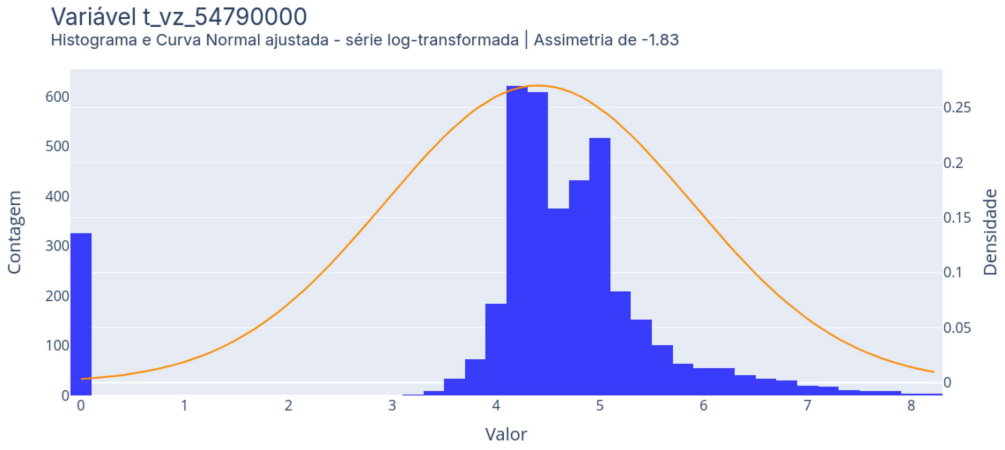
Fonte: O autor

Figura 3.39 – Dados originais para o rio Jequitinhonha.



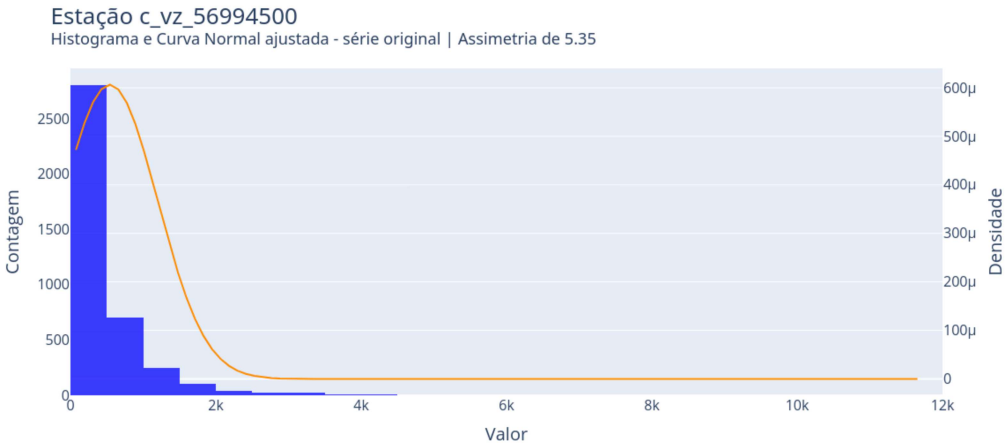
Fonte: O autor

Figura 3.40 – Dados log-transformados para o rio Jequitinhonha.



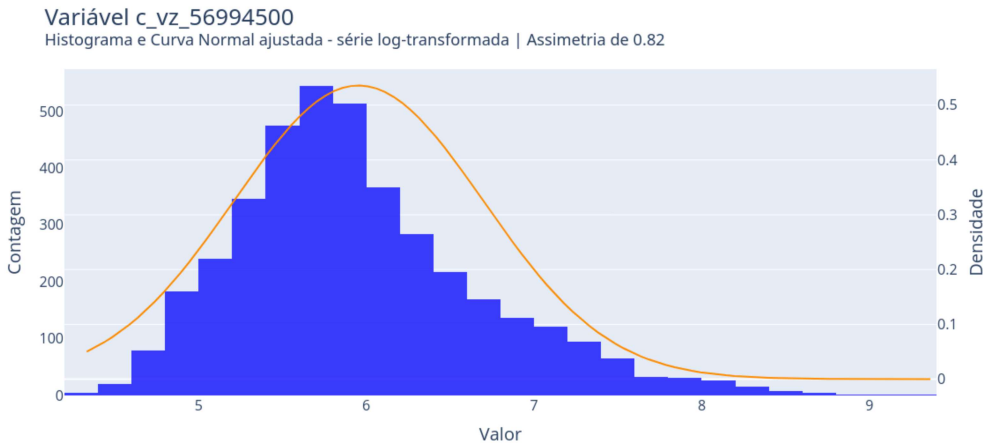
Fonte: O autor

Figura 3.41 – Dados originais para o rio Doce.



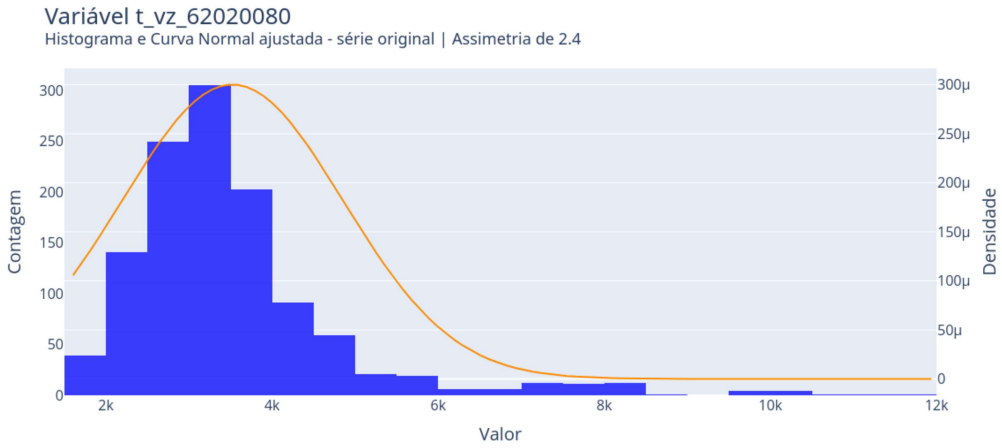
Fonte: O autor

Figura 3.42 – Dados log-transformados para o rio Doce.



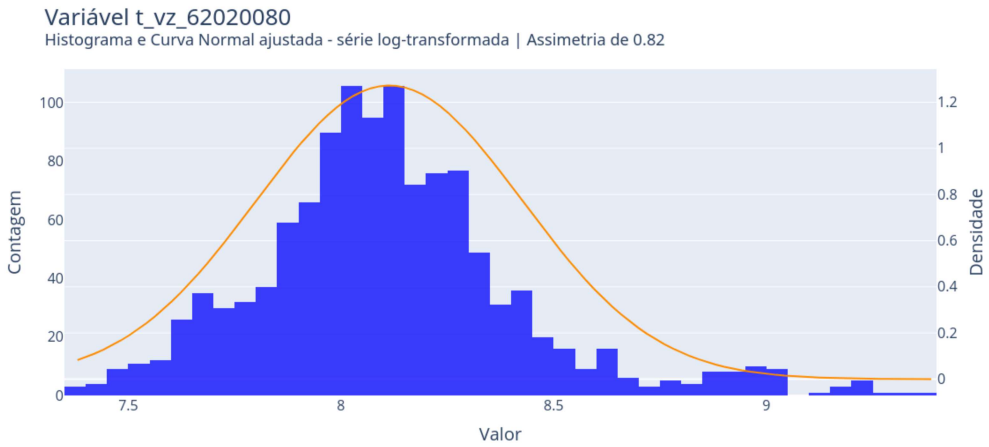
Fonte: O autor

Figura 3.43 – Dados originais para o rio Grande.



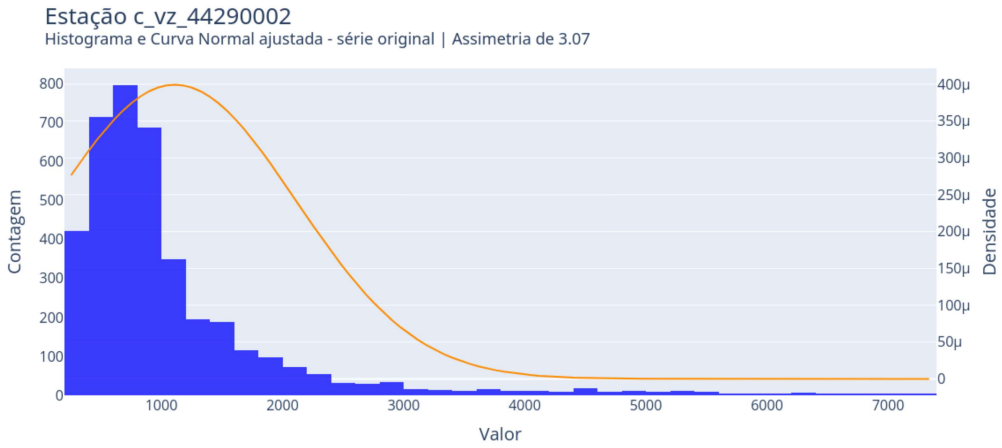
Fonte: O autor

Figura 3.44 – Dados log-transformados para o rio Grande.



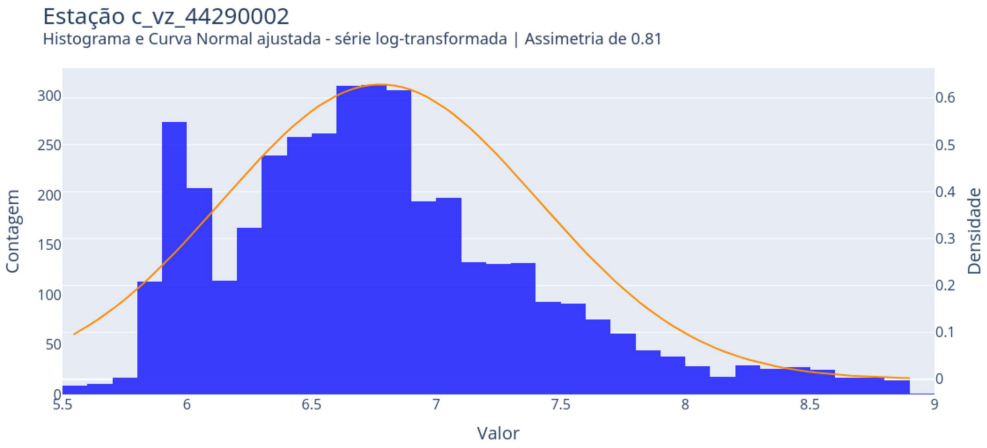
Fonte: O autor

Figura 3.45 – Dados originais para o rio São Francisco.



Fonte: O autor

Figura 3.46 – Dados log-transformados para o rio São Francisco.



Fonte: O autor

3.7 MÉTRICAS DE AVALIAÇÃO

Para avaliar o desempenho dos modelos de previsão utilizados neste estudo, foram adotadas quatro métricas: MAPE, PBIAS e KGE. A escolha dessas métricas baseia-se na necessidade de uma avaliação abrangente que considere diferentes aspectos da qualidade das previsões, como precisão, erro médio, tendência e correlação.

- ***Mean Absolute Percentage Error*** – **MAPE**: A MAPE é uma métrica amplamente utilizada para medir a precisão das previsões em termos percentuais. O algoritmo calcula a média das diferenças absolutas entre os valores observados e previstos, normalizadas pelos valores observados. O bom da métrica MAPE é a sua facilidade de interpretação, já que expressa o erro em termos percentuais, e por ser "livre de escala", ou seja, independente da escala dos dados, tornando os resultados comparáveis entre diferentes séries temporais e modelos. Contudo, a MAPE pode ser sensível a valores muito baixos e esta característica deve ser considerada ao interpretar os resultados. Quanto mais próximo de 0, melhor.(49)

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{O_i - P_i}{O_i} \right| \tag{3.1}$$

  - ◇ O_i valores observados
  - ◇ P_i valores previstos
  - ◇ n o número total de observações

- ***Percent Bias*** – **PBIAS**: A PBIAS avalia o viés das previsões, ou seja, a tendência do modelo em superestimar (PBIAS positivo) ou subestimar (PBIAS negativo) os valores observados. Ela expressa a diferença percentual entre a soma dos valores previstos e observados, permitindo identificar se o modelo apresenta uma tendência sistemática de erro. Um valor de PBIAS próximo de zero indica que o modelo não possui viés significativo. Não se espera que esta métrica seja 0, senão indicaria que a previsão foi exatamente o valor observado, mas ao mostrar o viés das previsões, isso tem impacto diretamente nas decisões de gestão de recursos hídricos.

$$PBIAS = 100 \times \frac{\sum_{i=1}^{n} (P_i - O_i)}{\sum_{i=1}^{n} O_i} \tag{3.2}$$

  - ◇ O_i valores observados
  - ◇ P_i valores previstos
  - ◇ n o número total de observações

- ***Kling-Gupta Efficiency*** – **KGE**: A KGE fornece uma avaliação integrada do desempenho do modelo, considerando simultaneamente três componentes: correlação, viés e variabilidade relativa entre os valores previstos e observados. A KGE é uma

métrica robusta que combina esses três fatores de forma equilibrada, fornecendo um entendimento geral da qualidade das previsões. Essa métrica é especialmente útil em estudos hidrológicos, pois tem capacidade de capturar a complexidade das relações entre variáveis hidrológicas de maneira mais eficaz do que métricas tradicionais focadas em um único aspecto. Quanto mais próximo de 1, melhor o desempenho do modelo.(44)

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{3.3}$$

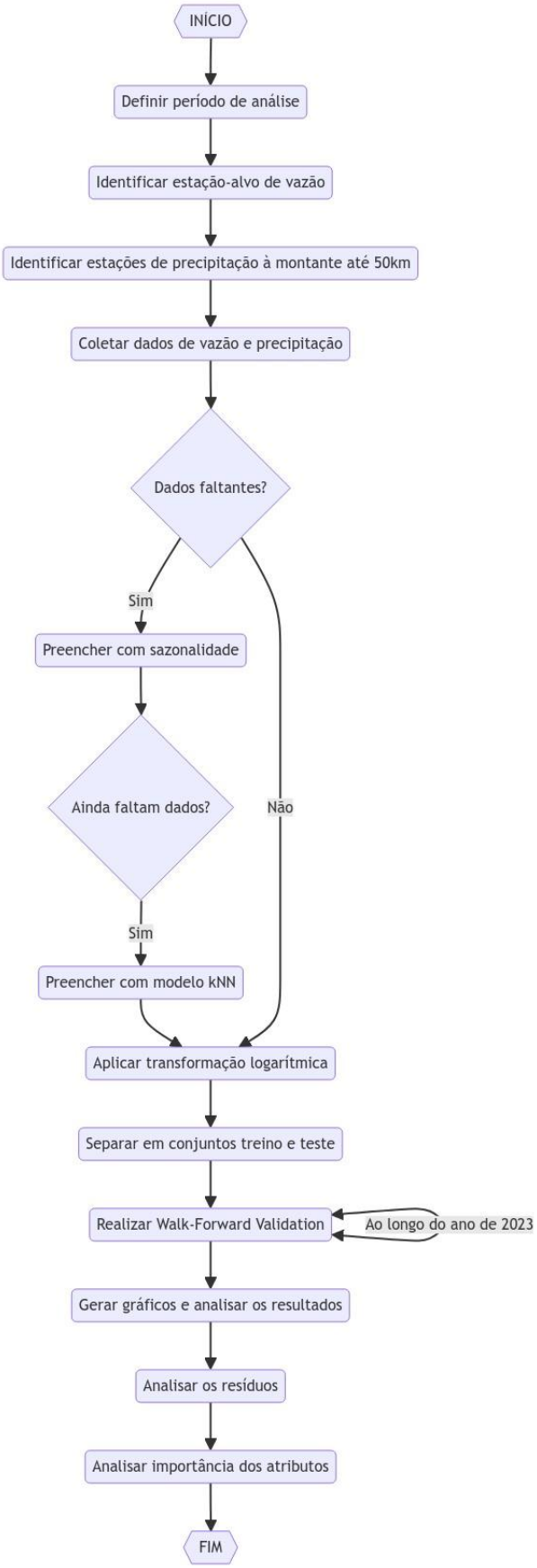
- r é o coeficiente de correlação linear entre os valores observados e previstos
- $\alpha = \frac{\sigma_p}{\sigma_o}$ é a variabilidade relativa, sendo σ_p o desvio-padrão das previsões e σ_o o desvio-padrão das observações
- $\beta = \frac{\mu_p}{\mu_o}$ é o viés, em que μ_p é a média dos valores previstos e μ_o a média dos valores observados

Os modelos serão qualificados quanto ao seu desempenho seguindo a seguinte ordem de métricas: 1. KGE, 2. MAPE, 3. PBIAS e 4. Cobertura empírica do intervalo de previsão, que será apresenta adiante.

3.8 MODELO PROPOSTO

Tudo detalhado até aqui, agora é preciso descrever o fluxo de trabalho. A figura 3.47 apresenta o fluxograma completo do trabalho para rápida depreensão. O texto que se segue descreve os passos.

Figura 3.47 – Fluxo de trabalho.



Fonte: O autor

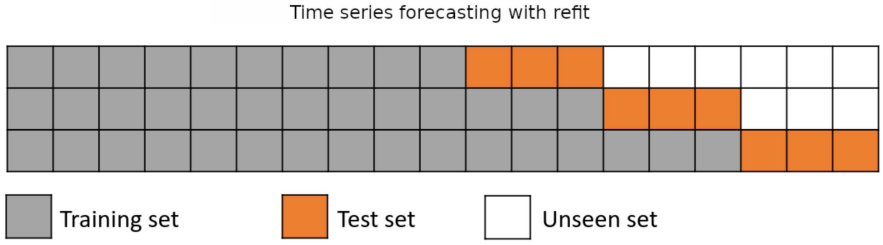
Primeiramente, é necessário especificar o período histórico de análise. Após estipulado o período, deve-se identificar as estações de vazão para as quais serão realizadas as previsões.

Identificadas estas estações de vazão, verifica-se em linha reta, em sentido montante da estação de vazão, as estações de precipitação a 50km nas sub-bacias que deságuam no corpo hídrico analisado. Para este passo foi utilizado o programa QGis com *shapefiles* das ottobacias (90). Para o *shapefile* de todas as estações do estado de Minas Gerais foi preciso acesso ao sistema da empresa Rhama Analysis. Sistema fechado este, no caso.

Aqui uma importante etapa: aplicar transformação logarítmica nos dados de vazão e precipitação para tratar o fenômeno de "cauda longa" dos dados.

O conjunto teste corresponde aos dados do último ano, 2023, e os demais correspondem ao conjunto treino. Realizar ajuste ('*fit*') e previsão ('*predict*'), dia-a-dia, durante um ano inteiro, para avaliar a qualidade das previsões e a estabilidade dos modelos, simulando o uso futuro de um sistema, em que o usuário vai treinar e prever (*fit*/*predict*) diariamente. Este procedimento é referido na literatura como *Walk-Forward Validation*, WFV, e pode ser por uma janela deslizante ou por janela expandida, que foi a técnica aplicada neste trabalho. Além disso, tem duas formas de realizar as previsões: criando vários modelos que irão prever cada um dos dias futuros ('*direct forecasting*'), o que geraria 365 modelos distintos e, consequentemente, consumiria muita memória, ou o que foi adotado neste trabalho que foi realizar reajustes ('*refit*') ao modelo inicial num processo recursivo ('*recursive forecasting*'). Por isso os modelos no trabalho são considerados dinâmicos, como informado na seção da Taxonomia da Série Temporal. Esse processo WFV permite capturar mudanças ao longo do tempo e avaliar a capacidade do modelo de generalizar para o futuro de forma mais realista. Perceba o funcionamento nas figuras 3.48, 3.49, 3.50, 3.51.

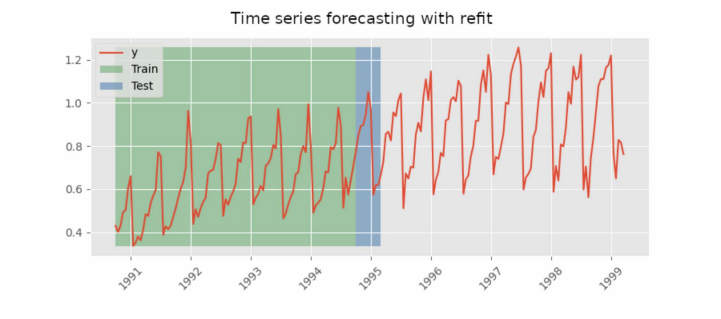
Figura 3.48 – Diagrama mostrando a divisão dos dados de treino/teste com *refit*.



Fonte: Rodrigo e Ortiz (30)

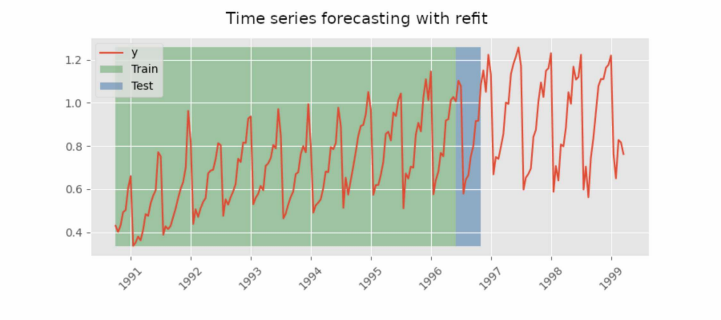
Foram calculados intervalos de previsão ('*prediction intervals*') de 95%. A interpretação do intervalo de previsão é: calculados os valores inferior e superior do intervalo de previsão de 95% ('*lo-95*' e '*hi-95*'), o valor real observado futuro tem 95% de probabilidade de estar dentro destes limites.(48) Apesar de ser usado intervalo de previsão de 95% na análise, nem sempre o valor real é capturado. Existe uma incerteza quanto à previsão. Esta

Figura 3.49 – WFV com janela expandida e *refit* - imagem 1.



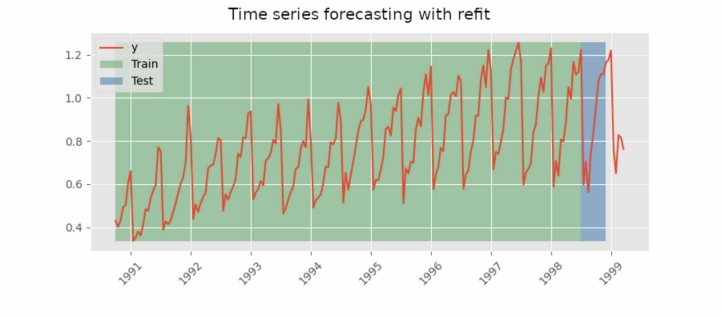
Fonte: Rodrigo e Ortiz (30)

Figura 3.50 – WFV com janela expandida e *refit* - imagem 2.



Fonte: Rodrigo e Ortiz (30)

Figura 3.51 – WFV com janela expandida e *refit* - imagem 3.

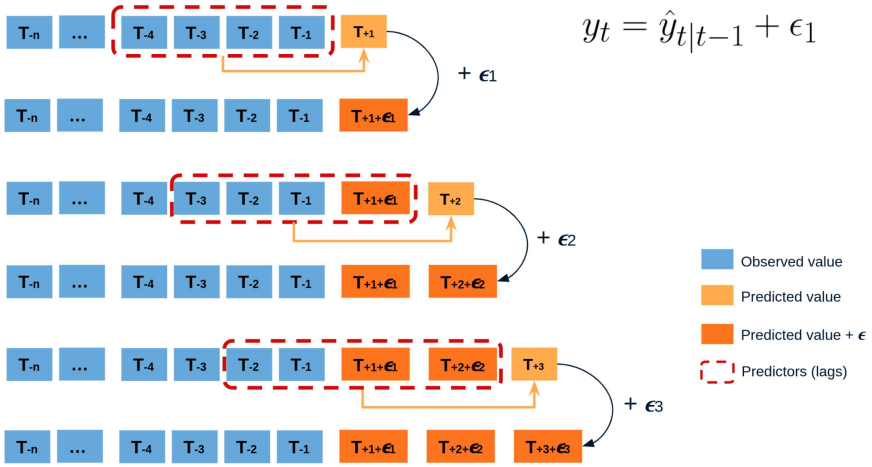


Fonte: Rodrigo e Ortiz (30)

incerteza advém da própria natureza dos dados hidrológicos e também por assumir que os erros do passado se manterão semelhantes nas previsões. Estes erros são os chamados "*in sample residuals*", resíduos das previsões realizadas no passo de treinamento dos dados. É com estes resíduos, que são sorteados da massa de resíduos e geram novas prováveis previsões, que, no final, comporão o cálculo do intervalo de previsão – figura 3.52.

Um fenômeno que pode acontecer é os intervalos de previsão serem estreitos.(47) Isso porque os modelos não conseguem capturar todas as fontes de incerteza a partir dos dados, a saber:

Figura 3.52 – Diagrama de como se calcula o intervalo de previsão.



Fonte: Rodrigo e Ortiz (31)

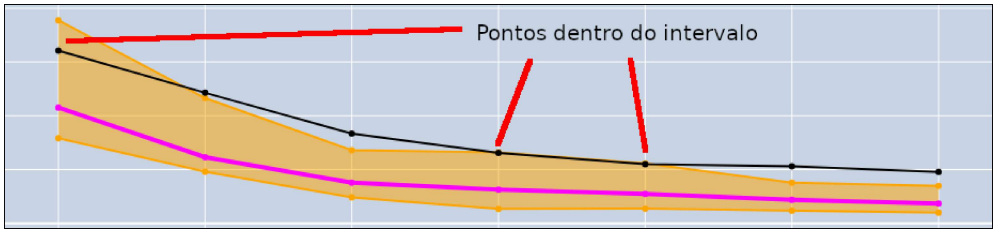
- O termo de erro aleatório
- As estimativas dos parâmetros dos modelos
- A escolha do modelo para trabalhar os dados históricos
- A continuação, no futuro, do processo que gerou os dados históricos

Ao produzir intervalos de previsão para modelos de séries temporais, geralmente apenas a primeira dessas fontes é levada em consideração. Portanto, foi criada uma métrica para avaliar a qualidade dos intervalos de previsão, que verificou a cobertura empírica dos intervalos. Não se confiou apenas na cobertura esperada para os intervalos.

Ficou assim: para cada dia previsto foi feito um teste lógico para verificar se o valor real observado ('y_true') esteve acima do valor mínimo do intervalo ('lo-95') **e** abaixo do valor máximo ('hi-95'). Para quando estes dois testes lógicos fossem verdadeiros, adicionava o valor 1 a um vetor e 0 quando um destes testes lógicos falhasse, ou seja, o 'y_true' estava de algum modo fora do intervalo. Ao final, calculou a média deste vetor e multiplicou por 100 para uma saída percentual. Se o número final fosse 95% a 100% então significa que os intervalos de previsão capturaram perfeitamente o valor real futuro. Se abaixo de 95%, significa que em algum ponto da previsão o valor real ficou para fora destes limites, ou abaixo do valor mínimo, ou acima do valor máximo. Veja uma demonstração.

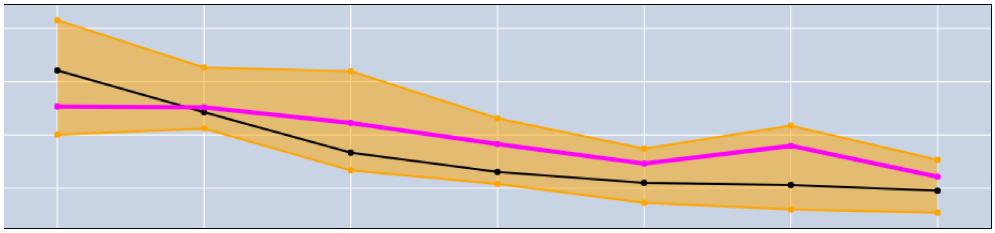
Aplicando a função detalhada anteriormente à figura 3.53 retornou $42,86\%$, os 3 pontos da linha preta, os valores reais observados, de um total de 7. Isso significa que o intervalo de previsão pretendido de 95% só conseguiu, empiricamente, capturar cerca de 43% dos valores reais do horizonte de previsão em questão. Na figura 3.54, por sua vez, o intervalo capturou 100% dos valores observados. Ambas figuras mostram previsão

Figura 3.53 – Previsão com valor fora dos limites.



Fonte: O autor

Figura 3.54 – Previsão completamente nos limites.



Fonte: O autor

de 7 dias, e sem considerar a linha rosa, que foi a previsão que os modelos geraram, o modelo da figura 3.54 teve um desempenho melhor, segundo essa métrica criada. É mais uma questão a ser considerada quando avaliar a qualidade dos modelos, mas deve ser empregada com parcimônia. Isso porque os intervalos de previsão podem assumir faixas de valores muito elevados. Se, por exemplo, o modelo calcular intervalos de previsão que vão de $0m^3/s$ a 200 mil m^3/s, então qualquer valor observado seria capturado, de fato, mas isso não ajudaria em nada na tomada de decisão.

Encerrada essa etapa, procede-se com a análise dos resíduos. Os modelos não-lineares *CatBoost* e *Random Forest* não necessitam que haja normalidade, nem para os dados de entrada, nem para os resíduos. Diferentemente do modelo *Linear Regression*, em que é necessário realizar tal avaliação em busca de compreender o funcionamento do modelo.

Analisar os resíduos permite verificar a estabilidade do modelo quanto às previsões. Uma distribuição normal dos resíduos é um indício de que os coeficientes do modelo conseguiram capturar as relações temporais dos dados de entrada. Além disso, analisar os resíduos permite avaliar se o modelo assumiu uma correta relação entre a variável alvo e as variáveis independentes.(3)

A última análise é avaliação da importância das variáveis. Apenas os melhores resultados serão analisados de cada modelo. Serão apresentados os valores SHAP das variáveis dos modelos e discussão sobre o comportamento de cada uma delas para a qualidade dos modelos.

SHapley Additive exPlanations – SHAP values – constitui-se de uma metodolo-

gia robusta, derivada da teoria dos jogos cooperativos, com o propósito de elucidar o funcionamento de modelos de aprendizado de máquina, frequentemente caracterizados como "caixas pretas". A premissa fundamental reside na atribuição de um valor a cada característica preditiva, quantificando sua contribuição individual para a predição final do modelo, considerando todas as interações possíveis com demais características. Essa abordagem viabiliza a compreensão dos fatores determinantes na tomada de decisão do modelo e como cada variável impacta o resultado, seja de forma positiva ou negativa, promovendo assim interpretabilidade e transparência (66).

Não consta no fluxograma do modelo geral proposto, mas para todos os modelos (*CatBoost*, *Random Forest* e *Linear Regression*) teve também uma rodada de execução com os dados originais sem transformação logarítmica, apenas havendo aplicação de escalonamento de dados usando MinMax para o *Linear Regression*. Mas isso é obrigatório para este tipo de modelo, pois a escala dos dados é muito discrepante, com vazão em m^3/s e precipitação em mm/dia. Contudo, o escalonamento não modifica a distribuição dos dados, mantendo o mesmo comportamento dos dados originais. Essa característica é importante. Todas as comparações serão importantes para chegar em uma melhor interpretação dos fenômenos e dar direcionamentos sobre quais modelos e como utilizá-los pensando em termos de gestão dos recursos hídricos.

O ambiente computacional em que os experimentos foram executados é:

Tabela 3.11 – Configurações do computador usado nos experimentos.

Configuração	**Descrição**
CPU	Ryzen 7 5700 (8 núcleos/16 *threads*)
GPU	RTX 3060
Memória RAM	32 GB 3200MHz
Disco	SSD NVMe
Kernel	Linux 6.8
Python	versão 3.9

Fonte: O autor

Os resultados serão apresentados no capítulo 4.

4 RESULTADOS E DISCUSSÃO

4.1 DESEMPENHO DOS MODELOS

Os resultados obtidos pelos modelos apresentaram variações significativas em termos de precisão, tanto nas previsões pontuais quanto nos intervalos de confiança, refletindo diferentes graus de precisão para cada cenário analisado. A análise dos resíduos revelou-se particularmente interessante, pois evidenciou comportamentos distintos em função da sazonalidade, com variações marcantes observadas em diferentes períodos do ano.

4.1.1 Rio Jequitinhonha

Iniciando por esta bacia hidrográfica, os resultados revelaram-se bastante satisfatórios, especialmente no que se refere ao modelo LR – a partir daqui os modelos serão referidos por suas siglas em inglês –, que se destacou com previsões precisas, tanto em termos pontuais quanto nos intervalos de confiança. Este modelo demonstrou uma capacidade robusta de capturar a dinâmica hidrológica do rio, refletindo precisão nas métricas e desempenho consistente, aliado a um tempo de execução significativamente reduzido. Em contrapartida, o modelo (simples) *Seasonal Naive* (SN) apresentou resultados abaixo das expectativas em todas as situações avaliadas – figura 4.1.

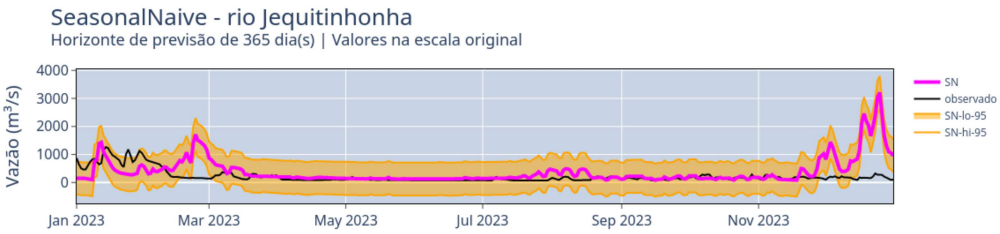
Com um MAPE extremamente elevado, de 150%, os resultados indicaram um viés significativo de superestimação, conforme evidenciado pelo PBIAS. Verificando o KGE, ficou negativo. Em outras palavras, isso significa que o modelo não apenas falha em capturar a variabilidade dos dados observados, mas também introduz erros que o tornam menos eficaz do que uma abordagem simplista, como utilizar a média histórica. Embora a análise superficial da qualidade dos intervalos de confiança pudesse sugerir um desempenho satisfatório do modelo, uma inspeção mais detalhada revela uma incongruência: os valores inferiores do intervalo (lo-95) foram calculados abaixo de 0, o que não faz sentido para o rio em questão.

Considerando o desempenho insatisfatório do modelo SN, a análise foi encerrada neste ponto, sem proceder com a avaliação dos resíduos ou a análise da importância das variáveis. O modelo foi incluído apenas para fins comparativos. A análise mais aprofundada será dedicada aos modelos mais complexos, que apresentaram desempenho superior.

Os resultados obtidos utilizando o modelo LR mostraram-se bastante promissores – figura 4.2. Nesta primeira avaliação, os dados foram log-transformados. Isso é verificável pelo título com o uso de “destransformados”. Os dados foram log-transformados e retornados para a escala original para desenhar o gráfico e ficarem acessíveis para quem lê. Essa dinâmica no título se manterá por todo trabalho.

Partindo pelo KGE calculado, o resultado mostrou-se excelente. Recordando:

Figura 4.1 – Resultado do *Seasonal Naive* no teste *Walk-Forward Validation*.



Modelo	MAPE	KGE	PBIAS (%)
SeasonalNaive	1.50031	-0.50459	75.54145
SeasonalNaive-lo-95	3.88613	-1.34311	-194.85927
SeasonalNaive-hi-95	5.76158	-2.69605	345.94216

Cobertura empírica do intervalo de previsão
86.3 %

Fonte: O autor.

quanto mais próxima de 1, melhor. Este valor sugere que o modelo foi eficaz na previsão do comportamento hidrológico do sistema em análise, oferecendo previsões que estão bem alinhadas com os dados observados. O MAPE de $14,5\%$ sugere que o modelo tem uma precisão média absoluta razoável e é bastante confiável. O modelo apresentou um viés sistemático de subestimar os resultados, conforme aponta o PBIAS de $-0,97\%$. Considerando todas as métricas, o resultado indica que o modelo teve um desempenho global ótimo. O KGE alto é particularmente indicativa de um bom ajuste global, com o modelo capturando bem tanto a dinâmica quanto a magnitude dos dados observados.

Considerando a qualidade dos intervalos de previsão, a cobertura observada de $98,08\%$ excedeu o intervalo teórico calculado de 95%, o que, à primeira vista, poderia ser interpretado como um desempenho satisfatório. No entanto, o limite superior do intervalo (hi-95) mostrou-se excessivamente elevado nos meses de janeiro e fevereiro, o que pode comprometer a interpretação dos resultados. Isso ocorre porque intervalos de previsão excessivamente amplos podem capturar praticamente qualquer valor observado, reduzindo a utilidade prática da previsão.

É importante destacar que os intervalos de previsão são calculados a partir dos erros do modelo durante a etapa de treinamento – *bootstrap in-sample residuals*. Para uma análise mais aprofundada desse comportamento, é necessário examinar os dados anteriores a 2023. A figura 4.3 revela que, nos últimos dias de 2022, houve observações significativamente elevadas. É provável que os erros associados a esse período mais próximo tenham influenciado a amplitude dos intervalos de previsão subsequentes. Essa

interpretação é suportada pela observação de que o comportamento ao final de 2023 foi mais estável, possivelmente devido à ausência de eventos ruidosos imediatamente anteriores – meses de maio a agosto.

Figura 4.2 – *Walk-Forward Validation* para o modelo Regressão Linear.

Modelo	MAPE	KGE	PBIAS (%)
LinearRegression	0.14576	0.95461	-0.87104
LinearRegression-lo-95	0.40864	0.35472	-43.68866
LinearRegression-hi-95	0.88114	-0.64261	99.41937

Cobertura empírica do intervalo de previsão
98.08 %

Fonte: O autor.

Figura 4.3 – Detalhe do trecho final dos dados, em 2022, usados para treinamento.

Fonte: O autor.

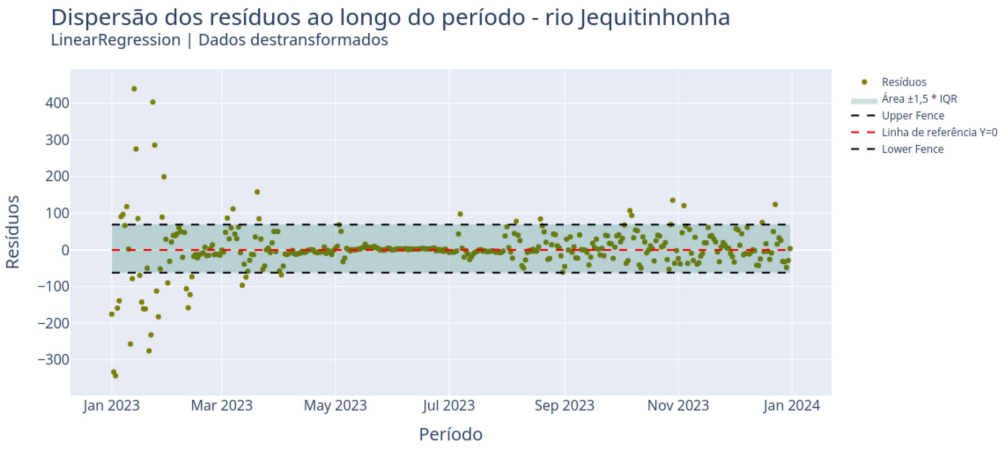
A área sombreada demarcada na figura 4.4 são os valores calculados para *lower fence* e *upper fence* – como se vê em um *boxplot* – a partir do primeiro $(-12,83)$ e terceiro $(19,98)$ quartil, que aqui resultaram em $-62,06$ e $69,21$, respectivamente. Entre estes valores é onde se espera que os resíduos estejam distribuídos (houve uma prevalência de

87, 4% dos resíduos nesta área), o que fica de fora dessas faixas pode ser interpretado como um *outlier* ou erro. Considerando que os dados de treinamento não foram tratados para valores *outliers*, isso parece estar se refletindo neste resultado, mesmo com a transformação logarítmica. O modelo não captou corretamente os valores elevados nos dados observados, por isso resíduos tão grandes assim, ainda que tenha apresentado um comportamento geral bom, como visto nas métricas anteriormente. Outro fator a se considerar também é que por estar na escala log, qualquer pequena variação nesta escala, quando retornada para a escala original, pode gerar números elevados.

Nesta mesma figura 4.4 vê-se como os resíduos estão dispersos ao longo do tempo. Isso ajuda a compreender a estabilidade do modelo. No início do ano e final do ano, exatamente quando os eventos de chuva mais ocorrem, o comportamento geral de vazões torna-se incerto e o modelo tende a se espalhar mais. No início do ano, como discutido anteriormente, pode ser que devido às vazões elevadas do final do ano de 2022, isso tenha causado ruídos em excesso na previsão do modelo. Quando houve vazões mais moderadas, os resíduos foram também mais moderados, como pode observar no final do ano de 2023, em que a influência imediata é o meio do ano, meses de inverno, e início de primavera. Quando da estação de baixa dos rios, os meses de outono e inverno, os resíduos estiveram bem alinhados em torno do 0.

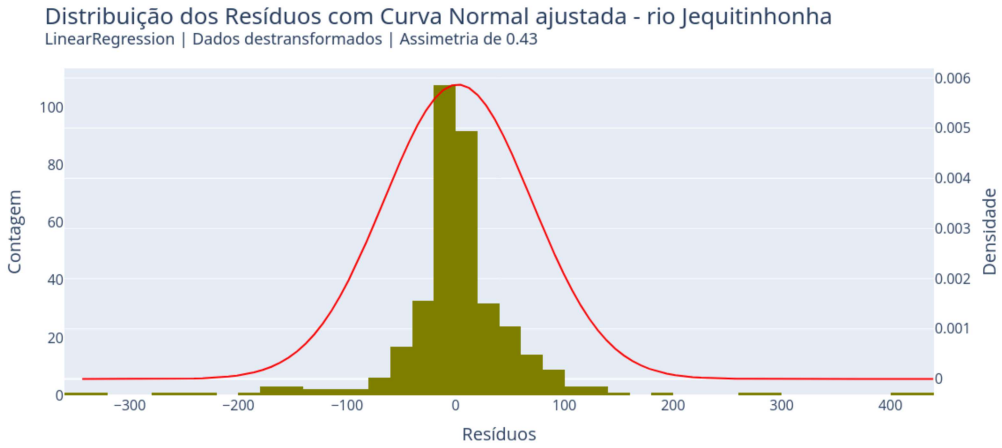
Para encerrar essa análise, a figura 4.5 apresenta que os resíduos estão próximos da normalidade quanto à distribuição destes em torno de 0, com uma assimetria de $0,48$. O valor positivo para assimetria sugere que houve casos em que o modelo subestimou os valores na previsão – existe uma cauda à direita do centro dos dados –, o que combina com a métrica PBIAS. Mas cabe destacar que os resíduos não estão significativamente desviados em nenhuma direção, apresentam distribuição concentrada em torno de 0 e este é um comportamento desejado. A análise da função de autocorrelação (ACF) está na figura 4.6 e aqui é analisado se existe independência ou dependência temporal entre os resíduos. A área sombreada representa a faixa ideal de permanência dos resíduos, onde se espera que a maioria dos resíduos esteja concentrada caso não haja autocorrelação significativa. Algumas *lags* estão fora destes limites (picos), mais precisamente, *lags* que estão mais próximas da *lag* de referência. Este comportamento indica que o modelo pode ser refinado para melhor capturar dinâmicas temporais, possivelmente incorporando termos de tendência. Mas no aspecto geral, está bom o comportamento dos resíduos. Para melhorar os resíduos com valores tão elevados, um tratamento de *outliers* pode verter bons resultados.

Figura 4.4 – Resíduos da previsão ao longo do tempo.



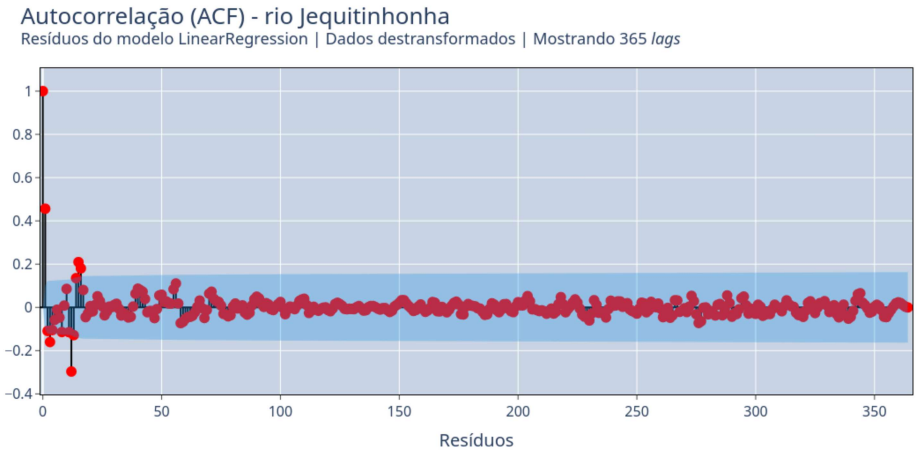
Fonte: O autor.

Figura 4.5 – Histograma dos resíduos.



Fonte: O autor.

Figura 4.6 – Resíduos da previsão ao longo do tempo.



Fonte: O autor.

Passando agora à análise dos modelos principais deste trabalho, cujos resultados serão comparados ao modelo de referência – o modelo LR –, optou-se por realizar a análise de forma concomitante, uma vez que ambos os modelos apresentaram comportamentos similares, embora o modelo RF tenha demonstrado um desempenho superior ao CB. Neste ponto, os resultados dos modelos não-lineares passaram pela log-transformação.

Na métrica KGE, comparando-se os resultados com o comportamento visto na figura 4.2, observa-se que ambos os modelos não conseguiram superar o modelo de referência, com destaque para o CB, que apresentou um desempenho consideravelmente inferior – figuras 4.7 e 4.8. No entanto, ao considerar a previsão percentual média, avaliada pela métrica MAPE, ambos os modelos baseados em árvores apresentaram melhorias em relação ao modelo de referência, evidenciando um desempenho superior em termos de erro percentual médio.

O KGE, relembrando, combina três aspectos fundamentais: variabilidade, viés e correlação entre os dados observados e previstos. Considerando esses fatores, com os dados log-transformados, o modelo LR capturou de forma mais eficaz os três aspectos mencionados. É provável que a linearização tenha sido um fator determinante para o bom desempenho do modelo LR olhando por esta métrica, uma vez que, ao analisar o MAPE, os modelos não-lineares, CB com $0,13$ e RF com $0,12$, apresentaram desempenhos melhores. Um desempenho médio percentual melhor pode dever-se à resiliência dos modelos não-lineares à sua robustez diante de valores discrepantes, aos quais o modelo linear é mais sensível.

Observa-se que tanto o CB quanto o RF apresentaram intervalos de previsão menos amplos no início do ano, em comparação com o modelo LR, mesmo impactados pelas vazões elevadas no final de 2022, conforme figura 4.3. Em termos de previsão pontual, ambos os modelos não-lineares demonstraram-se robustos, no entanto, ao analisar a cobertura empírica dos intervalos de previsão, o modelo RF mostrou-se consideravelmente defasado em relação aos modelos LR e CB. Isso pode ser explicado pela possível "otimização" excessiva do modelo ao calcular os intervalos, ao presumir uma repetição dos eventos e erros passados, resultando em intervalos estreitos. Esse fenômeno é explicado por Hyndman (47). Apesar disso, a cobertura empírica de 80% do modelo RF ainda pode ser considerada satisfatória, e a cobertura de 86% obtida pelo CB demonstra um bom desempenho.

Figura 4.7 – *Walk-Forward Validation* para o modelo *CatBoost*.

Modelo	MAPE	KGE	PBIAS (%)
CatBoost	0.13031	0.84101	-6.2082
CatBoost-lo-95	0.23899	0.56573	-26.96171
CatBoost-hi-95	0.59863	0.30777	51.31453

Cobertura empírica do intervalo de previsão
86.85 %

Fonte: O autor.

Figura 4.8 – *Walk-Forward Validation* para o modelo *Random Forest*.

Modelo	MAPE	KGE	PBIAS (%)
RandomForest	0.12548	0.91602	-3.15394
RandomForest-lo-95	0.23379	0.60227	-25.15944
RandomForest-hi-95	0.35827	0.47249	34.02366

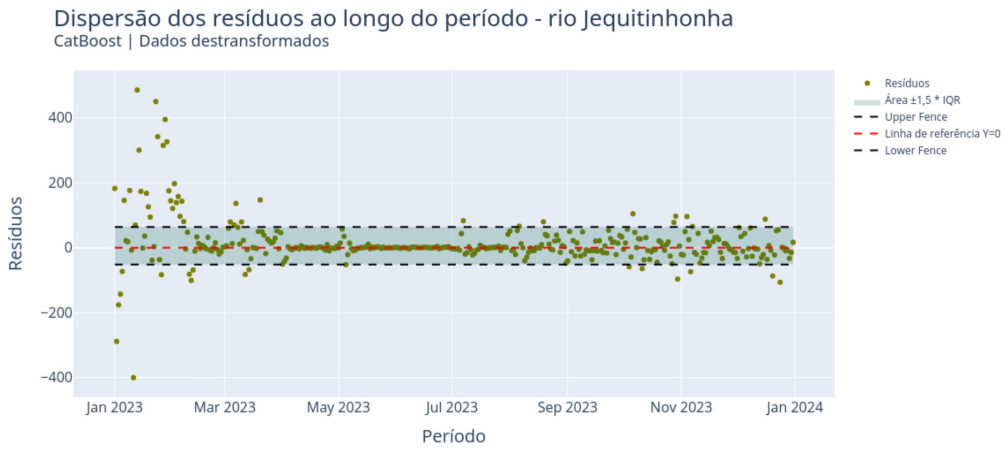
Cobertura empírica do intervalo de previsão
80.82 %

Fonte: O autor.

Os resíduos em modelos não-lineares podem ser mais difíceis de interpretar porque esses modelos capturam interações e padrões complexos. Porém, mesmo assim, detectar padrões sistemáticos nos resíduos ajuda a elucidar se os modelos conseguiram capturar todas as nuances dos dados.

A dispersão ao longo do tempo foi praticamente a mesma para ambos os modelos, com resíduos de valor elevado mais presentes no início da série – figuras 4.9 e 4.10. Aqui vale a interpretação usada para o modelo de referência, de que os valores elevados aferidos em final de 2022 possam ter interferido. Houve uma prevalência de $84,11\%$ e $86,03\%$, respectivamente CB e RF, dos resíduos na área sombreada, correspondente a $\pm 1,5 \times IQR$, que é onde se espera que todos os resíduos estejam – *upper fence* e *lower fence* do *boxplot*. Uma leve melhor prevalência para o modelo RF, que pode ser vista também em menos dispersão de resíduos para fora dos limites da área sombreada no início do ano.

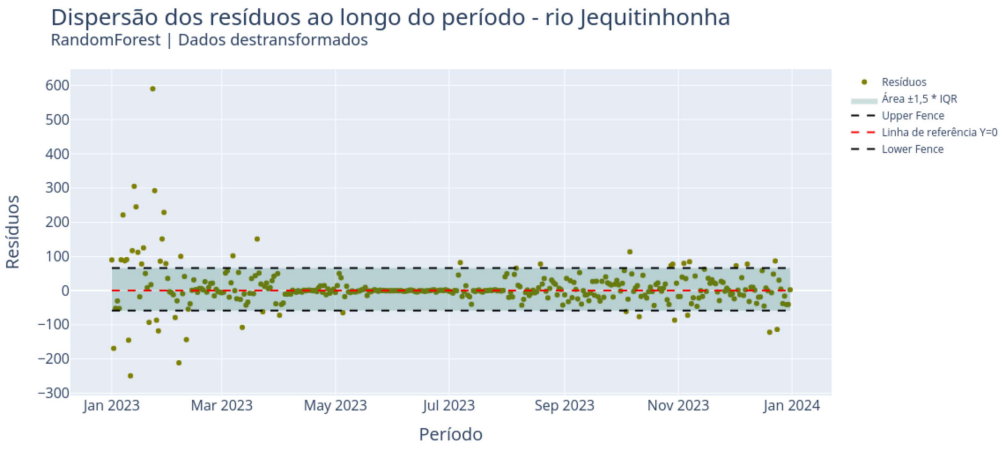
Figura 4.9 – Dispersão dos resíduos ao longo do ano.



Fonte: O autor.

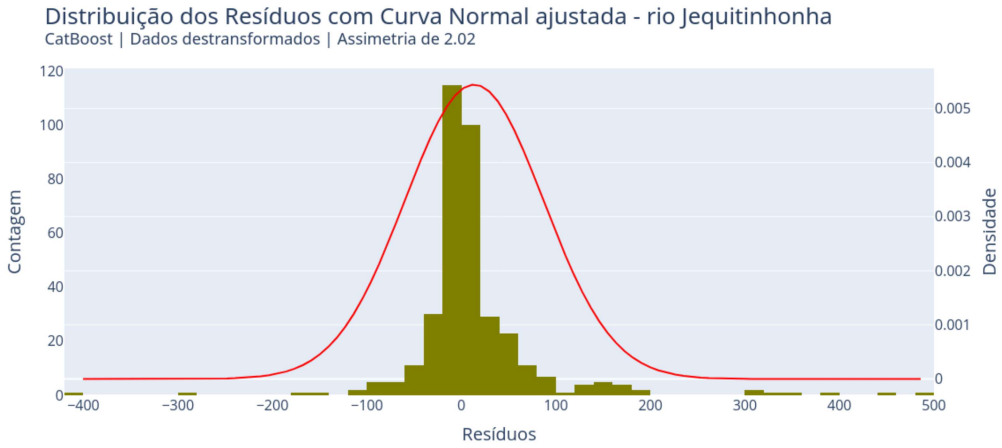
Em termos de assimetria da distribuição dos resíduos, ambos modelos apresentaram valores elevados, com assimetria positiva indicando cauda à direita. Com valores de $2,02$ para o modelo CB e $3,05$ para o RF, isso mostra tendência de subestimar as previsões, como visto anteriormente pelo PBIAS de ambos os modelos.

Figura 4.10 – Dispersão dos resíduos ao longo do ano.



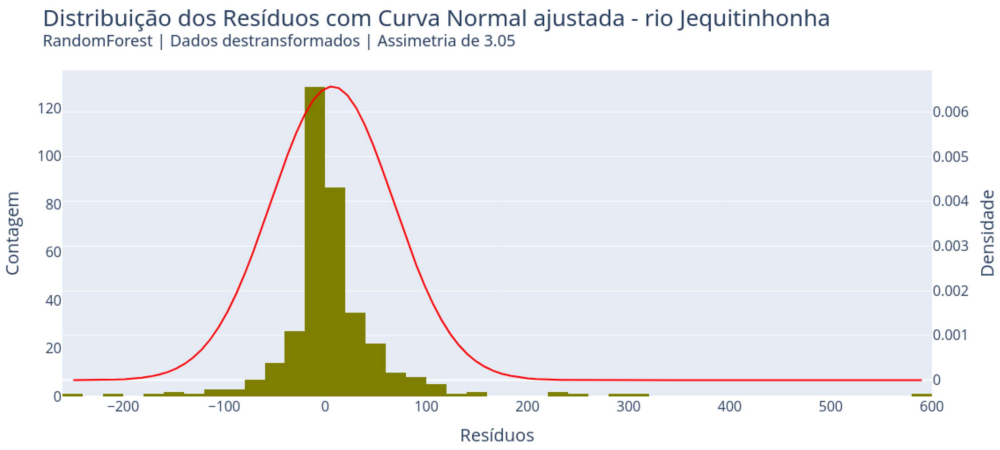
Fonte: O autor.

Figura 4.11 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

Figura 4.12 – Histograma e curva-normal dos resíduos.

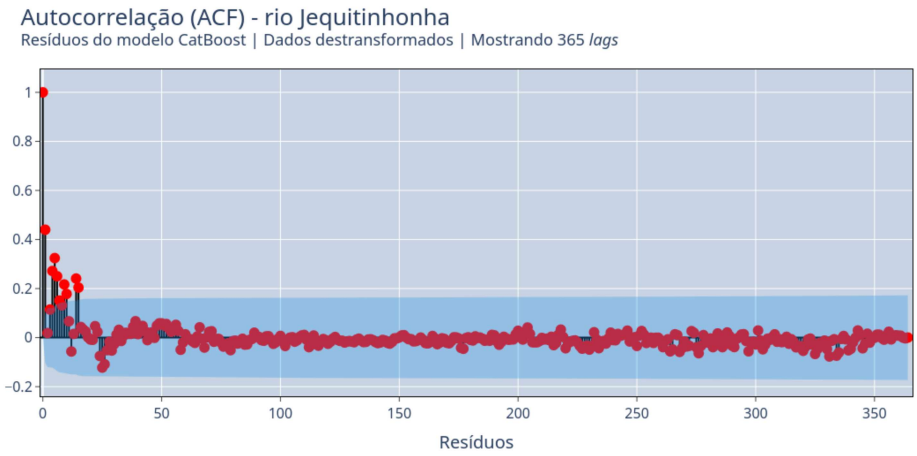


Fonte: O autor.

Observando os gráficos de autocorrelação dos modelos, houve um comportamento parecido nas *lags* iniciais. Nas primeiras *lags*, os gráficos mostram pontos de autocorrelação que ultrapassam a faixa de confiança – sombreada em azul –, indicando que os resíduos possuem dependência temporal significativa nos primeiros períodos. Esse padrão sugere que os modelos não conseguiram capturar totalmente a estrutura temporal dos dados, resultando em resíduos correlacionados em curtos períodos. Uma possível explicação pode ser a presença de componentes sazonais ou padrões de curto prazo que não foram completamente modelados e capturados pelos modelos.

A partir de aproximadamente a *lag* 20, a autocorrelação dos resíduos cai para valores próximos de 0 e permanece dentro da faixa de confiança para as *lags* seguintes até a *lag* 365. Este é um comportamento esperado para um bom modelo, onde os resíduos devem ser independentes e distribuídos aleatoriamente ao longo do tempo. A presença de resíduos com baixa autocorrelação em *lags* mais próximas – ao início – indica que, em períodos mais longos, os modelos não apresentam problemas de dependência temporal.

Figura 4.13 – Gráfico ACF dos resíduos.

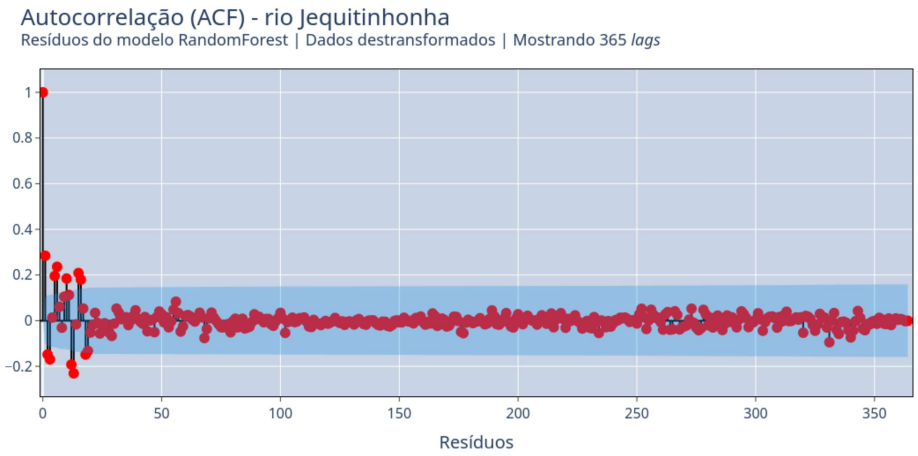


Fonte: O autor.

Agora os resultados sem emprego de log-transformação, com os dados originais. Salientando que para o modelo LR não é correto aplicar os dados originais na escala em que se encontram, foi preciso realizar normalização antes empregando algoritmo MinMax. Este algoritmo coloca todos os dados em valores entre 0 e 1, sendo o valor mínimo correspondente de cada série temporal passado para 0 e o valor máximo, consequentemente, é passado para 1. Com os demais valores é feito, basicamente, uma regra de três para achar o valor correspondente na escala entre 0 e 1.

Dito isso, o resultado para os dados originais ficou levemente inferior em relação ao modelo com os dados log-transformados. Em todas as métricas que se avaliar o resultado ficou piorado. Um comportamento destacado foi as faixas nos valores dos intervalos de previsão. Diferentemente do comportamento anterior – figura 4.2 –, em que houve uma

Figura 4.14 – Gráfico ACF dos resíduos.



Fonte: O autor.

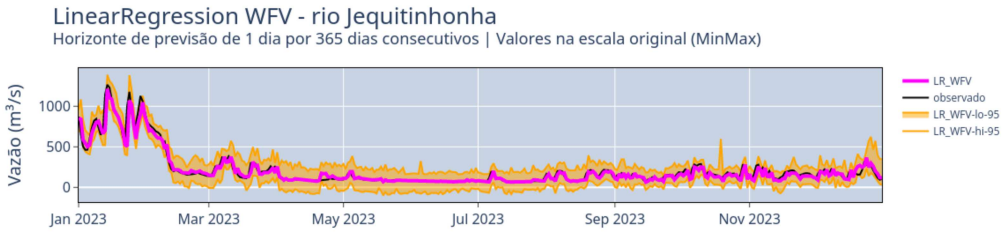
prevalência de faixas elevadas no início do ano, neste resultado o início do ano esteve, digamos, bem alinhado. Ao decorrer do ano que os intervalos incrementaram, a partir do mês de fevereiro, e foram aumentando até o fim do ano. Isso pode ser um indicativo de que o modelo teve problemas na estabilidade de longo prazo, acumulando as incertezas no período.

Ainda que tenha apresentado um KGE inferior, bem como nas demais métricas, quando se observa o comportamento dos resíduos, houve uma prevalência de $91,78\%$ dos resíduos na área sombreada no gráfico 4.16. A faixa entre os valores mínimo e máximo da faixa aumentaram em relação ao gráfico apresentado anteriormente. Pode-se verificar que no início do ano menos *outliers* estão presentes, ainda que na porção final do ano tenha aparecido alguns a mais que não foram vistos na análise anterior – figura 4.4. Para os dados originais, pela análise dos resíduos, o modelo mostrou boa estabilidade nas previsões pontuais. Quando se considera os intervalos de previsão, precisa considerar com cuidado, visto que o incremento dos valores superiores (hi-95) e valores de vazão $0m^3/s$ na faixa inferior (lo-95) indicam instabilidade de longo prazo.

Caminhando para o fim, o histograma e curva-normal apresentaram assimetria consideravelmente superior (valor de $3,37$) ao resultado de antes, com presença de cauda longa à direita – figura 4.17. O modelo teve, desta forma, tendência de subestimar os resultados, o que é verificável pelo PBIAS. Na figura 4.18 é possível perceber picos fora do intervalo de confiança. Após cerca de 10 *lags*, os pontos de autocorrelação ficam dentro desta faixa, o que significa que a maioria dos resíduos após esse ponto não está significativamente correlacionada com valores anteriores. Estes pontos vermelhos no início do gráfico podem indicar que há correlações significativas entre os resíduos nas *lags* mais próximas, o que sugere, nos primeiros períodos, que os resíduos estão correlacionados com os valores anteriores. Isso é indicativo, novamente, de que o modelo pode não estar capturando

totalmente a estrutura temporal dos dados e ser necessário ajustá-lo, adicionar variáveis que expliquem essa dependência temporal, ainda que variáveis de valor acumulado tenham sido inseridas exatamente na intenção de capturar tais comportamentos. Contudo, claramente precisaria aprimorar.

Figura 4.15 – *Walk-Forward Validation* para o modelo Regressão Linear.

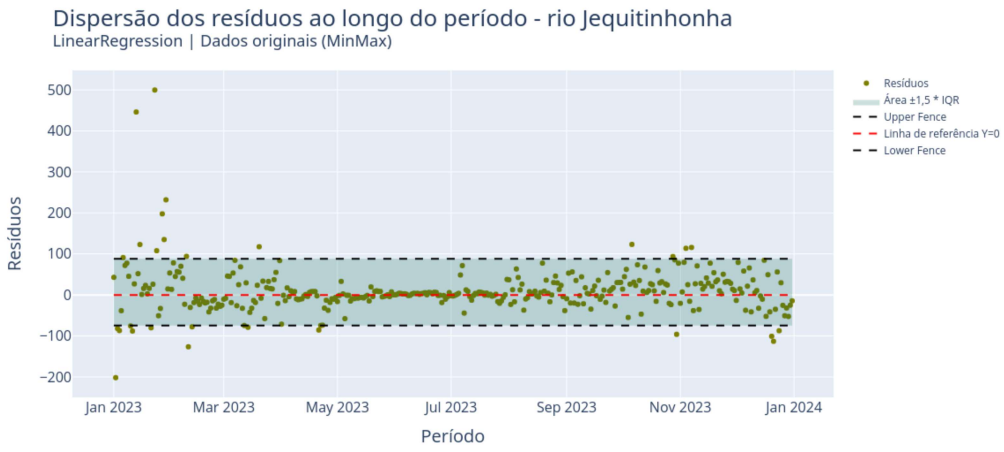


Modelo	MAPE	KGE	PBIAS (%)
LinearRegression	0.15144	0.91575	-4.30195
LinearRegression-lo-95	0.93932	0.3984	-59.79748
LinearRegression-hi-95	0.97542	0.41307	58.37905

Cobertura empírica do intervalo de previsão
96.99 %

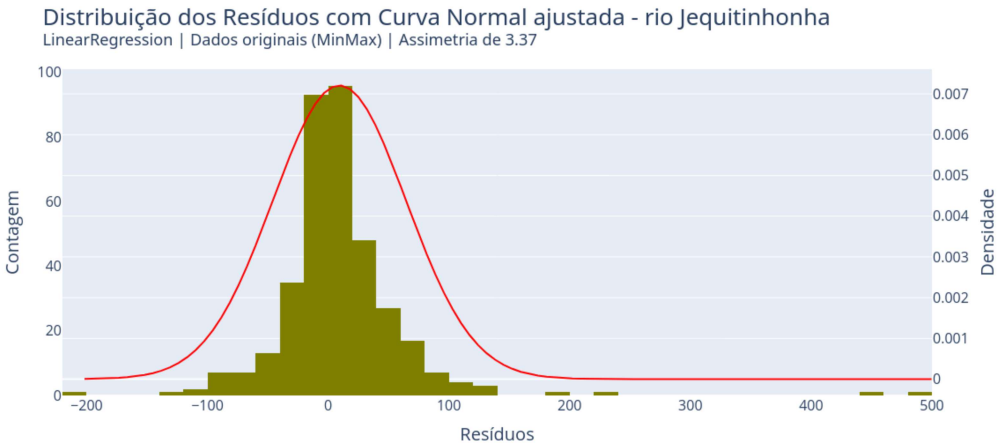
Fonte: O autor.

Figura 4.16 – Dispersão dos resíduos ao longo do ano.



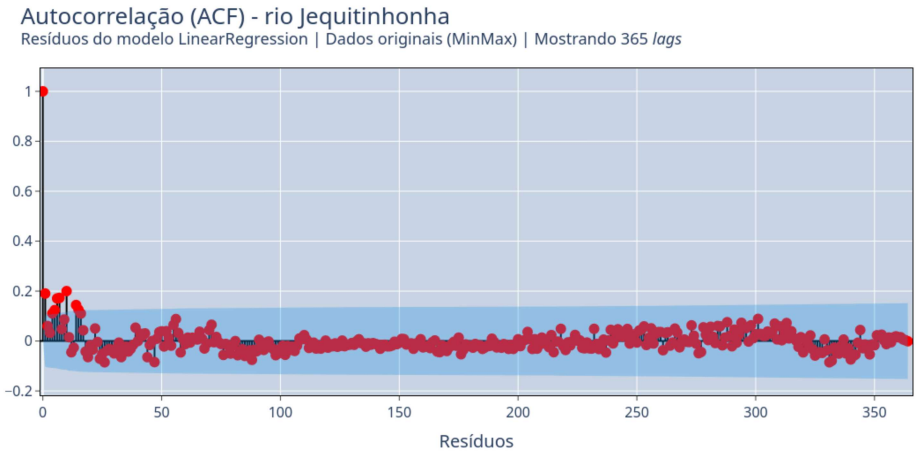
Fonte: O autor.

Figura 4.17 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

Figura 4.18 – Gráfico ACF dos resíduos.

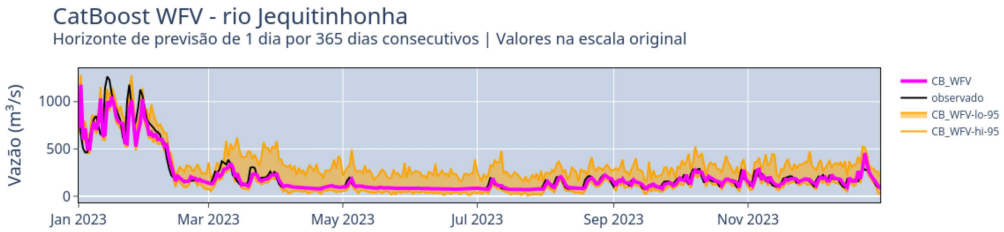


Fonte: O autor.

Relembrando que a log-transformação lineariza os dados e para o modelo LR ela realmente resultou em melhora dos resultados. Esperava-se que o mesmo pudesse, eventualmente, ocorrer com estes modelos não-lineares, pois simplificaria as relações entre as variáveis. Contudo, aconteceu o oposto. Permanecer com os dados originais, sem nenhum tipo de transformação, foi o que fez ambos, CB e RF, se comportarem melhor.

Veja o comportamento após WFV. Houve melhoras em quase todas as métricas nos dois modelos, exceto para o erro médio percentual, em que ambos tiveram leve piora – $0,133$ para o modelo CB (antes foi $0,130$) e $0,14$ para o RF ($0,12$ antes). É ínfimo, porém não descartável. Mas o KGE, uma métrica mais robusta, melhorou bastante entre o resultado anterior e este resultado em tela. Aplicando os modelos CB e RF nos dados originais, melhorou também o viés sistemático, PBIAS. O modelo CB continuou mostrando tendência de subestimar as previsões, mas aqui o modelo RF inverteu, passando a apresentar tendência de superestimar as previsões. A cobertura empírica dos intervalos de previsão melhoraram, sendo o CB quase $2,5\%$ melhor, e o RF melhorou 3% nos cálculos dos intervalos. A respeito dos intervalos de previsão, os modelos mostraram tendência de aumento das faixas ao longo do ano. Diferentemente da análise log-transformada, aqui os modelos parecem ter demostrado incerteza quanto aos erros calculados em eventos anteriores, ainda que no início do ano, para ambos, pareça ter ficado mais estável o cálculo dos intervalos.

Figura 4.19 – *Walk-Forward Validation* para o modelo *CatBoost*.

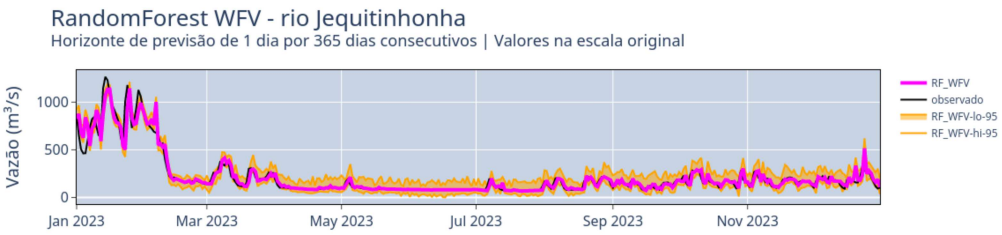


Modelo	MAPE	KGE	PBIAS (%)
CatBoost	0.13346	0.88914	-3.77912
CatBoost-lo-95	0.3677	0.73184	-24.66957
CatBoost-hi-95	1.23007	0.29189	69.72264

Cobertura empírica do intervalo de previsão
89.32 %

Fonte: O autor.

Figura 4.20 – *Walk-Forward Validation* para o modelo *Random Forest*.



Modelo	MAPE	KGE	PBIAS (%)
RandomForest	0.14003	0.95193	1.24504
RandomForest-lo-95	0.36217	0.77689	-21.59261
RandomForest-hi-95	0.71244	0.5863	40.85158

Cobertura empírica do intervalo de previsão
83.84 %

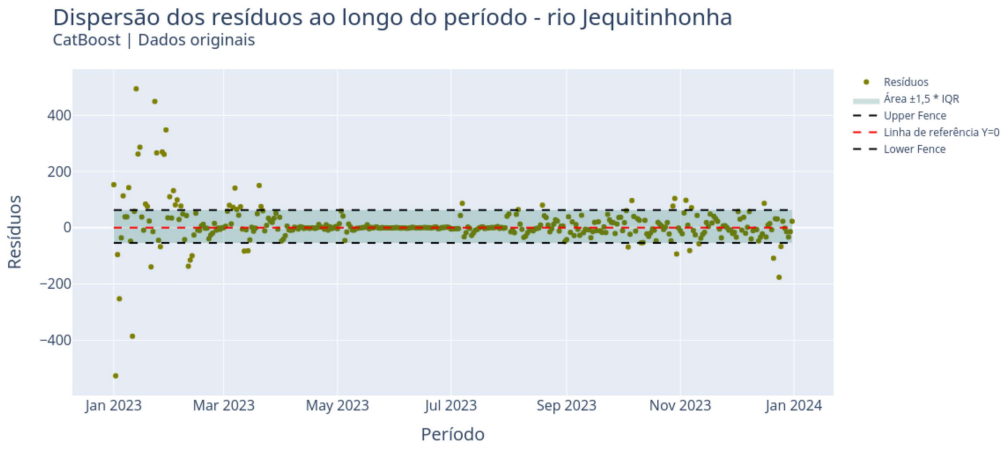
Fonte: O autor.

A maioria dos resíduos está concentrada em torno da linha de referência $y = 0$, indicando que os modelos mantiveram um desempenho geral estável ao longo do tempo. Porém, há *outliers* significativos no início de 2023 – valores de resíduos acima de 400 e abaixo de -400 para o CB e acima de 500 e abaixo de -300 para o RF. Esses *outliers* ocorreram em períodos de vazões mais elevadas durante o início do ano, momentos de maiores picos de vazão, indicando que ambos tiveram dificuldades em prever alguns eventos específicos, resultando em erros grandes.

Para o modelo CB, após fevereiro de 2023, e março para o modelo RF, os resíduos parecem estar melhor distribuídos dentro da área delimitada pelas faixas, com menos *outliers*, sugerindo que os modelos ajustaram-se melhor aos dados ao longo do tempo e permaneceram estáveis.

Em geral, ambos os modelos apresentaram um padrão de melhora ao longo do ano, com resíduos mais bem distribuídos após o primeiro trimestre de 2023. No entanto, os *outliers* iniciais indicam que os modelos podem ter dificuldades com eventos específicos de sazonalidade ou anomalias no início do ano. O CB parece ter uma leve vantagem em termos de previsões mais consistentes ao longo do tempo, com valores menos extremos. Houve uma prevalência de $85,75\%$ dos resíduos dentro da área sombreada do gráfico 4.21, ao passo que o modelo RF obteve $84,11\%$ de prevalência – gráfico 4.22. Ajustes de hiperparâmetros podem ajudar os modelos nos períodos mais desafiadores do ano, como no início do ano, ou adição de componentes de tendência e sazonalidade localizada.

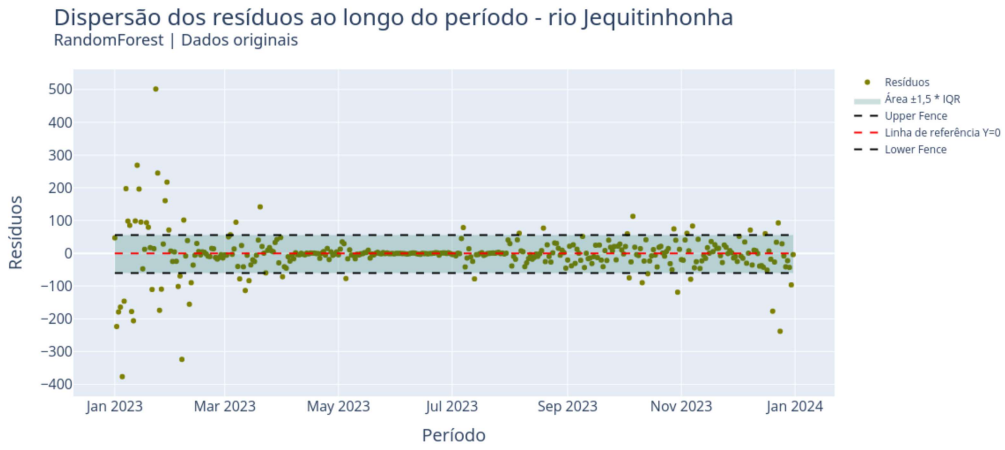
Figura 4.21 – Dispersão dos resíduos ao longo do ano.



Fonte: O autor.

A distribuição dos resíduos do modelo CB – figura 4.23 – apresenta uma assimetria positiva de $0,66$, o que indica que a cauda direita da distribuição é mais longa que a esquerda. A presença de resíduos elevados tanto na cauda direita – acima de 200 – quanto na cauda esquerda – abaixo de -100 – indica que o modelo teve dificuldades em prever com precisão alguns valores mais extremos. A curva normal ajustada em vermelho não

Figura 4.22 – Dispersão dos resíduos ao longo do ano.



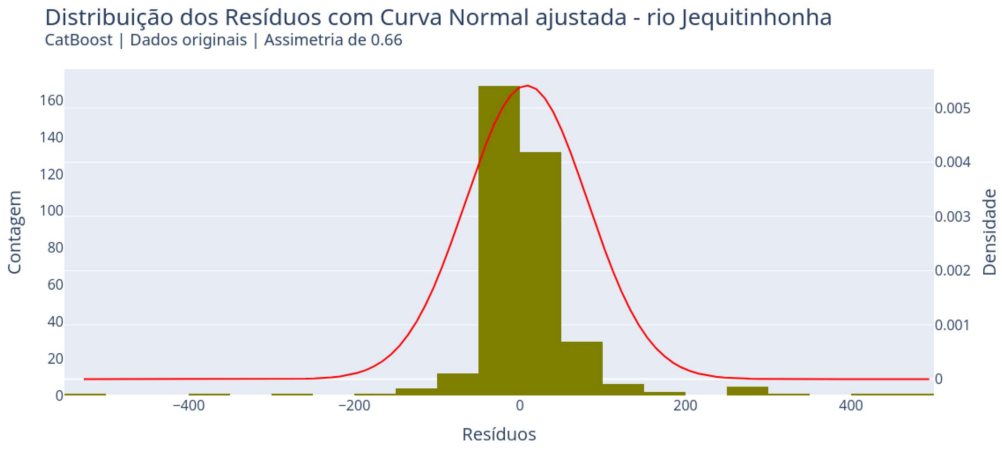
Fonte: O autor.

se adequa perfeitamente à distribuição observada, especialmente nas caudas, deixando claro que a distribuição dos resíduos não é perfeitamente normal. Esse comportamento é corroborado pelo valor da assimetria. Contudo, a maioria dos resíduos está concentrada em torno de $y = 0$, com uma contagem elevada próxima a 0. Isso é um bom sinal, pois sugere que o modelo fez previsões razoavelmente precisas para grande parte das observações.

A distribuição dos resíduos do modelo RF – figura 4.24 – apresenta uma assimetria positiva de $0,49$, indicando que a cauda direita é um pouco mais longa, mas a assimetria é menos pronunciada do que no modelo CB. A curva normal ajustada se aproxima mais da distribuição observada do que no gráfico do CB, o que quer dizer que os resíduos do RF seguem uma distribuição mais próxima de uma distribuição normal, apesar da pequena assimetria positiva. Assim como no gráfico do CB, a parcela maior dos resíduos está concentrada em torno de 0, com a maioria das previsões sendo razoavelmente precisas, porém, o modelo RF apresenta um adensamento ligeiramente mais equilibrado.

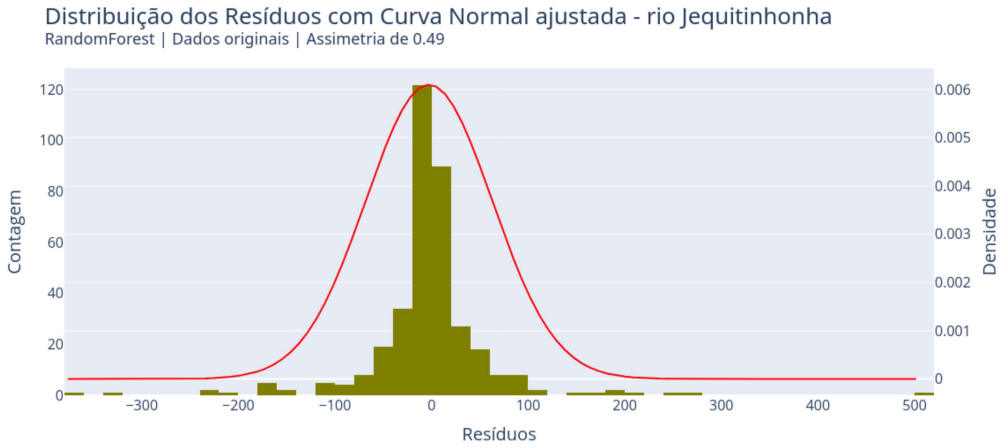
O RF tem um comportamento de resíduos mais próximo de uma distribuição normal, caracterizando-o como mais estável e robusto em termos de previsões para a maioria dos dados. O CB mostrou sensibilidade a eventos extremos, com uma assimetria maior, indicando que o modelo enfrenta dificuldades em capturar corretamente estes valores mais extremos, resultando em erros maiores. Mas, em geral, ambos os modelos apresentaram uma boa concentração de resíduos em torno de 0 – praticamente indistinguível –, mostrando que a maioria das previsões está correta.

Figura 4.23 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

Figura 4.24 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

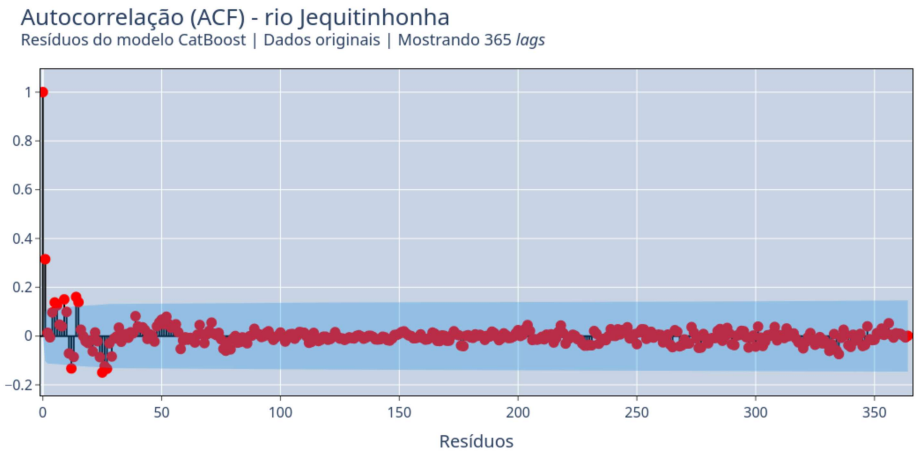
Nas primeiras *lags*, há autocorrelações significativas que ultrapassam a faixa de confiança, especialmente nas primeiras 15 *lags*. Para curtos intervalos de tempo, os resíduos do modelo CB estão correlacionados – figura 4.25. A presença dessas correlações positivas nas primeiras *lags* pode indicar que há padrões temporais de curto prazo nos dados que o modelo CB não capturou adequadamente. No entanto, após essa *lag* 15, as autocorrelações caem para valores dentro da faixa de confiança de 95% e permanecem próximas de 0. Isso mostra que, para intervalos maiores, os resíduos não apresentam correlação significativa, o que é um sinal positivo de que, em termos de longo prazo, o modelo está capturando a variabilidade dos dados de forma satisfatória.

Assim como no gráfico do CB, o modelo RF também apresenta autocorrelações significativas nas primeiras *lags*, até perto de 20 *lags* – figura 4.26. Houve um comportamento levemente pior aqui – levou mais 5 *lags* para estabilizar – neste modelo do que em relação ao CB, indicando, semelhantemente, uma dependência temporal de curto prazo que o modelo RF não capturou completamente, resultando em resíduos correlacionados nas primeiras *lags*. A partir da *lag* 20, as autocorrelações do RF caem para dentro da faixa de confiança e permanecem próximas de 0, indicando que, para previsões de longo prazo, o modelo RF também está capturando bem a estrutura dos dados e os resíduos se comportam de maneira aleatória.

Ambos os modelos CB e RF apresentam autocorrelações significativas nas primeiras *lags*, denotando que os modelos apresentaram problemas com a dependência temporal de curto prazo nos dados. No entanto, o CB teve uma presença um pouco menor de correlações significativas em comparação com o RF, o que indica um desempenho ligeiramente melhor no curto prazo.

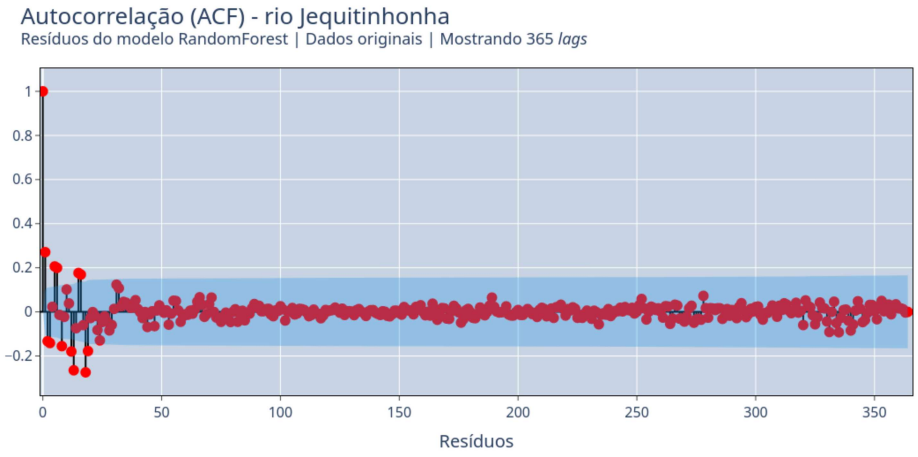
Para *lags* maiores (acima de 20), tanto o CB quanto o RF apresentaram resíduos que se comportaram de maneira aleatória, com autocorrelações dentro da faixa de confiança e próximas de 0. Considerando as previsões de longo prazo, ambos os modelos estão funcionando de forma satisfatória e não apresentaram padrões remanescentes nos resíduos.

Figura 4.25 – Gráfico ACF dos resíduos.



Fonte: O autor.

Figura 4.26 – Gráfico ACF dos resíduos.



Fonte: O autor.

Para o rio Jequitinhonha a análise termina aqui. Aplicar a log-transformação melhorou o resultado do modelo LR de quando não foi aplicada essa transformação, e os modelos CB e RF (não-lineares) foram melhores quando os dados não foram transformados. Aqui, para este rio, foi este o resultado, mais adiante será visto para os demais rios.

Uma coisa que precisa ficar claro aqui: só serão discutidos para os outros rios os melhores resultados. Para o rio Jequitinhonha foi mostrada toda linha de trabalho para que ficasse claro que toda comparação foi realizada entre resultados log-transformados e não transformados, tudo foi estudado e analisado. Mas pelo bem da concisão – e deve ter ficado evidente as voltas dadas no texto mostrando exatamente cada resultado – todo resultado não apresentado aqui é porque não foram considerados bons o suficiente. E não será mais incluída a análise para o modelo SN.

Adiante.

4.1.2 Rio Doce

Os resultados para o rio Doce apresentaram-se ótimos, considerando-se todas as métricas do trabalho. Isso possivelmente se deve à qualidade dos dados de treinamento, em que não foi preciso preencher tantos dados faltantes de vazão. Relembre: de 4017 registros, apenas 3 faltavam.

O modelo CB apresentou um desempenho bom, com KGE de $0,89$ sugerindo alta correlação entre os valores previstos e observados. Quando se considera o MAPE, foi o mais baixo dentre os três modelos, contudo, mesmo assim, indica uma precisão elevada nas previsões médias, o que demonstra robustez do modelo. O PBIAS de -3.73% mostra um leve viés geral para subestimar as previsões, e a cobertura empírica de $91,23\%$ indica que o intervalo de previsão capturou bem a variabilidade dos dados observados.

O gráfico 4.27 revela ainda que o modelo acompanha bem os picos de vazão e que os intervalos de previsão foram particularmente estreitos quando no período de vazões elevadas e relativamente amplos nos períodos de menores vazões, podendo indicar que as medições de precipitação nos períodos chuvosos conferiram leve melhoramento na precisão do comportamento médio do modelo. Em outras palavras: o modelo fica "otimista" quando os dados de chuva não estão zerados e possui mais informações, o que ocorre nos períodos de outono e inverno. Tanto que nestes períodos mencionados, quando há uma profusão de valores 0 na precipitação, os intervalos de previsão ficaram amplos, o que denota a tentativa do modelo de "acertar" a medida. Comportamento este que o modelo RF também apresentou em relação aos intervalos de previsão – visto em 4.29. A cobertura empírica dos intervalos de previsão de ambos os modelos ficaram próximas, com uma leve vantagem para o modelo CB – o modelo RF obteve $88,77\%$.

O modelo LR, mesmo sendo um modelo simples, apresentou um desempenho notável, com MAPE ligeiramente maior, de $0,084$, do que a do CB, mas ainda assim baixo.

O KGE foi o maior dos três modelos, alcançando $0,94$, mostrando uma ótima correlação e menor erro de variância entre os valores observados e os previstos. O PBIAS de $-2,75\%$ sugere um viés de subestimação, sendo o "caminho do meio" entre os modelos. A cobertura empírica de $96,71\%$ mostra que o intervalo de previsão captura quase a totalidade da variabilidade dos dados observados – visível no gráfico 4.28. O gráfico também mostra que o LR também consegue capturar os picos de vazão de forma eficaz, com intervalos de previsão mais estreitos ao longo do ano se comparados ao CB e ao RF, porém mais amplos no período inicial do ano.

O modelo RF também teve um bom desempenho, com um MAPE intermediário entre os outros dois modelos e um KGE de $0,93$, ficando muito próximo do LR. O PBIAS de $-1,44\%$ indica o menor viés de subestimação entre os três modelos, no entanto, a cobertura empírica foi a menor, com $88,77\%$, sugerindo que os intervalos de previsão não abrangem tanto a variabilidade observada quanto nos outros modelos. Isso é perceptível principalmente no início da série, nos meses de janeiro, fevereiro e março, ainda que tenha captado bem os picos de vazão, um comportamento semelhante ao CB em termos de previsões.

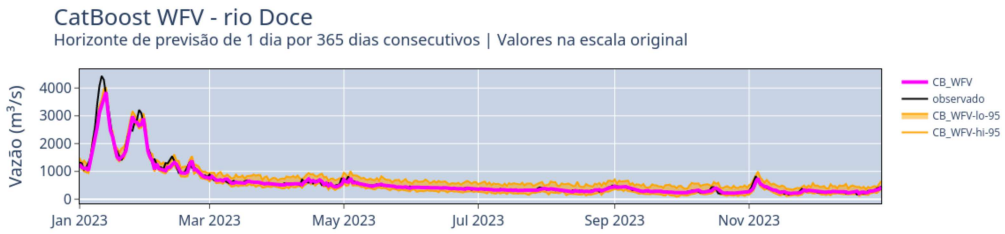
Em termos de precisão média, o CB apresentou o melhor desempenho – valor de $0,082$ -, seguido de perto por RF – valor de $0,083$ – e LR – MAPE de $0,084$. A diferença foi muito pequena, na ordem de 10^{-3}, mostrando que todos os modelos tiveram alta precisão. Analisando o KGE – a mais complexa das métricas –, o modelo LR foi o melhor nesse critério, seguido por RF e CB. Isso sugere que, apesar de ser um modelo linear, a regressão conseguiu capturar bem a dinâmica não-linear das vazões do rio. Mas, novamente, os valores ficaram muito próximos, não havendo grande discrepância entre as medições.

No gráfico 4.30 é possível observar que no início do período – janeiro e fevereiro de 2023 –, os resíduos estão mais dispersos, com muitos valores além dos limites, tanto na parte superior quanto na inferior. Isso denota que o modelo teve dificuldades em prever corretamente durante essa fase inicial do ano, que pode estar associada a picos de vazão ou condições hidrológicas atípicas. Após março de 2023, os resíduos começam a se concentrar dentro dos intervalos, de onde se conclui que o modelo se estabilizou e começou a prever com mais precisão, com menos erros extremos, ou seja, o CB foi capaz de capturar bem o padrão de vazões ao longo do ano, exceto em eventos de pico no início.

Da mesma forma que o CB, o modelo LR mostra uma maior dispersão dos resíduos no início do período, particularmente nos meses de janeiro e fevereiro de 2023 – figura 4.31. A partir de março de 2023 os resíduos do LR também se tornam mais estáveis e permanecem concentrados dentro da área sombreada, indicando uma melhor desempenho com o passar do tempo e em períodos de amenidades nas vazões.

E, finalmente, o modelo RF – figura 4.32. O comportamento dos resíduos do RF ao longo do tempo é similar ao dos outros dois modelos anteriores. Há uma maior dispersão

Figura 4.27 – *Walk-Forward Validation* para o modelo *CatBoost*.



Modelo	MAPE	KGE	PBIAS (%)
CatBoost	0.08207	0.8933	-3.73972
CatBoost-lo-95	0.26334	0.77828	-19.96104
CatBoost-hi-95	0.50795	0.69261	29.28936

Cobertura empírica do intervalo de previsão
91.23 %

Fonte: O autor.

Figura 4.28 – *Walk-Forward Validation* para o modelo *Linear Regression*.

Modelo	MAPE	KGE	PBIAS (%)
LinearRegression	0.08434	0.94579	-2.75287
LinearRegression-lo-95	0.23034	0.64556	-24.09201
LinearRegression-hi-95	0.27906	0.66321	24.97568

Cobertura empírica do intervalo de previsão
96.71 %

Fonte: O autor.

no início do ano (janeiro e fevereiro de 2023), mas a partir de março os resíduos se tornam mais estáveis, com a maioria dos valores dentro da área destacada. Indica que o RF também teve dificuldades em capturar os eventos extremos no início do ano e que também apresentou bom desempenho para o restante do período.

Figura 4.29 – *Walk-Forward Validation* para o modelo *Random Forest*.

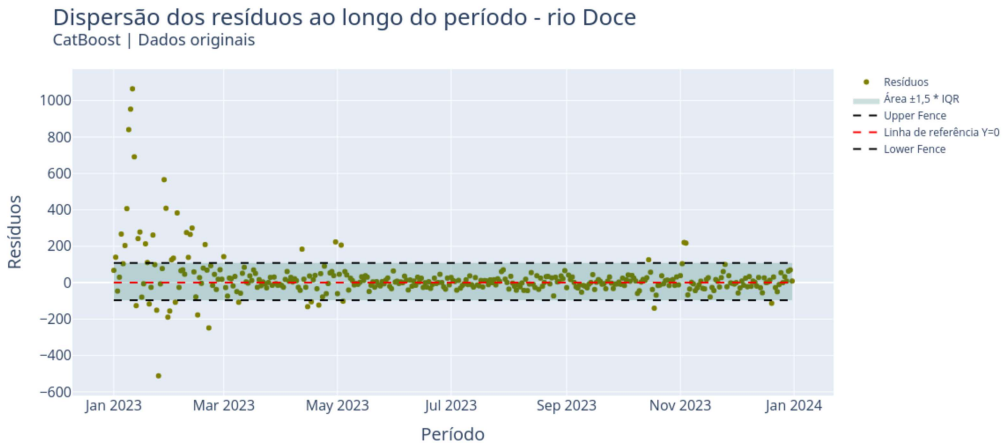
Modelo	MAPE	KGE	PBIAS (%)
RandomForest	0.08355	0.9375	-1.44227
RandomForest-lo-95	0.23089	0.82843	-16.07935
RandomForest-hi-95	0.43947	0.73666	24.84179

Cobertura empírica do intervalo de previsão
88.77 %

Fonte: O autor.

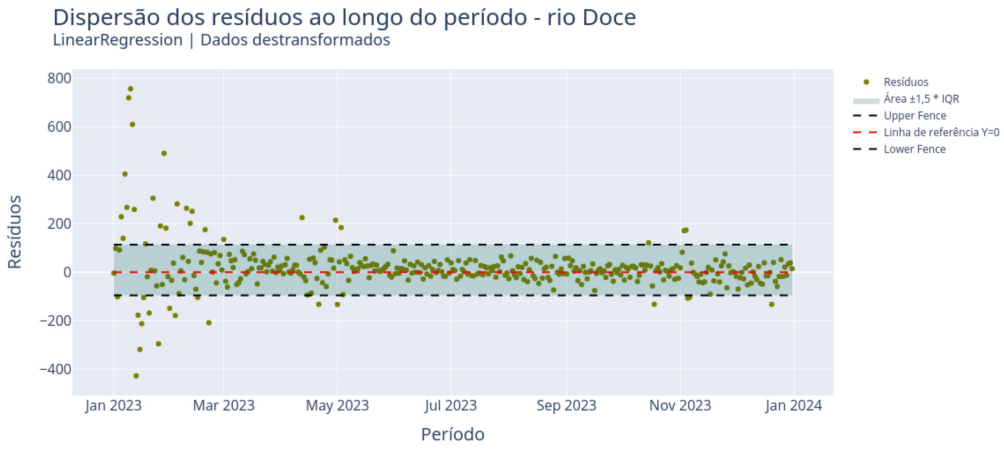
Os três modelos enfrentaram desafios semelhantes durante o período inicial do ano – janeiro e fevereiro –, comportamento quase indistinguível entre eles, mas todos estabilizaram a partir de março, sem que tenha um modelo que realmente se destaque frente aos demais. O comportamento dos resíduos ao longo do tempo indica que, para períodos com picos de vazão, pode ser necessário um refinamento adicional nos modelos ou mesmo refinamento nos dados de entrada para melhor capturar estes eventos.

Figura 4.30 – Dispersão dos resíduos ao longo do ano.



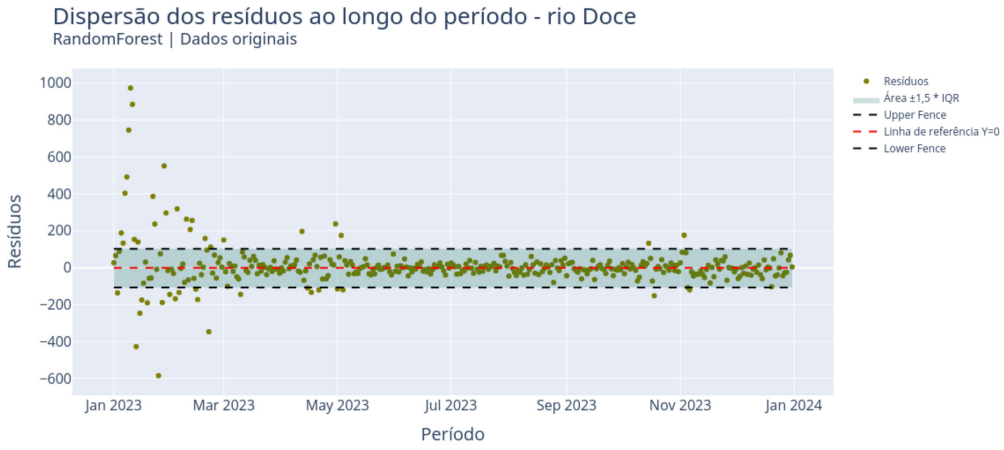
Fonte: O autor.

Figura 4.31 – Dispersão dos resíduos ao longo do ano.



Fonte: O autor.

Figura 4.32 – Dispersão dos resíduos ao longo do ano.

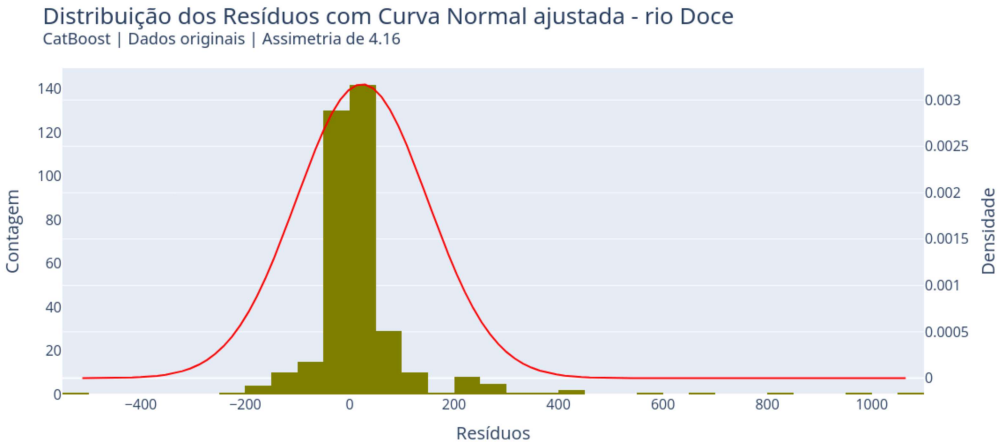


Fonte: O autor.

A assimetria dos resíduos para o modelo CB – figura 4.33 – é de $4,16$, o que indica uma distribuição altamente assimétrica à direita e positiva. Isso quer dizer que os erros positivos são mais frequentes ou mais intensos do que os erros negativos, o que combina com o viés sistemático do modelo em que o PBIAS indicou subestimação – visto em 4.27. Posto que o cálculo do resíduo é "*observado* − *previsto*" e trazendo o gráfico 4.30 à vista aqui novamente, percebe-se que esta cauda longa de valores positivos ocorreu principalmente no início do ano, exatamente quando as variações nas vazões foram mais intensas. Este fenômeno também ocorreu com os outros modelos – figuras 4.34 e 4.35. Contudo, para o modelo LR com dados log-transformados, a assimetria se mostrou menor – de $2,71$ – do que em relação aos modelos não-lineares, restando a assimetria de $3,24$ para o modelo RF. Nesta avaliação, o modelo CB apresentou a pior distribuição de resíduos, abrindo margem para melhorias. O mesmo, no entanto, vale para os modelos LR e RF que também apresentaram assimetria elevada e carecem de aprimoramentos.

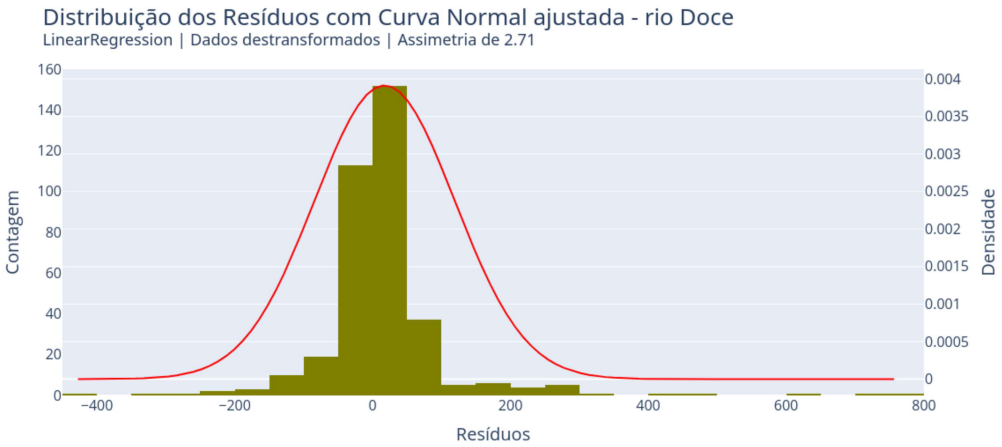
A curva normal está deslocada à direita, devido à assimetria dos dados, para todos os modelos, porém um pouco menos para o modelo LR. Todos os modelos possuem resíduos concentrados, majoritariamente, entre -200 e 200, contudo, para os modelos não-lineares CB e RF, a variação total dos resíduos oscila entre aproximadamente -550 e 1000, ao passo que o modelo LR, mostrando melhor comportamento, oscila entre -400 e aproximadamente 750.

Figura 4.33 – Histograma e curva-normal dos resíduos.



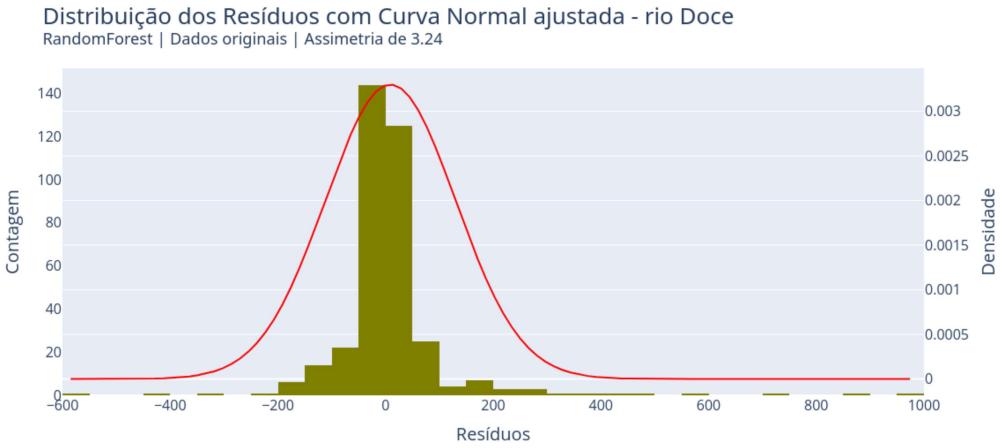
Fonte: O autor.

Figura 4.34 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

Figura 4.35 – Histograma e curva-normal dos resíduos.



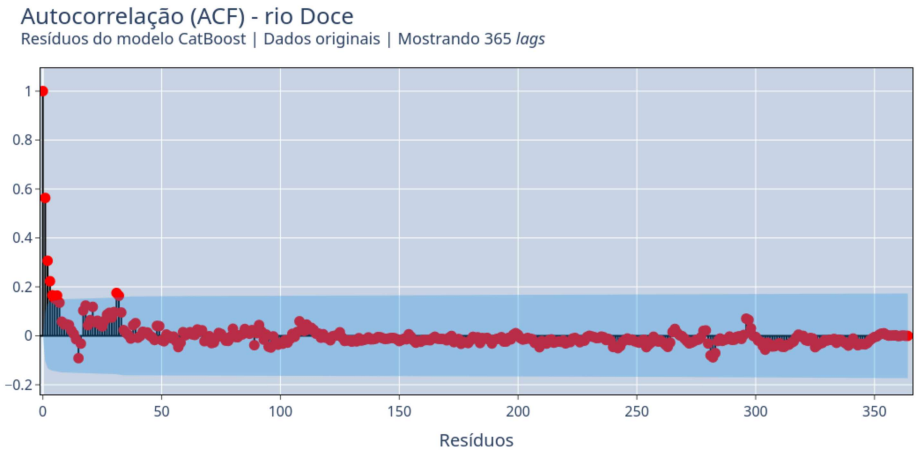
Fonte: O autor.

O CB apresenta autocorrelação significativa nas primeiras *lags* – até aproximadamente a *lag* 10 –, com picos consideráveis de correlação positiva, mostrando que o modelo está deixando dependências temporais nos resíduos – figura 4.36. Após as previsões imediatas, os resíduos ainda carregam dependências de eventos passados, o que não deveria ocorrer. O LR também mostra autocorrelação nas primeiras *lags*, repetindo o comportamento do CB, embora de menor magnitude quando comparado a este – figura 4.37. Ainda que o modelo LR também esteja com problemas nas dependências de curto prazo, ele o faz de forma um pouco diferente que o CB. Por sua vez, o RF apresenta picos de autocorrelação menores que CB, igualmente baixo quanto o LR – até a *lag* 2 – nas primeiras *lags* – figura 4.38. No entanto, ainda persiste alguma autocorrelação, com uma magnitude baixa, sugerindo que o RF está deixando passar para os resíduos uma dinâmica temporal.

Avançando no estudo, a autocorrelação do modelo CB diminui rapidamente após as primeiras *lags* e por volta da *lag* 10, a maioria dos resíduos já se estabiliza dentro do intervalo de confiança – a área azul esmaecida. Isso é um sinal de que, a partir de *lags* maiores, os resíduos são independentes, um comportamento esperado e positivo para previsões de longo prazo. Perceba, no entanto, que tanto o LR quanto o RF apresentam uma dissipação rápida da autocorrelação, diferentemente do CB, e seus resíduos se estabilizam rapidamente dentro do intervalo de confiança após as primeiras *lags*.

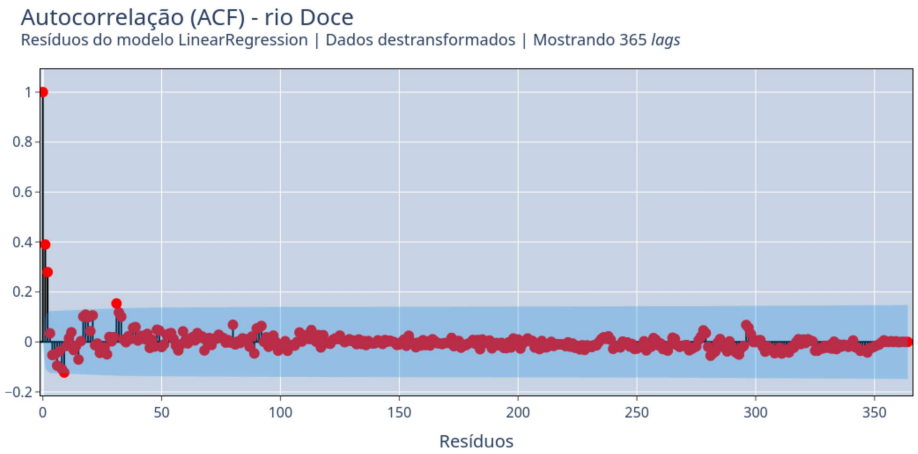
Todos os modelos mostraram alguma autocorrelação nos resíduos de curto prazo, indicando que nenhum deles conseguiu eliminar completamente as dependências temporais de curto prazo nas previsões. Esse fenômeno é mais acentuado no CB, seguido por LR e RF, que apresentaram a menor autocorrelação nas primeiras *lags*.

Figura 4.36 – Gráfico ACF dos resíduos.



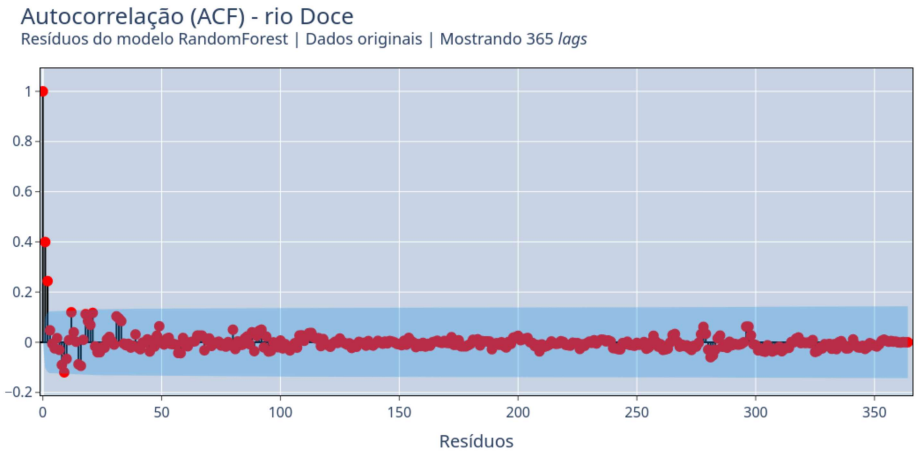
Fonte: O autor.

Figura 4.37 – Gráfico ACF dos resíduos.



Fonte: O autor.

Figura 4.38 – Gráfico ACF dos resíduos.



Fonte: O autor.

Toda essa análise demonstrou aspectos importantes acerca dos modelos. Todos apresentaram uma tendência de subestimar os picos, com o modelo LR tendo uma oscilação levemente menor – de -400 a 750 –, o período do ano que foi complicado para os modelos foi o período inicial do ano, verão, quando as vazões são fortemente impactadas pelas oscilações nas precipitações e o gráfico ACF deixou claro que os resíduos carregam dependência temporal de curto prazo. Isso sugere a necessidade de inclusão de mais variáveis explicativas para lidar com esses eventos extremos, refinamento nos hiperparâmetros – para os modelos não-lineares –, ou mesmo de refinamento nas técnicas de modelagem – isso não pode ser descartado.

4.1.3 Rio Grande

O rio Grande foi o único que a análise foi feita em uma massa reduzida de dados. Conforme já detalhado, apenas dados de setembro de 2020 até dezembro de 2023. Feita esta ressalva, os resultados mostraram-se bons e os modelos foram estáveis quando da análise dos resíduos. Perceba que aqui, para o modelo LR, os melhores resultados foram com os dados escalados MinMax e não os dados log-transformados.

Iniciando pelo KGE, o RF se saiu melhor, seguido próximo pelo LR – figuras 4.41 e 4.40, respectivamente. Ambos tiveram valores superiores ao CB – figura 4.39 –, indicando que capturaram melhor correlação, viés e variabilidade nos dados previstos de vazão ao longo do período.

Sobre o erro médio – MAPE –, o RF e o CB apresentaram os menores valores e próximos entre si, ou seja, ambos tiveram comportamento semelhante – valores de $0,123$ e $0,126$, respectivamente. O modelo LR ficou ligeiramente ruim, $0,133$, mas de toda forma, isso os coloca com resultados muito próximos, sugerindo que todos os três modelos ofereceram previsões de boa qualidade.

O LR apresentou o menor PBIAS, apenas $-1,99\%$, ou seja, suas previsões foram mais equilibradas, com menos tendência neste caso de subestimar os valores de vazão. Os modelos CB e RF também têm PBIAS relativamente baixos – respectivamente $-4,44\%$ e $-2,18\%$ –, mas o CB apresentou um valor negativo pior, deixando clara uma subestimação elevada comparada aos outros modelos.

A cobertura empírica do LR é significativamente maior do que a dos outros dois modelos, com $84,38\%$, mostrando que o intervalo de previsão capturou com mais precisão a variabilidade real presente nas observações. Tanto o CB ($63,56\%$) quanto o RF ($63,01\%$) têm valores de cobertura mais baixos, que mostram, por sua vez, intervalos de previsão foram mais estreitos e não conseguiram captar toda a variabilidade observada nos dados reais.

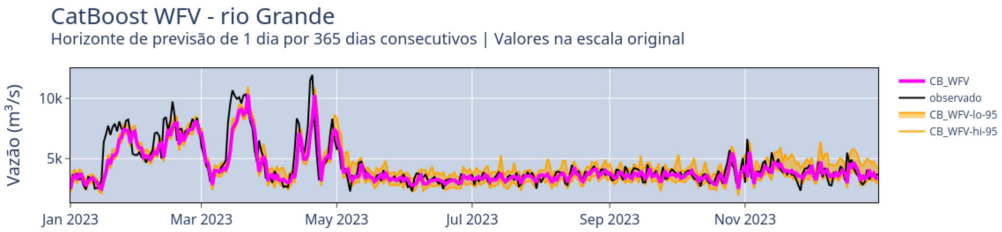
Em ambos os modelos não-lineares a previsão segue de perto a série observada, mas a subestimação das vazões mais altas é evidente. O intervalo de previsão é estreito em

momentos de pico, o que contribui para a baixa cobertura empírica calculada. Os maiores desvios nas previsões ocorrem em períodos de alta vazão, por volta de março e maio de 2023, onde os modelos apresentaram dificuldade em acompanhar as flutuações nas vazões.

O modelo LR consegue capturar razoavelmente bem as flutuações de vazão, mas também apresenta comportamento tendencioso de subestimar os picos mais extremos, da mesma forma que os modelos CB e RF. Contudo, o intervalo de previsão foi o mais amplo dos três modelos, resultando em uma cobertura empírica melhor.

O RF foi o modelo que obteve o melhor equilíbrio entre precisão (MAPE), correlação/variabilidade (KGE) e viés (PBIAS), mostrando um desempenho consistente. No entanto, a cobertura empírica baixa sugere que o intervalo de previsão não captura toda a incerteza associada aos dados, é uma característica que pode melhorar. O LR, embora tenha uma MAPE ligeiramente maior, apresentou o melhor valor de PBIAS, denotando que ele tem a menor tendência a viés sistemático entre os modelos. Além disso, sua maior cobertura empírica sugere que o modelo captura melhor a variabilidade dos dados. Finalmente, o CB, ainda que seja um modelo robusto, mostrou desempenho inferior em termos de KGE e PBIAS, mostrando dificuldades em capturar corretamente as variações sazonais e os picos extremos de vazão. Contudo, vale destacar, este é um modelo altamente personalizável e um procedimento de otimização dos hiperparâmetros pode aprimorar os resultados.

Figura 4.39 – *Walk-Forward Validation* para o modelo *CatBoost*.

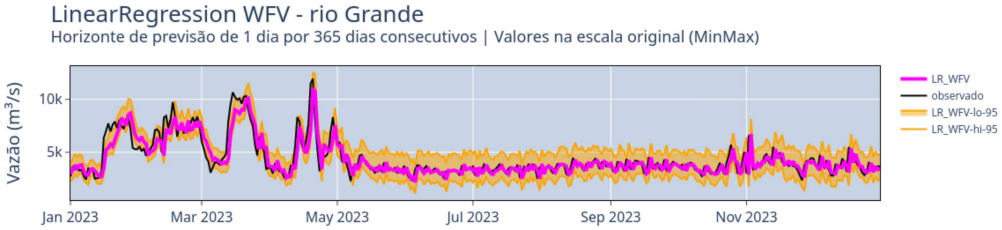


Modelo	MAPE	KGE	PBIAS (%)
CatBoost	0.12605	0.76212	-4.44499
CatBoost-lo-95	0.16782	0.73785	-14.28403
CatBoost-hi-95	0.20847	0.72743	10.6298

Cobertura empírica do intervalo de previsão
63.56 %

Fonte: O autor.

Figura 4.40 – *Walk-Forward Validation* para o modelo *Linear Regression*.



Modelo	MAPE	KGE	PBIAS (%)
LinearRegression	0.13353	0.83365	-1.99105
LinearRegression-lo-95	0.27533	0.71266	-24.84153
LinearRegression-hi-95	0.3772	0.64855	29.36936

Cobertura empírica do intervalo de previsão
84.38 %

Fonte: O autor.

Figura 4.41 – *Walk-Forward Validation* para o modelo *Random Forest*.

Modelo	MAPE	KGE	PBIAS (%)
RandomForest	0.123	0.83524	-2.18939
RandomForest-lo-95	0.17597	0.80847	-13.23295
RandomForest-hi-95	0.20471	0.76869	11.33873

Cobertura empírica do intervalo de previsão
63.01 %

Fonte: O autor.

A análise dos gráficos de dispersão dos resíduos ao longo do tempo para os modelos revela informações importantes sobre a evolução do desempenho de cada modelo em termos de erro durante o período analisado para o rio Grande. Aqui é possível identificar padrões de erro, presença de *outliers* e a estabilidade dos resíduos ao longo do tempo.

De início é possível notar as largas faixas de valores em que os resíduos se concentraram. O CB teve valores de *lower fence* e *upper fence* de $-1328,72$ e $1492,10$, respectivamente, uma amplitude de 2820, o que lhe conferiu uma prevalência de $89,32\%$ dos resíduos dentro destes limites. De antemão, o menor valor dentre os modelos. Continuando para o modelo LR, tendo $-1628,38$ para o valor da faixa mínima e $1687,81$ para a máxima – amplitude de 3315 – teve prevalência de $92,60\%$ dos resíduos. Isso parece realmente bom, mas com uma ampla faixa assim deixa indicado que o modelo oscilou muito nas previsões realizadas, demostrando incerteza. Finalmente, o RF, ficou "no meio do caminho" entre os modelos anteriores, tendo valores de $-1422,62$ para *lower fence* e $1531,92$ para *upper fence*, perfazendo uma amplitude de 2953, o modelo teve prevalência de $90,68\%$ dos resíduos dentro da área sombreada.

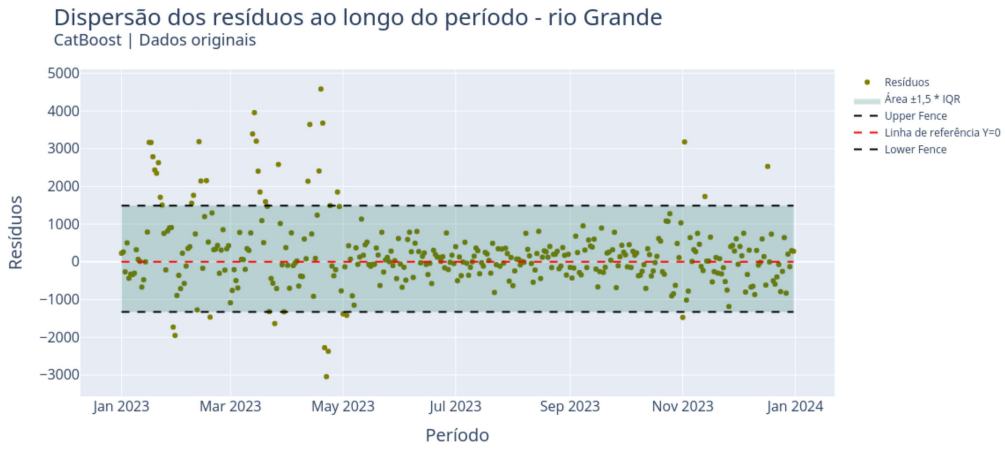
O modelo CB demonstra uma concentração dos resíduos dentro da faixa de $\pm 1000 m^3/s$ na maior parte do período estudado – figura 4.42. Contudo, nos primeiros meses, de janeiro a maio de 2023, observa-se um maior espalhamento dos resíduos, com vários pontos excedendo os $2000 m^3/s$. A presença de *outliers* entre março e maio de 2023 indica que o modelo enfrentou dificuldades para lidar com picos de vazão nesse período, ocasionando erros elevados. Apesar disso, após o pico de erro em maio, os resíduos tornaram-se mais constantes, com variações menores e mais previsíveis, embora alguns *outliers* persistam no final do ano, demonstrando, desta forma, que houve desafios na captura de variações extremas.

Em contraste, o LR apresenta uma dispersão dos resíduos significativamente maior ao longo do tempo quando comparada ao CB – figura 4.43 e é possível inferir isso olhando para a larga amplitude dos limites da área sombreada. Mesmo nos períodos de menor erro, como de julho a novembro de 2023, os resíduos ainda apresentam uma variação considerável, ainda muito dispersos, distantes da reta de referência $y = 0$. A estabilidade dos resíduos da LR ficou, assim, comprometida e é verificável pela elevada dispersão mesmo em períodos em que a série de dados observados apresentou vazões estáveis e mais moderadas. Essa variabilidade reflete a limitação do LR em capturar a complexidade dos dados de vazão, com dados de treinamento muito ruidosos – figura 3.24, revelando que o LR não lidou adequadamente com eventos extremos de vazão.

Por sua vez, o RF apresenta uma dispersão dos resíduos semelhante à do CB – figura 4.44 –, porém com uma quantidade menor de *outliers* – 35 pontos fora da área demarcada contra 39 pontos fora para o modelo CB –, considerando acima de *upper fence* e abaixo de *lower fence*. A maioria dos resíduos do RF está confinada à faixa delimitada,

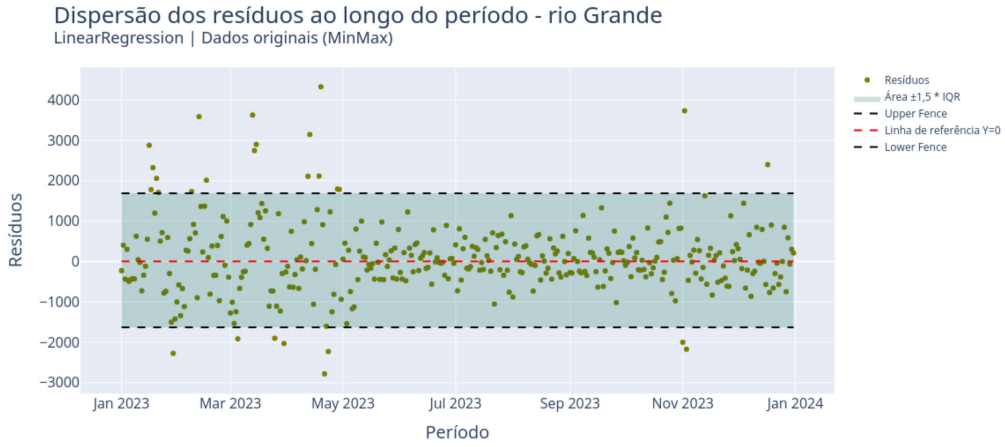
com algumas exceções nos meses de março e abril e também os meses de novembro e dezembro de 2023, onde o modelo aparentou maior dificuldade em prever corretamente as vazões. A presença reduzida de *outliers* no RF, em comparação ao CB, bem como uma amplitude menor comparada ao LR, sugere uma maior robustez na captura das variações de vazão ao longo do tempo, embora alguns *outliers* ocorram durante períodos de maior instabilidade nas vazões.

Figura 4.42 – Dispersão dos resíduos ao longo do ano.



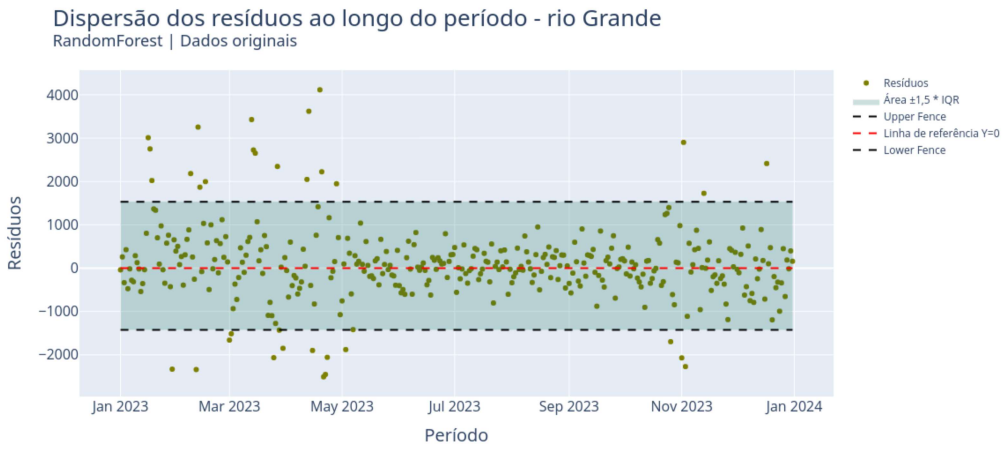
Fonte: O autor.

Figura 4.43 – Dispersão dos resíduos ao longo do ano.



Fonte: O autor.

Figura 4.44 – Dispersão dos resíduos ao longo do ano.



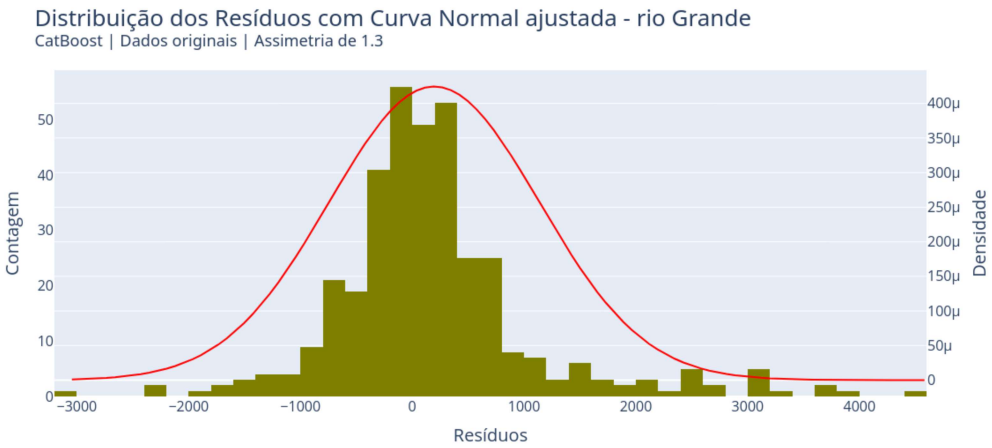
Fonte: O autor.

O CB apresentou assimetria de $1,3$, o que significa uma distribuição levemente assimétrica para a direita. Embora a porção predominante dos resíduos estivesse concentrada próximo a 0, houve uma quantidade considerável de erros positivos, especialmente para valores acima de 2000. A curva normal ajuda a demonstrar que os resíduos seguem uma forma próxima da normalidade, porém com uma cauda longa à direita. Esse comportamento significa que o CB tende a cometer alguns erros maiores nas previsões de altas vazões, subestimando estas vazões mais extremas – figura 4.45.

Em contraste, o LR apresentou uma assimetria de $0,97$, ou seja, a distribuição dos resíduos no modelo LR foi quase simétrica, com um leve desvio à direita. Os erros estiveram razoavelmente bem concentrados em torno de 0, apesar da grande variabilidade mencionada anteriormente, com uma leve tendência de subestimativas. Contudo, ainda houve resíduos significativos acima de $1500m^3/s$, mostrando que o LR não capturou adequadamente as variações maiores nas vazões – figura 4.46.

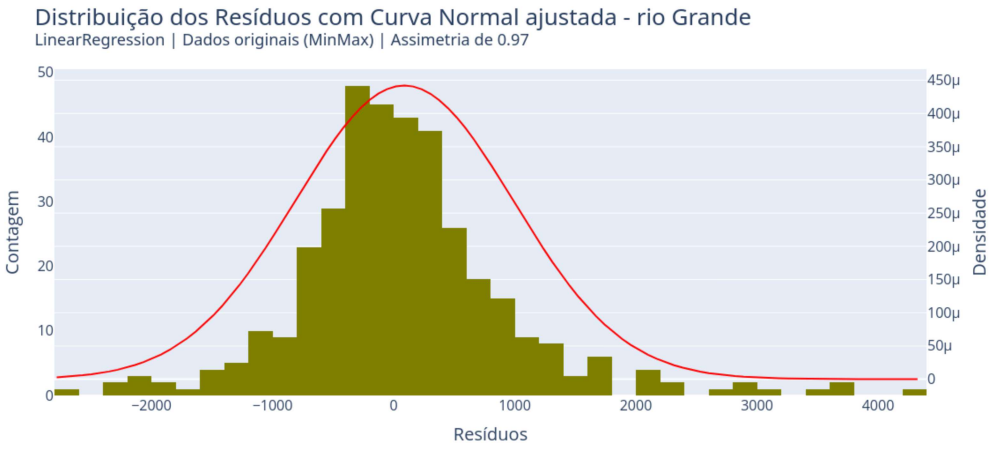
O RF apresentou a menor assimetria entre os três modelos, com um valor de $0,82$, a distribuição de resíduos mais simétrica entre todos. O RF lidou melhor com os erros, distribuindo-os de maneira mais uniforme em torno de 0, concentrando os resíduos, em sua maior parte, na faixa $\pm 1000m^3/s$, e teve a distribuição dos resíduos mais precisa, seguindo a forma da curva normal em comparação com o CB e o LR. Isso indica que o modelo RF teve uma maior capacidade de gerar previsões com erros menores e menos dispersos, o que ajuda a reforçar o resultado visto pelas métricas KGE, $0,83$, e MAPE, $0,12$. Este comportamento e entendimento sugerem ao modelo RF uma maior robustez na gestão de valores extremos de vazão – figura 4.47.

Figura 4.45 – Histograma e curva-normal dos resíduos.



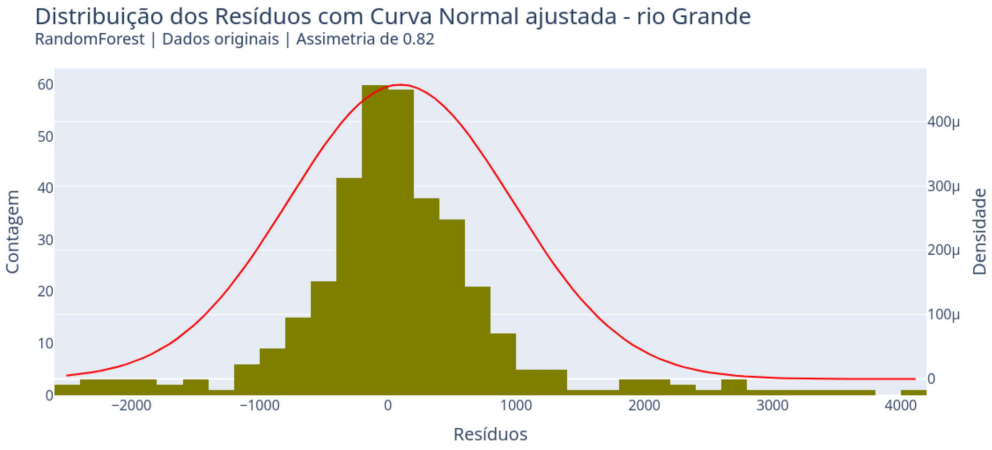
Fonte: O autor.

Figura 4.46 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

Figura 4.47 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

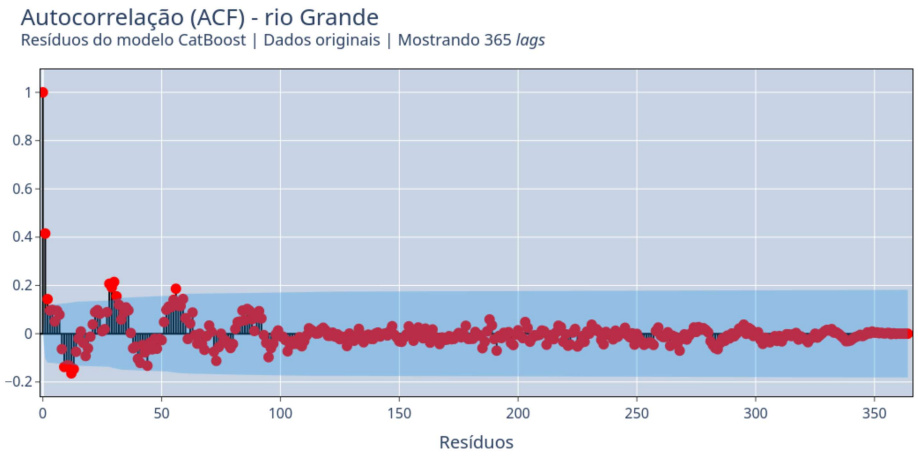
Os gráficos ACF de todos os modelos demonstraram que as primeiras 30 *lags* apresentaram autocorrelação significativa e positiva. Este comportamento mostra a presença de uma estrutura nos resíduos que não foi completamente capturada pelos modelos, sugerindo que ainda mantiveram dependências temporais não modeladas de forma adequada. A partir da *lag* 30, aproximadamente *lag* 45 para o LR, as correlações caíram dentro da faixa de confiança, sinalizando uma melhora na independência dos resíduos. No entanto, as flutuações entre autocorrelações positivas e negativas dentro dessa faixa ainda foram perceptíveis.

Com o modelo RF foi menos, mas para os modelos CB e LR é possível reparar inclusive um comportamento sazonal nas *lags* iniciais do gráfico, deixando ainda mais evidente que houve comportamento temporal que ficou remanescente nos resíduos.

As conclusões gerais indicam que todos os modelos apresentaram algum nível de autocorrelação significativa nas primeiras *lags*, nenhum deles conseguiu modelar completamente as dependências temporais presentes nos dados. Isso aponta para a possibilidade de melhorias, como, por exemplo, a incorporação de mecanismos para lidar com sazonalidade ou dependência temporal de forma mais direta, por exemplo, *features* de *lag* mais robustas. Cabe destacar ainda que a qualidade dos dados de treinamento também exerce influência neste resultado por parte dos modelos, basta comparar com os outros rios.

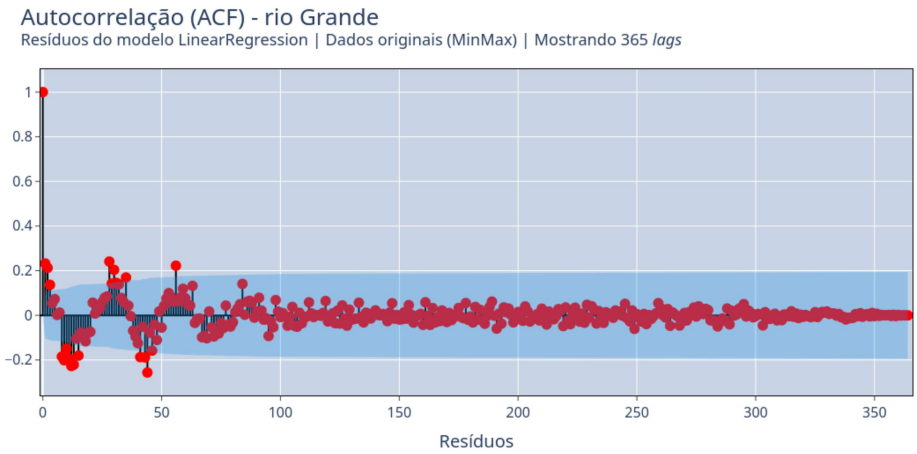
Resultados visíveis nas figuras 4.48, 4.49 e 4.50.

Figura 4.48 – Gráfico ACF dos resíduos.



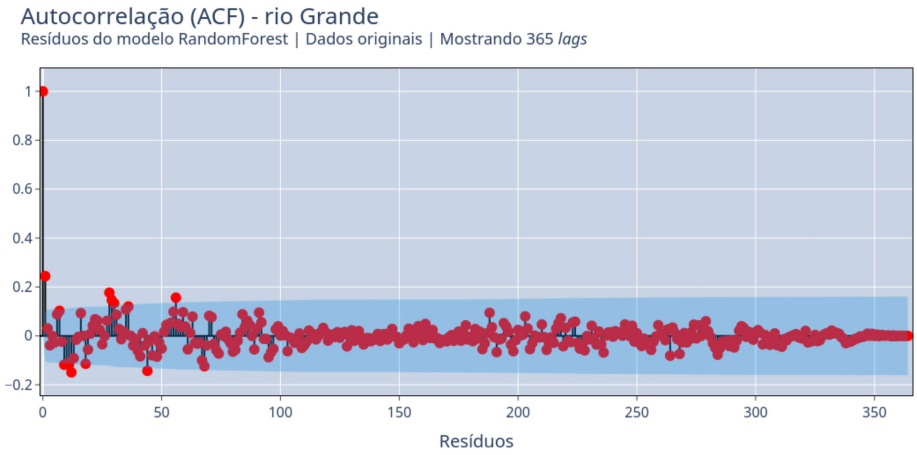
Fonte: O autor.

Figura 4.49 – Gráfico ACF dos resíduos.



Fonte: O autor.

Figura 4.50 – Gráfico ACF dos resíduos.



Fonte: O autor.

4.1.4 Rio São Francisco

O CB apresentou um KGE de $0,93$ – figura 4.51 –, apresentando assim uma boa eficiência nas previsões e demonstrando que o modelo conseguiu replicar o comportamento hidrológico observado de forma consistente. O LR destacou-se com o maior valor de KGE, atingindo $0,98$ – figura 4.52 –, o que corroborou sua elevada capacidade preditiva e eficiência em relação aos dados observados. O RF apresentou um KGE de $0,97$, o que demonstra uma ótima qualidade de previsão, que se aproximou consideravelmente dos valores observados, embora ligeiramente inferior ao do modelo LR – figura 4.53.

Em seguida, a métrica MAPE foi analisada. O CB exibiu um MAPE de $0,032$, o que é um erro percentual médio muito baixo. O LR apresentou um MAPE de $0,027$ e o RF $0,025$, os dois modelos com o melhor resultado, sendo o RF levemente melhor. Mas em termos práticos, ambos os modelos tiveram a mesma qualidade no erro médio das previsões. Considerando o resultado da KGE, até este momento, os dois modelos ficaram quase idênticos.

Quanto ao PBIAS, o CB apresentou um valor de $-2,61\%$, o que é um leve viés negativo, indicando que o modelo tende a subestimar as vazões previstas. O LR registrou um PBIAS mediano entre os modelos, com $-0,98\%$, um viés sistemático de subestimar as previsões, mas tão baixo que é quase irrelevante. Um desempenho excelente. O RF exibiu um PBIAS de $-0,80\%$, o menor viés dentre todos os modelos, ainda com subestimação, porém, assim como o LR, baixíssimo. As previsões praticamente "acertaram" os valores observados. Em suma, neste ponto, todos os modelos apresentaram viés sistemático de subestimação, com o CB ficando com o resultado que pode ser chamado de "pior".

Por fim, quanto à cobertura empírica. O CB cobriu $90,68\%$ dos dados, demonstrando uma boa capacidade de captura das incertezas nas previsões. O LR alcançou a maior cobertura empírica, com $96,71\%$, tendo o intervalo de previsão com o melhor desempenho, o modelo capturou quase todos os eventos da série de vazões observadas. O RF, por sua vez, apresentou a menor cobertura empírica, com $82,47\%$, o que indica que o intervalo de previsão gerado pelo RF foi mais estreito e possivelmente não capturou todas as variações na série de vazão.

O LR destacou-se como o modelo com maior eficiência na replicação do comportamento hidrológico observado, evidenciado pelo maior valor de KGE, MAPE e PBIAS entre os modelos e pela cobertura empírica mais alta. Porém, de forma geral, todos os modelos apresentaram resultados competitivos e para os dados do rio São Francisco se mostraram altamente eficientes.

Foi observado que a área sombreada nos gráficos de dispersão dos resíduos permaneceu bastante estreita para todos os modelos. Um bom comportamento geral, deixando indicado que os modelos realizaram previsões bastante próximas aos valores observados.

Figura 4.51 – *Walk-Forward Validation* para o modelo *CatBoost*.

Modelo	MAPE	KGE	PBIAS (%)
CatBoost	0.03296	0.93396	-2.61473
CatBoost-lo-95	0.15104	0.85745	-12.40588
CatBoost-hi-95	0.20589	0.86271	12.31072

Cobertura empírica do intervalo de previsão
90.68 %

Fonte: O autor.

Figura 4.52 – *Walk-Forward Validation* para o modelo *Linear Regression*.

Modelo	MAPE	KGE	PBIAS (%)
LinearRegression	0.02707	0.98472	-0.98255
LinearRegression-lo-95	0.09813	0.86799	-9.36665
LinearRegression-hi-95	0.11665	0.87176	10.12609

Cobertura empírica do intervalo de previsão
96.71 %

Fonte: O autor.

Especificamente, o modelo CB apresentou um limite *upper fence* de $103,87$ e *lower fence* de $-72,34$, resultando em uma amplitude de $176,21$, com prevalência dos resíduos dentro desses limites de $83,84\%$. Contudo, o CB exibiu resíduos discrepantes no início do ano, especialmente em janeiro e fevereiro, antes de estabilizar a partir de março,

Figura 4.53 – *Walk-Forward Validation* para o modelo *Random Forest*.

Modelo	MAPE	KGE	PBIAS (%)
RandomForest	0.02571	0.97754	-0.80974
RandomForest-lo-95	0.06906	0.94219	-5.12336
RandomForest-hi-95	0.10486	0.93372	6.14794

Cobertura empírica do intervalo de previsão
82.47 %

Fonte: O autor.

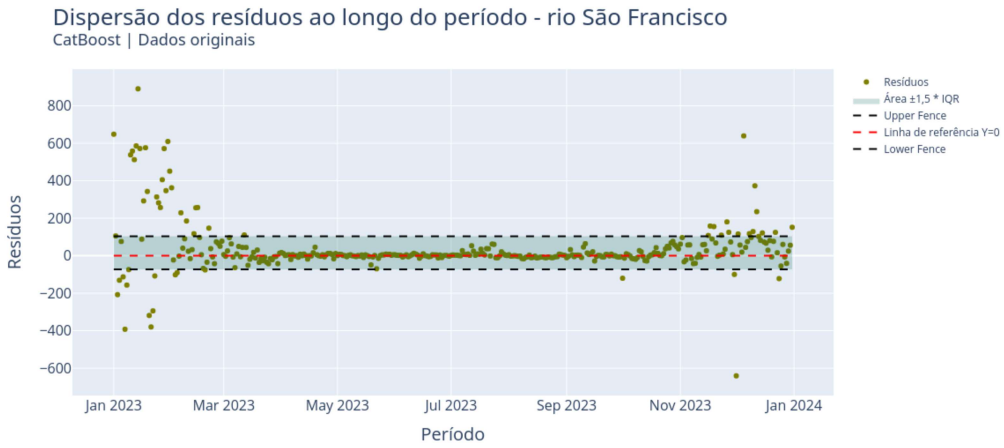
restringindo-se à área de tolerância – figura 4.54. Na porção final do ano, o modelo voltou a apresentar alguns valores discrepantes, embora a vasta maioria dos resíduos permanecesse próxima do limite superior.

Por sua vez, o modelo LR apresentou um *upper fence* de $67,78$ e um *lower fence* de $-48,93$, com uma amplitude de $116,71$ e uma prevalência de $81,92\%$. Embora o LR também tenha exibido resíduos discrepantes no início do ano, a quantidade foi visualmente menor comparada ao CB e ao RF, e a estabilização dos resíduos ocorreu apenas a partir de março. Na porção final do período, o LR mostrou um aumento no número de resíduos fora das faixas estabelecidas, tanto acima quanto abaixo dos limites, indicando uma menor robustez em comparação ao CB para este período. A amplitude dos resíduos do LR variou entre -600 e 600, sendo menor quando comparada aos outros modelos. Adicionalmente, o LR apresentou mais *outliers* acima do limite superior, similar ao RF, e menos *outliers* abaixo do limite inferior – figura 4.55.

O RF exibiu um limite superior de $46,83$ e inferior de $-39,87$, resultando em uma amplitude de $86,70$ e prevalência de $74,52\%$ dos resíduos na área sombreada, demonstrando uma distribuição de resíduos bastante controlada em torno da linha de referência 0. No entanto, o RF também apresentou diversos *outliers* no início do ano, de janeiro a março, e demorou para estabilizar na porção central do período, semelhante ao comportamento do LR. Na porção final do período, o RF voltou a apresentar muitos *outliers*, e um maior espalhamento, refletindo desafios na previsão de vazões extremas. A amplitude dos resíduos do RF variou entre -600 e 600, igualando-se à da LR – figura 4.56.

De maneira geral, os modelos demonstraram que, durante períodos de variações elevadas nas precipitações e, consequentemente, nas vazões, apresentaram resultados incertos, evidenciados pela quantidade de *outliers* tanto no início quanto no final do ano, com um resultado levemente melhor para o CB que teve menos *outliers*. Ao que tudo indica, houve um certo atraso dos modelos em capturar eventos extremos, levando-os a dificuldades em responder rapidamente às mudanças abruptas nas condições hidrológicas. O CB apresentou uma boa prevalência dos resíduos na área demarcada, apesar de algumas discrepâncias em períodos de alta variabilidade. O LR, embora com uma prevalência menor, apresentou também menor amplitude, houve equilíbrio entre estes dois aspectos, embora também tenha tido problemas com vazões extremas. O RF, apesar de sua estabilidade geral, também enfrentou desafios semelhantes em períodos de alta variabilidade. Estes resultados destacam a necessidade de aprimoramentos nos modelos para melhor capturar e responder a eventos hidrológicos extremos, potencialmente através da incorporação de mecanismos adicionais para detecção de picos de vazão.

Figura 4.54 – Dispersão dos resíduos ao longo do ano.

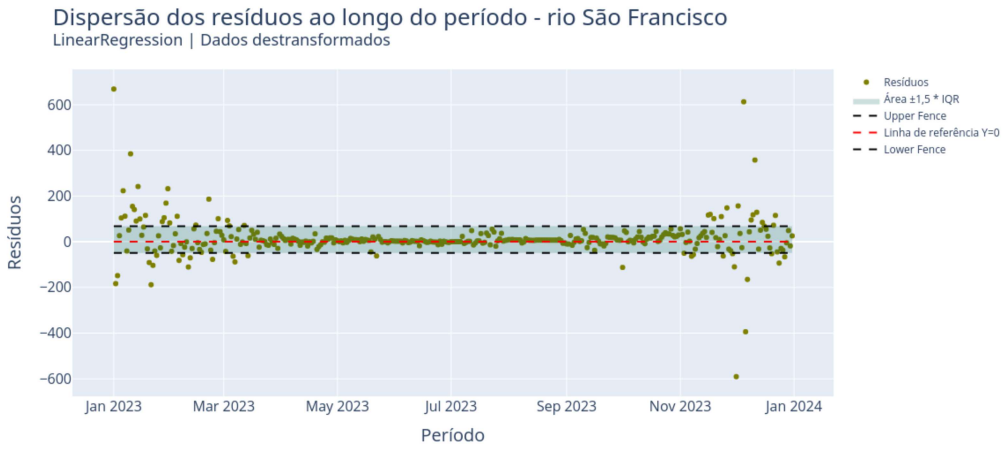


Fonte: O autor.

A análise comparativa dos resíduos dos modelos de previsão para o rio São Francisco revelou comportamentos semelhantes entre os diferentes modelos – figuras 4.57, 4.58 e 4.59. Todos os modelos demonstraram uma distribuição dos resíduos com cauda longa à direita, indicando assimetria positiva, com uma concentração central em torno de 0, em uma faixa que se estende de -200 a $200 m^3/s$. Essa característica sugere que, geralmente, os resíduos estiveram próximos ao valor esperado, mas houve uma tendência de ocorrência de erros discrepantes.

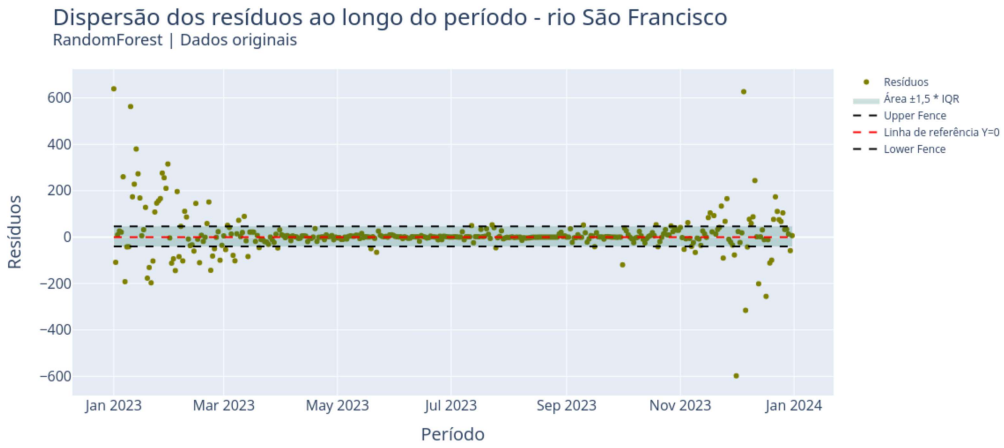
No entanto, o modelo LR apresentou uma menor quantidade de resíduos elevados a partir do valor de $300 m^3/s$ quando comparado aos outros modelos, ainda que o modelo RF tenha tido resultado parecido. Este resultado do modelo LR reflete, assim, uma menor instabilidade quanto à presença de *outliers* – mas é patente que todos tiveram *outliers*,

Figura 4.55 – Dispersão dos resíduos ao longo do ano.



Fonte: O autor.

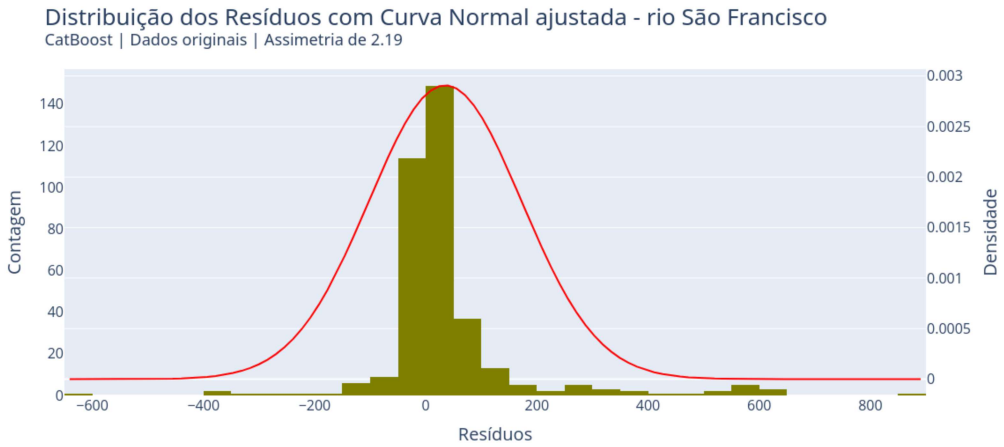
Figura 4.56 – Dispersão dos resíduos ao longo do ano.



Fonte: O autor.

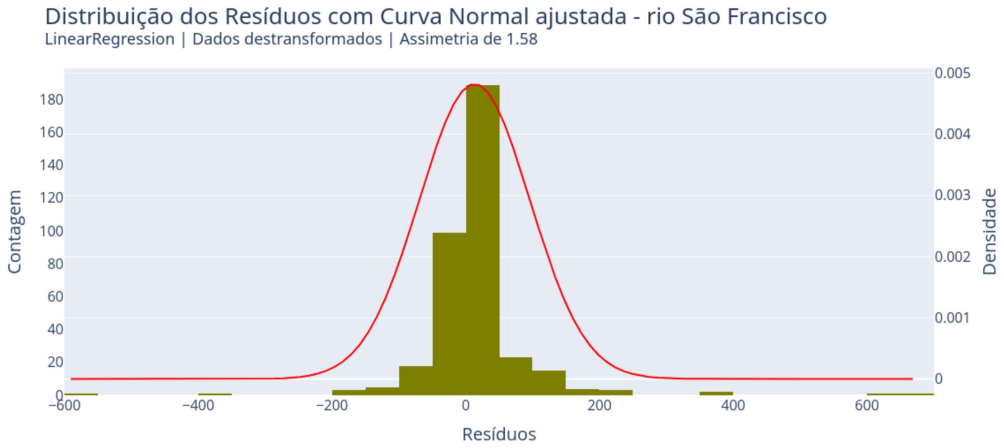
confirmados pelas análises anteriores. As medidas de assimetria aferidas foram de $2,19$ para o CB, $1,58$ para a LR e $1,75$ para o RF. Tais valores indicam que a distribuição dos resíduos do CB foi a mais assimétrica em relação aos outros modelos, enquanto a LR apresentou a menor assimetria, conferindo uma distribuição mais equilibrada dos resíduos.

Figura 4.57 – Histograma e curva-normal dos resíduos.



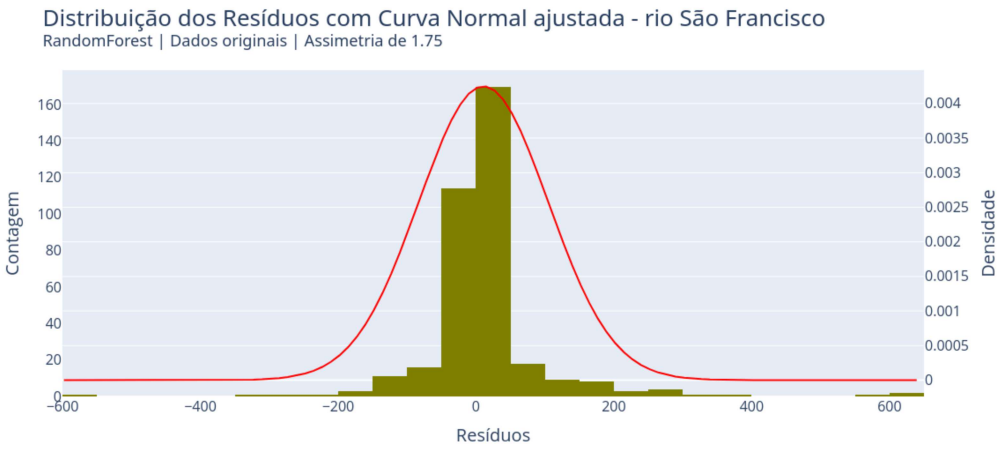
Fonte: O autor.

Figura 4.58 – Histograma e curva-normal dos resíduos.



Fonte: O autor.

Figura 4.59 – Histograma e curva-normal dos resíduos.



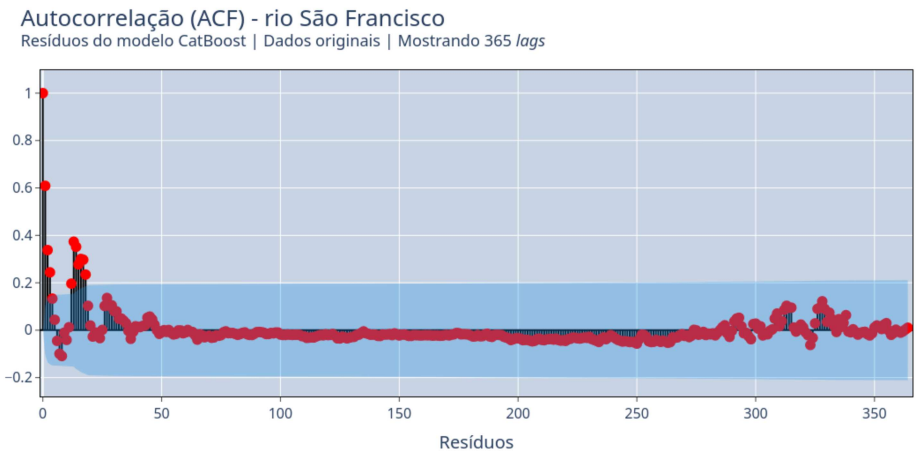
Fonte: O autor.

O modelo CB apresentou resíduos autocorrelacionados nas *lags* iniciais, estendendo-se aproximadamente até a *lag* 30 – figura 4.60. Posteriormente, os resíduos estabilizaram-se dentro da faixa de tolerância definida, demonstrando uma melhoria na independência dos erros. Notavelmente, mesmo no final do ano, quando os demais modelos exibiram resíduos fora da faixa ou no limite da mesma, o CB continuou a gerar resíduos estáveis, indicando robustez durante todo o período analisado. Nas *lags* iniciais, o CB demonstrou um comportamento sazonal nos resíduos, sugerindo que certos padrões sazonais próximos não foram adequadamente capturados pelo modelo, resultando em autocorrelações residuais persistentes.

Por sua vez, o modelo LR também exibiu alguns resíduos autocorrelacionados nas *lags* iniciais e também apresentou resíduos fora da faixa de tolerância no final do período. Contudo, a análise geral indicou que o comportamento do LR foi relativamente estável, suportado por uma faixa de tolerância estreita. Essa estabilidade sugere que, apesar das autocorrelações observadas, o modelo manteve uma consistência na distribuição dos resíduos ao longo do tempo, minimizando a ocorrência de erros sistemáticos – figura 4.61.

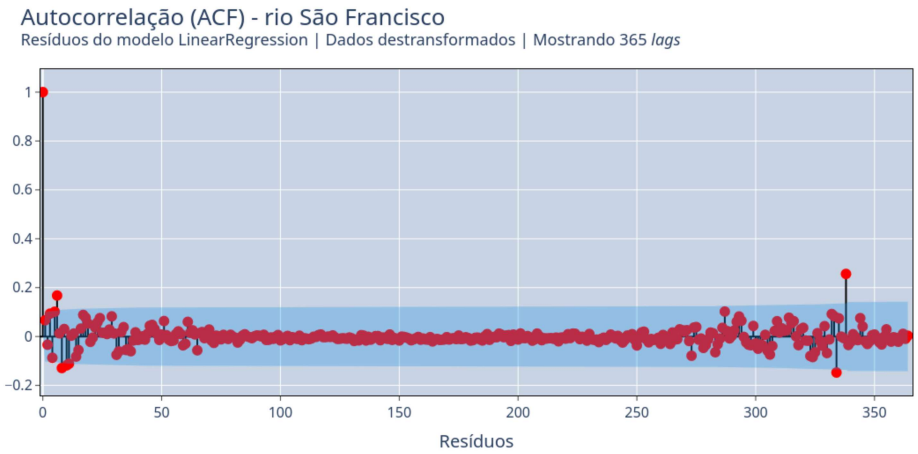
Finalmente, o modelo RF mostrou-se comportamentalmente semelhante ao modelo LR, apresentando também resíduos autocorrelacionados nas *lags* iniciais. No entanto, o RF não registrou escapadas de resíduos fora da faixa de tolerância no final do período, ficando alguns pontos bem no limite. Além disso, o RF manteve uma faixa de tolerância igualmente estreita, refletindo uma alta qualidade e estabilidade nos resíduos gerados. Juntamente com o modelo LR, o RF destacou-se como um dos modelos que apresentaram maior qualidade e estabilidade na autocorrelação dos resíduos – figura 4.62.

Figura 4.60 – Gráfico ACF dos resíduos.



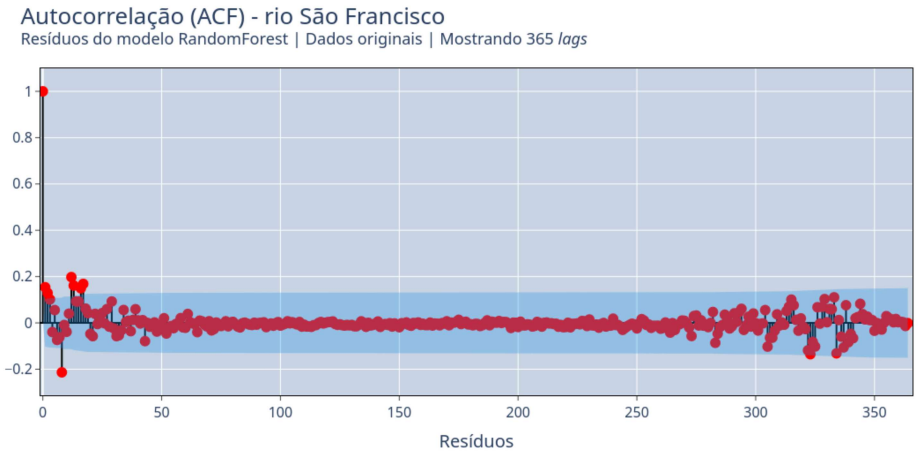
Fonte: O autor.

Figura 4.61 – Gráfico ACF dos resíduos.



Fonte: O autor.

Figura 4.62 – Gráfico ACF dos resíduos.



Fonte: O autor.

4.2 IMPORTÂNCIA DAS VARIÁVEIS

A análise das variáveis nos modelos de aprendizado de máquina constitui uma etapa fundamental que impacta diretamente a eficácia, interpretabilidade e confiabilidade dos resultados obtidos.(67) Essa análise envolve a avaliação da importância e relevância das variáveis de entrada, *features*, no processo de previsão realizado pelo modelo. Entre os diversos métodos disponíveis para essa finalidade, a análise de valores SHAP destaca-se por sua abordagem robusta, proporcionando percepções mais profundas e precisas sobre a contribuição de cada variável.

Tal análise permite identificar quais variáveis possuem maior influência nas previsões do modelo. Métodos tradicionais, como a importância das variáveis, *feature importance*, em algoritmos baseados em árvores – por exemplo, *Random Forest* e *CatBoost* – ou técnicas de regressão regularizada – como Lasso e Ridge na Regressão Linear –, fornecem ideias sobre quais variáveis contribuem de forma significativa para a desempenho do modelo. Contudo, essa análise é baseada no cálculo de coeficientes – pesos – que certas variáveis tiveram e que geraram um determinado resultado, no caso da Regressão Linear, ou a contagem de quantas divisões uma determinada variável de entrada foi usada para a geração da árvore, no caso dos modelos baseados em árvore. É uma relação direta entre os atributos de entrada e os resultados, no seguinte sentido: se uma variável apareceu mais vezes na árvore, então ela é muito importante, se uma variável teve o maior peso, então ela deve ser a mais importante e isso pode gerar uma interpretação enviesada.

A incorporação dos valores SHAP aprimora a análise ao oferecer uma metodologia consistente e teoricamente fundamentada para explicar previsões individuais. SHAP atribui a cada variável um valor de importância para uma determinada previsão, baseado na teoria dos jogos cooperativos.(66) O método permite uma compreensão detalhada de como cada variável contribui para o resultado do modelo, tanto de forma global – resultado da previsão – quanto local – uma instância dentre todas as previsões. Diferentemente das métricas tradicionais de importância das variáveis, os valores SHAP consideram as interações entre as variáveis, proporcionando uma explicação clara e aditiva da previsão, o que facilita a interpretabilidade e transparência.

Além disso, a análise com valores SHAP aborda limitações de outros métodos de análise de variáveis ao garantir justiça e consistência na atribuição da importância das variáveis. SHAP satisfaz propriedades desejáveis como precisão local, ausência de viés e consistência, que nem sempre são garantidas por outros métodos. Isso torna o SHAP uma ferramenta robusta para a análise de variáveis, fornecendo explicações confiáveis que podem ser compreendidas e utilizadas por quaisquer usuários.(68)

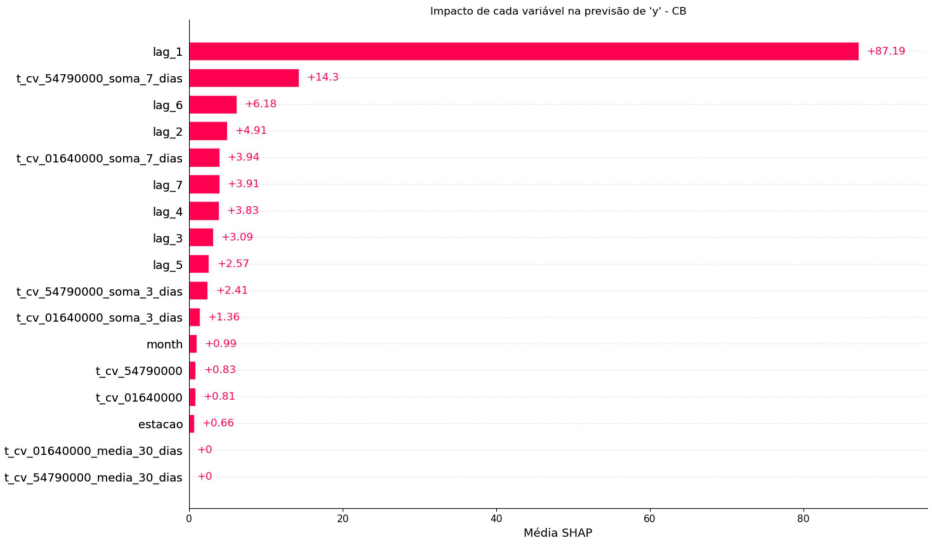
Contudo, uma coisa precisa ficar bastante clara: a análise SHAP não avalia causalidade, ou seja, realizar a avaliação de impacto de cada variável na previsão não significa dizer que tais variáveis mais importantes **causem** o evento estudado.

A comparação da importância das variáveis será realizada por modelo, visto que o que se busca apresentar aqui é o comportamento de cada um destes frente aos tipos de variáveis do problema, variáveis do tipo categóricas e contínuas. Se as variáveis são de um determinado rio ou de outro não importa para o modelo nem para a análise SHAP.

4.2.1 Catboost

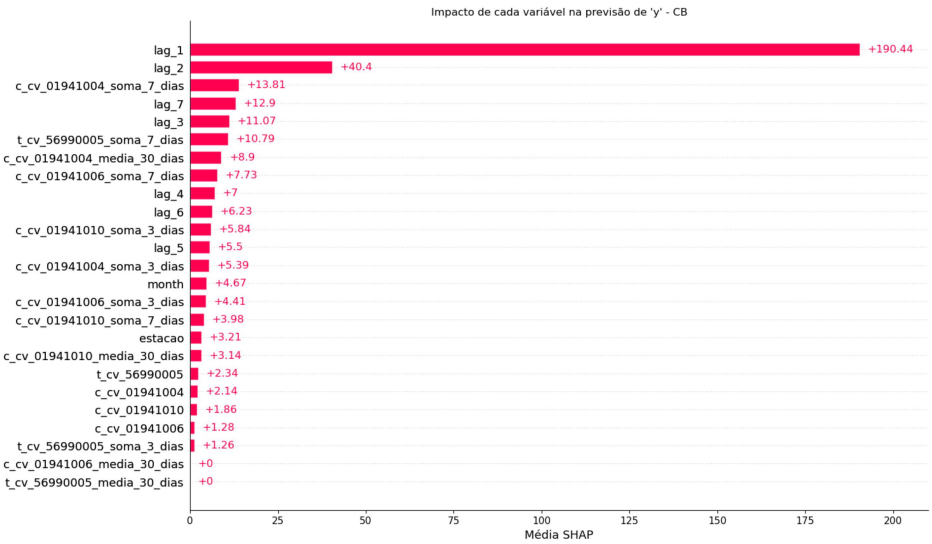
Os gráficos mostrados em 4.63, 4.64, 4.65 e 4.66 representam a média do valor absoluto SHAP para cada uma das variáveis, correspondente a cada rio estudado.

Figura 4.63 – Média dos valores SHAP - rio Jequitinhonha.



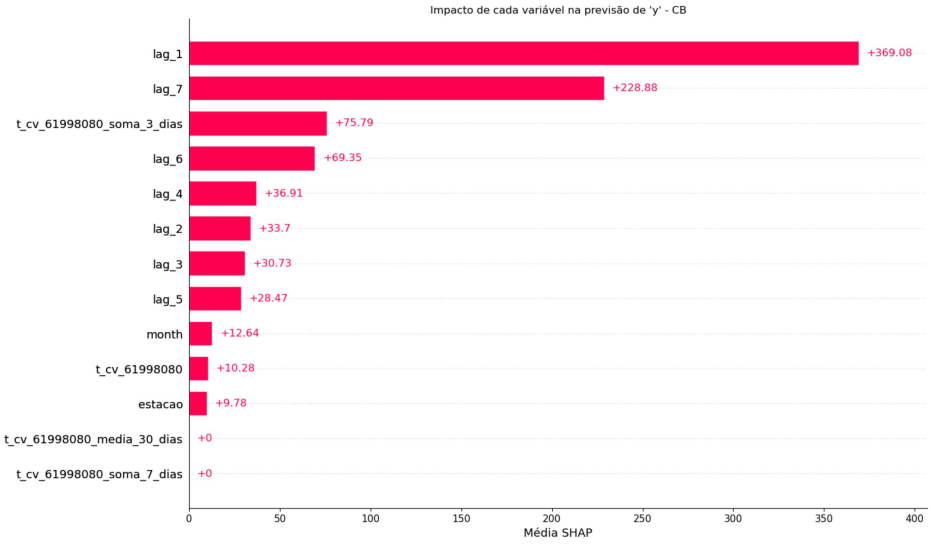
Fonte: O autor.

Figura 4.64 – Média dos valores SHAP - rio Doce.



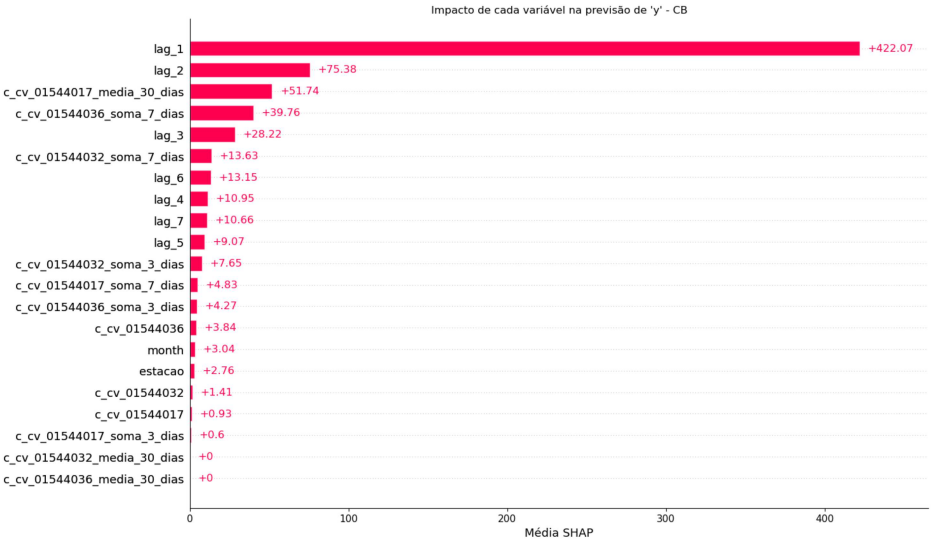
Fonte: O autor.

Figura 4.65 – Média dos valores SHAP - rio Grande.



Fonte: O autor.

Figura 4.66 – Média dos valores SHAP - rio São Francisco.



Fonte: O autor.

Para todos os rios, os valores defasados – *lags* – de vazão empenham maior importância nas previsões quando se observa as 10 variáveis de maior impacto. Isso é o que se espera, de fato, visto que as *lags* possuem forte autocorrelação. Esta natureza do problema e a forma de modelar o problema explica tamanha prevalência destas variáveis na importância das previsões. A lag_1 pode ser interpretada como "para prever a vazão de amanhã, basta que eu saiba a vazão de hoje".

A tentativa de criar variáveis de chuva de valores acumulados surtiu bom efeito para o modelo *CatBoost*. A não-linearidade na interação entre fenômenos de precipitação e vazão foi melhor capturada quando do uso de valores acumulados das chuvas. Apenas usando valores das estações não ficou bom para quaisquer dos rios, visto que, como já salientado anteriormente, as séries de chuva apresentaram muitos valores zerados, eventualmente por muito dias em sequência, e isso naturalmente não teria um peso relevante para realização das previsões – figuras 4.63, 4.64, 4.65 e 4.66.

As variáveis que ficaram na parte inferior do gráfico, se removidas, podem melhorar o comportamento geral dos modelos. Tomemos como exemplo o rio São Francisco. Uma nova rodada de experimento foi executada e foi feita a remoção das variáveis que tiveram média abaixo de 1. Foram as colunas 'c_cv_01544017', 'c_cv_01544017_soma_3_dias', 'c_cv_01544032_media_30_dias', 'c_cv_01544036_media_30_dias'. O resultado ficou melhor após a remoção – figura 4.67 – comparada com antes – figura 4.51. Com exceção da métrica MAPE, que houve uma piora de menos de 1%, nas demais métricas o modelo ficou melhor.

Tabela 4.1 – Resultados antes e depois da remoção das variáveis - rio São Francisco.

Situação	**KGE**	**MAPE**	**PBIAS**	**Cobertura empírica**
Antes	$0,93396$	$0,03296$	$-2,61473$	$90,68\%$
Depois	$0,96111$	$0,03344$	$-1,86549$	$91,23\%$

Fonte: O autor.

Agora como uma previsão é realizada pelo modelo, apresentando a interação das variáveis que levam a determinada previsão, neste caso, a previsão do dia 19 de julho de 2023. Este gráfico foi gerado com dados do rio Jequitinhonha – figura 4.68.

Um valor base é calculado sem que se considere a importância das variáveis, como se todas fossem neutras na previsão, e aqui este valor, também chamado de valor esperado – *expected value* – foi de $29,406$.

Perceba como as *lags* estão fortemente relacionados à previsão que o modelo faz, com a 'lag_1' tendo a maior relevância na previsão. Aqui esta variável apresentou valor negativo, digamos, 'puxou' a previsão para baixo – aquele número negativo de $-14,207$ é o valor SHAP e não está em unidade de vazão. As variáveis de chuva acumulada, com exceção da média de 30 dias, estiveram relevantes na previsão e isso é um bom indicativo

Figura 4.67 – *Walk-Forward Validation* para o modelo *CatBoost* após remoção de variáveis.

Modelo	MAPE	KGE	PBIAS (%)
CatBoost	0.03344	0.96111	-1.86549
CatBoost-lo-95	0.15922	0.87215	-12.24437
CatBoost-hi-95	0.23666	0.84607	14.63098

Cobertura empírica do intervalo de previsão
91.23 %

Fonte: O autor.

de que ao incluir estas variáveis, a modelagem proposta favoreceu este comportamento. Por outro lado, as variáveis categóricas apresentaram pouca relevância no resultado do modelo para este dia.

Este gráfico para o rio Jequitinhonha combina com o mostrado na figura 4.63, a ordem de importância, e reforça que as variáveis acumuladas da média de 30 podem, eventualmente, serem removidas dos dados de entrada, pois tanto no gráfico geral quanto neste de uma previsão específica, o impacto destas variáveis permaneceu zerado.

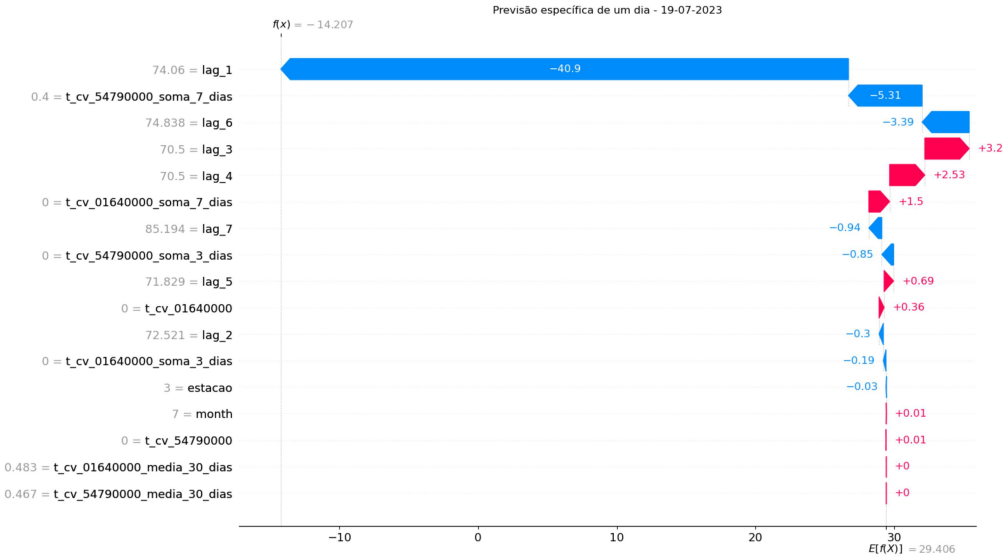
Em suma, foi mostrado por este gráfico que as variáveis *lag* têm um papel predominante na previsão para este dia específico, com 'lag_1' sendo o fator de maior impacto. As variáveis de tendência, assim como a sazonalidade – 'estacao' e 'month' –, têm impacto consideravelmente menor, ao passo que as variáveis de média mostraram-se irrelevantes.

4.2.2 Regressão linear

Para este modelo, como foi preciso aplicar a técnica de *OneHotEncoding*, apareceram diversas variáveis para mês e estação. Para deixar clara a leitura, a variável 'month_1.0' corresponde ao mês de janeiro e assim por diante até o mês de dezembro com 'month_12.0'. Para as variáveis da estação, a ordem é: 1 para o verão, 2 para outono, 3 é o inverno e, finalmente, 4 corresponde à primavera.

Dito isso, aqui também se manteve a primazia das *lags* conforme apresentado para o modelo CB. Em todos os gráficos – figuras 4.69, 4.70, 4.71 e 4.72 –, as variáveis defasadas

Figura 4.68 – *Waterfall plot* de uma previsão específica no rio Jequitinhonha.



Fonte: O autor.

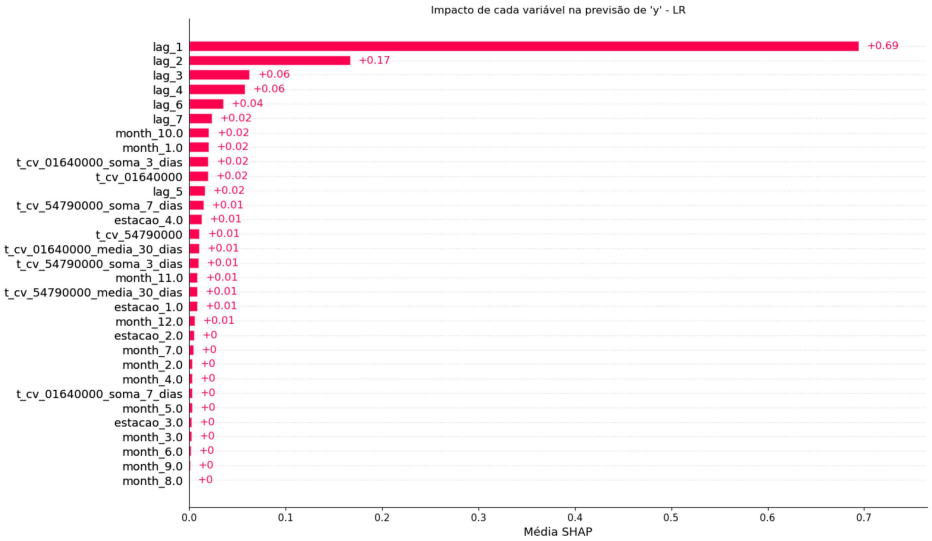
(lag_1, lag_2, etc.) têm uma importância significativa. A lag_1 aparece consistentemente como a variável mais importante, indicando que o valor de vazão de um dia anterior é altamente preditivo para o valor atual. Outras variáveis '*lag*' também aparecem entre as mais importantes, embora com menor impacto quando comparadas a lag_1.

Algo notável é que quando a qualidade dos dados de entrada não foi bom, como o caso do rio Grande, o valor de impacto da 'lag_1' ficou menos distante das demais variáveis - figura 4.71.

Atributos categóricos como '*month*' e 'estacao' têm um impacto, no geral, baixo em todos os rios, sugerindo que a sazonalidade não seria tão preditiva quanto os valores de atraso para o modelo linear. Outra questão também é que devido à modelagem aplicar *OneHotEncoding* isso possa ter diluído o impacto para as previsões.

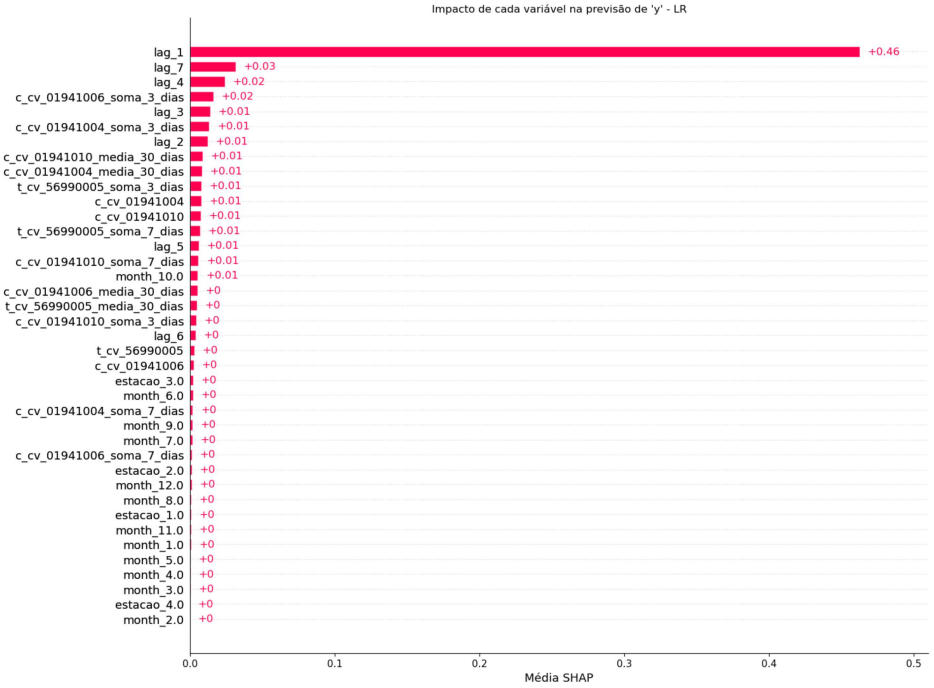
As variáveis que representam somas e médias ('soma_3_dias', 'soma_7_dias', 'media_30_dias', etc.) têm, em geral, impacto pequeno para todos os rios e ficaram eclipsadas pelas variáveis de vazão defasadas.

Figura 4.69 – Média dos valores SHAP - rio Jequitinhonha.



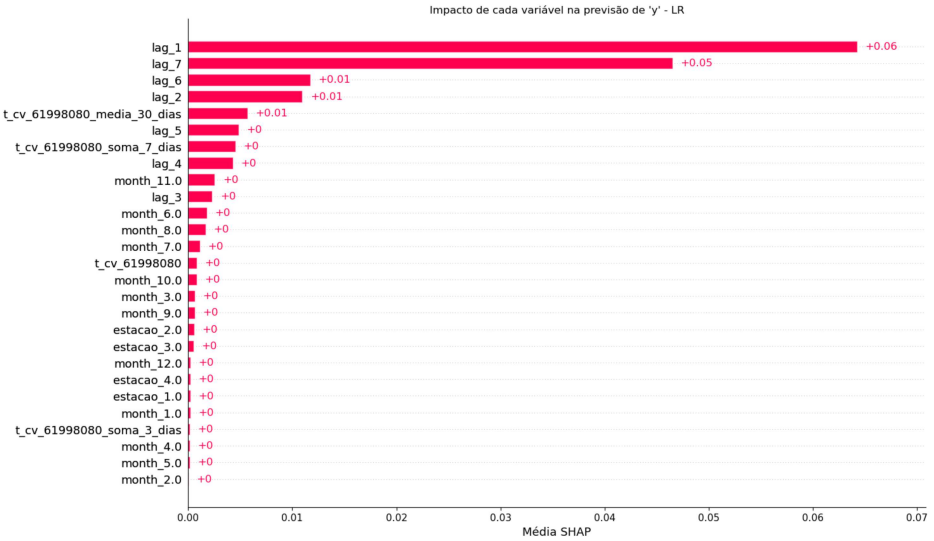
Fonte: O autor.

Figura 4.70 – Média dos valores SHAP - rio Doce.



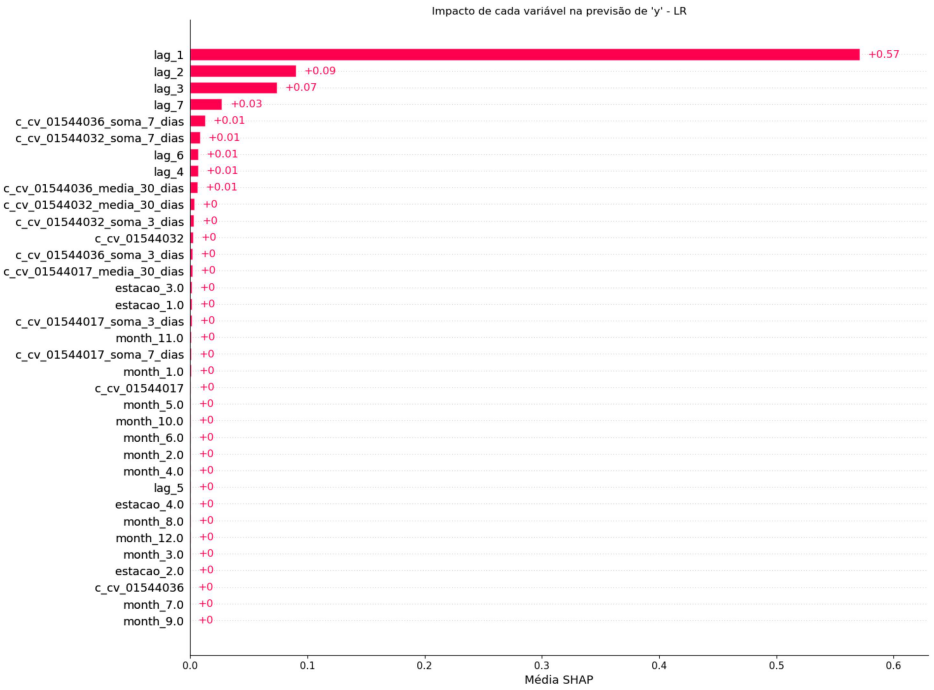
Fonte: O autor.

Figura 4.71 – Média dos valores SHAP - rio Grande.



Fonte: O autor.

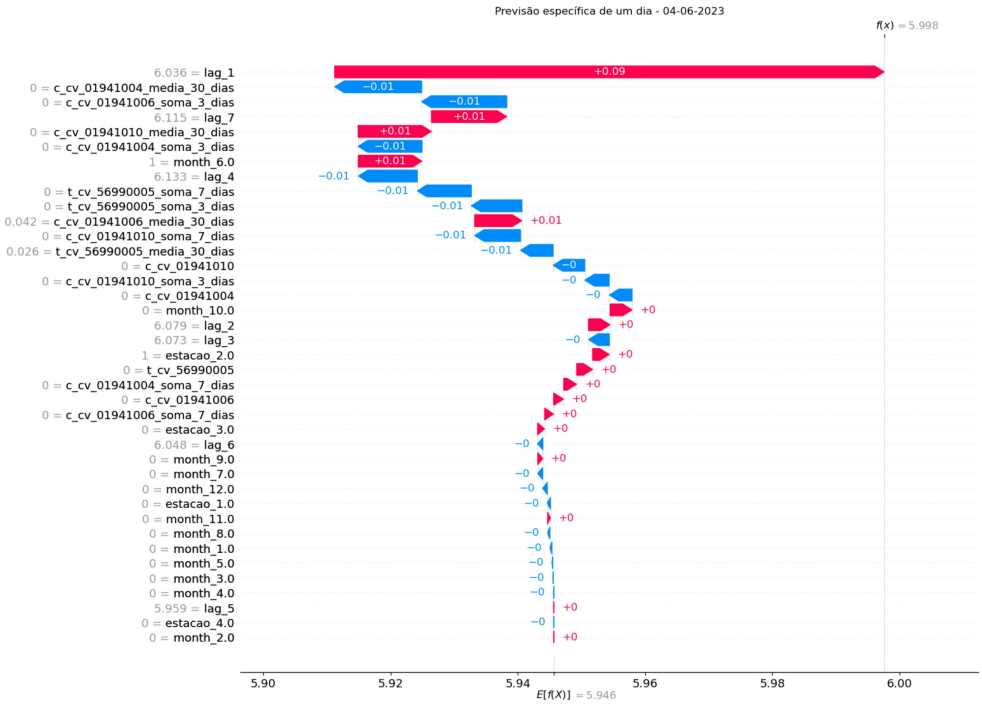
Figura 4.72 – Média dos valores SHAP - rio São Francisco.



Fonte: O autor.

A figura 4.73 mostra a dinâmica de uma previsão realizada, para o rio Doce. Este gráfico reforça a prevalência da variável lag_1, conforme demonstrado anteriormente em 4.70. As variáveis de tendência mostraram-se importantes neste resultado e o único atributo categórico relevante foi o mês de junho. Este mês é de outono, tempo de seca, tendência de vazões baixas, o que parece sugerir o impacto que as variáveis com esta característica tiveram para esta previsão, mesmo tais variáveis tendo valor 0.

Figura 4.73 – *Waterfall plot* de uma previsão específica no rio Doce.



Fonte: O autor.

De maneira geral, as variáveis categóricas desempenharam pouco ou mesmo nenhum efeito nas previsões. Desta forma, foi executado um experimento, para o rio Doce, sem estas variáveis, para averiguar se haveria alguma melhora no resultado. Mostrando a complexidade da análise de valores SHAP, a qual não demonstra causalidade entre as variáveis e o resultado do modelo, a nova execução apresentou uma melhora muito marginal, como apresentado na tabela 4.2 e na figura 4.74, com exceção apenas à cobertura empírica, que ficou pior.

Tabela 4.2 – Resultados antes e depois da remoção das variáveis - rio Doce.

Situação	**KGE**	**MAPE**	**PBIAS**	**Cobertura empírica**
Antes	$0,94579$	$0,08434$	$-2,75287$	$96,71\%$
Depois	$0,94787$	$0,08420$	$-2,51728$	$96,16\%$

Fonte: O autor.

Figura 4.74 – *Walk-Forward Validation* para o modelo *Linear Regression* após remoção de variáveis.

Modelo	MAPE	KGE	PBIAS (%)
LinearRegression	0.0842	0.94787	-2.51728
LinearRegression-lo-95	0.2307	0.65709	-23.80328
LinearRegression-hi-95	0.29192	0.61084	27.36471

Cobertura empírica do intervalo de previsão
96.16 %

Fonte: O autor.

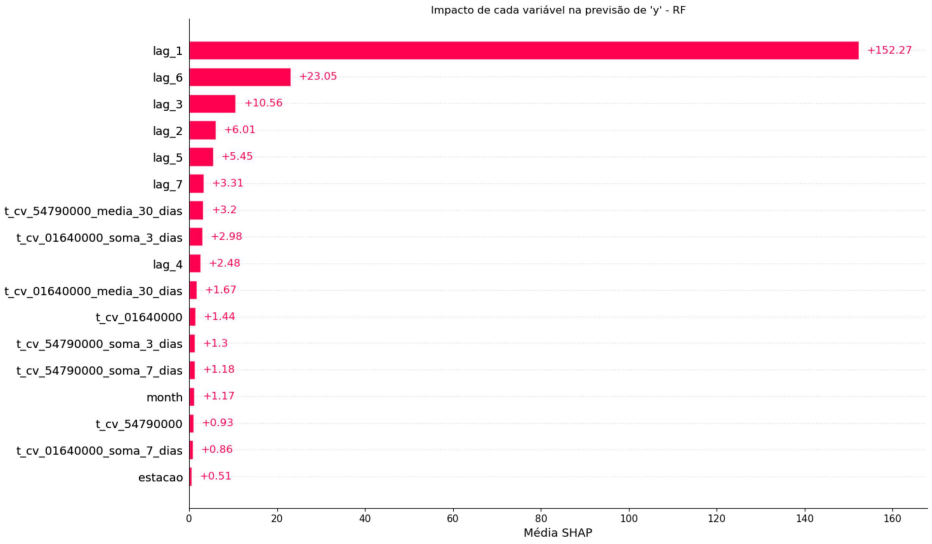
4.2.3 Random forest

Não houve variável de valor 0, todas as variáveis apresentaram algum impacto nas previsões, contudo, repete-se o comportamento dos modelos anteriores: os valores defasados de vazão se impõem frente às outras variáveis, sendo a lag_1 a variável que mais apresentou impacto. Isso indica que o modelo RF também depende muito dos valores recentes para prever o valor atual de vazão. Este entendimento é reforçado por outras variáveis defasadas além da lag_1, como lag_2, lag_3 e lag_6, que também apresentaram valores SHAP elevados, reforçando que as informações imediatas de dias anteriores são cruciais para a previsão.

Da mesma forma que a regressão linear, algumas variáveis de somas e médias de 3, 7 e 30 dias também apareceram nos gráficos do RF, mas com importância moderada, muito abaixo da importância que as variáveis *lags* apresentaram.

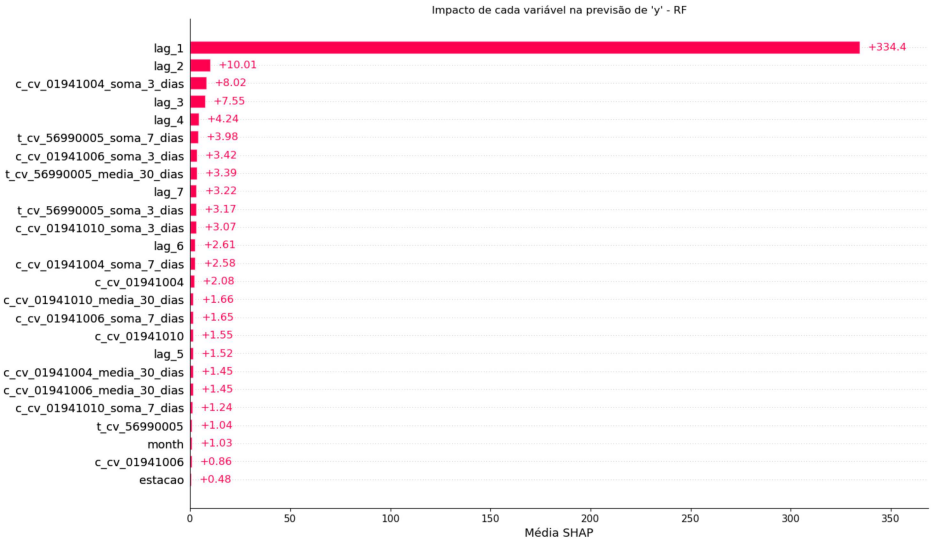
A consideração de todas as variáveis – nenhuma apresentou valor médio SHAP 0 – muito se deve à forma como o modelo *Random Forest* se comporta, criando árvores e sub-árvores até um limite máximo em que não se tenha mais variável para empregar na criação de sub-árvores. Mas ainda assim, as variáveis de sazonalidade, 'estacao' e 'month', continuaram com um impacto baixo. Mesmo no modelo RF, essas variáveis do tipo categóricas não possuíram valor preditivo tanto quanto relevante as *lags* de vazão.

Figura 4.75 – Média dos valores SHAP - rio Jequitinhonha.



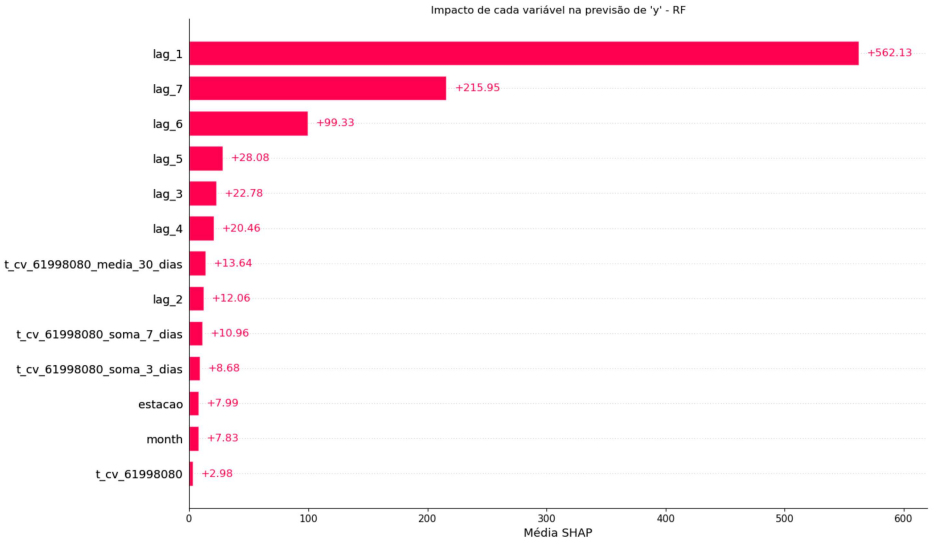
Fonte: O autor.

Figura 4.76 – Média dos valores SHAP - rio Doce.



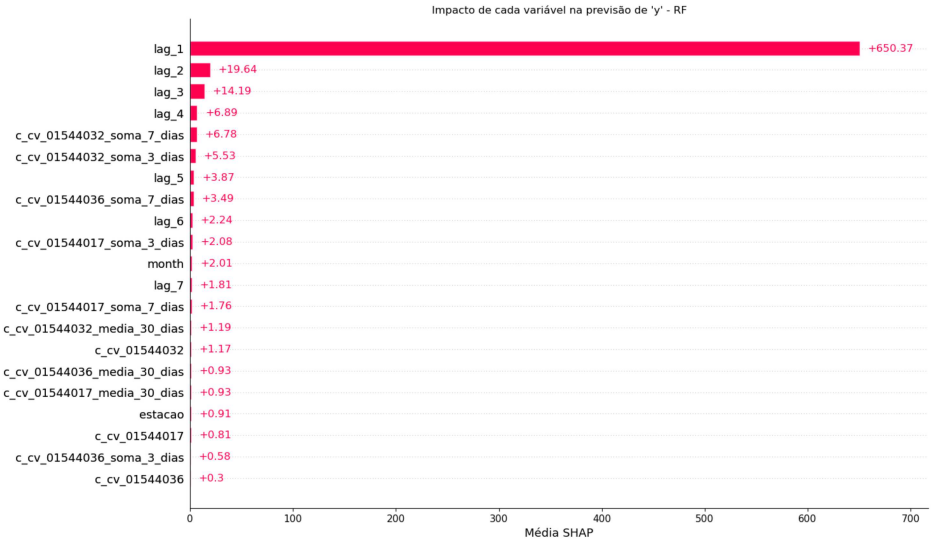
Fonte: O autor.

Figura 4.77 – Média dos valores SHAP - rio Grande.



Fonte: O autor.

Figura 4.78 – Média dos valores SHAP - rio São Francisco.

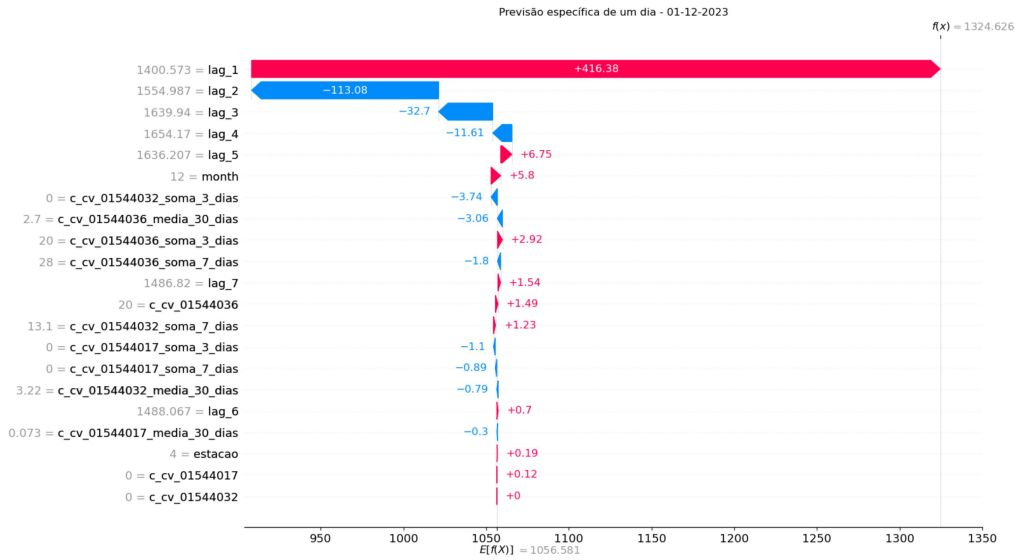


Fonte: O autor.

Começando por um valor base de $E[f(x)] = 1056,581$, após avaliados os impactos de cada variável, mais destacadamente as variáveis lag_1 e lag_2, a previsão final do modelo foi $f(x) = 1324,626$.

Esta previsão em específico corrobora com o gráfico 4.78, de que as variáveis mais importantes são as observações passadas da vazão. Mesmo as variáveis de tendência, de soma e média, não tiveram muito impacto na previsão. Neste modelo RF, assim como nos modelos anteriores, sempre as *lags* mostraram-se mais importantes para qualquer previsão e as demais variáveis apresentaram apenas um impacto marginal para o resultado.

Figura 4.79 – *Waterfall plot* de uma previsão específica no rio São Francisco.



Fonte: O autor.

Não foi incluído aqui o experimento com remoção de algumas variáveis, sejam as variáveis com média SHAP abaixo de 1 ou os atributos categóricos, conforme realizado anteriormente para os outros modelos. Isso deveu-se à pouca alteração no resultado, sendo mudanças na ordem de 10^{-4}, praticamente ínfimo, e desta forma procedeu-se com o seguimento do fluxo deste trabalho. Como o modelo *Random Forest* tem por característica a execução até o ponto de não haver mais subdivisões para novas árvores, e não houve mudanças nos hiperparâmetros do modelo, é provável que o melhor resultado possível, com pequenas oscilações, tenha sido alcançado averiguando todas as possibilidades.

4.3 ANÁLISE DOS RESULTADOS

A discussão sobre os resultados já foi amplamente debatida previamente, quando da apresentação dos resultados rio a rio. Aqui será feito um resumo sobre o comportamento geral dos modelos e a busca pelos motivos de porque, eventualmente, um modelo foi melhor em desempenho que outro. É bom mostrar, para cada rio, qual o modelo, ou quais os modelos, apresentaram melhor desempenho. A tabela 4.3 contém os resultados e destacado

os melhores. O negrito marca a melhor métrica e também o nome do modelo com melhor resultado.

Tabela 4.3 – Resultados dos modelos.

Rio Jequitinhonha				
Modelo	**KGE**	**MAPE**	**PBIAS**	**Cobertura empírica**
Linear Regression (log)	**0,95461**	0,14576	**-0,87104**	**98,08%**
CatBoost	0,88914	**0,13346**	−3,77912	89,32%
Random Forest	0,95193	0,14003	1,24504	83,84%
Rio Doce				
Modelo	**KGE**	**MAPE**	**PBIAS**	**Cobertura empírica**
Linear Regression (log)	**0,94579**	0,08434	−2,75287	**96,71%**
CatBoost	0,89330	**0,08207**	−3,73972	91,23%
Random Forest	0,93750	0,08355	**-1,44227**	88,77%
Rio Grande				
Modelo	**KGE**	**MAPE**	**PBIAS**	**Cobertura empírica**
Linear Regression	0,83365	0,13353	**-1,99105**	**84,38%**
CatBoost	0,76212	0,12605	−4,44499	63,56%
Random Forest	**0,83524**	**0,12300**	−2,18939	63,01%
Rio São Francisco				
Modelo	**KGE**	**MAPE**	**PBIAS**	**Cobertura empírica**
Linear Regression (log)	**0,98472**	0,02707	−0,98255	**96,71%**
CatBoost	0,93396	0,03296	−2,61473	90,68%
Random Forest	0,97754	**0,02571**	**-0,80974**	82,47%

Fonte: O autor.

Esta análise comparativa dos modelos de previsão de vazões nos rios avaliou o desempenho do LR, do CB e do RF em diferentes contextos hidrológicos. A avaliação considerou aspectos como eficiência preditiva, robustez frente a ruídos nos dados e capacidade de lidar com a complexidade das relações entre precipitação e vazão.

No que tange ao modelo LR, observou-se que este apresentou o melhor desempenho em três de quatro rios analisados. A aplicação da transformação logarítmica para linearização simplificou as relações complexas entre os dados de precipitação e vazão, relação esta bem estabelecida conforme discutido por Collischonn and Dornelles (12) no contexto do ciclo hidrológico, facilitando a modelagem. No entanto, o desempenho do LR foi significativamente influenciado pela presença de autocorrelação nas *lags* e pela multicolinearidade entre as variáveis. Esses fatores, combinados com a linearização dos dados, refletem a natureza dessa modelagem de séries temporais que emprega *lags* de vazão.

Em contraste, os modelos não-lineares CB e RF apresentaram desempenhos distintos. O RF se destacou ao obter os melhores resultados nos três aspectos aferidos pelo KGE e

no MAPE apenas para o rio Grande, seguido próximo pelo modelo LR. Este desempenho é particularmente notável considerando que os dados do rio Grande foram considerados os de pior qualidade, caracterizados por altos níveis de ruído e menor quantidade. Isso indica que o RF possui uma capacidade robusta de extrair conhecimento mesmo em condições adversas, superando desafios associados à qualidade dos dados.

Por outro lado, o modelo CB demonstrou um desempenho geral inferior em comparação ao RF. Embora o CB utilize o algoritmo *Ordered Boosting* e seja também um algoritmo baseado em árvores de decisão, tal qual o RF, cujas divisões se dão de maneira direcionada e eficiente, exploram menos o espaço de informações. Talvez seja neste aspecto que o RF supere o CB, pois aquele realiza subdivisões até o máximo, onde não se pode mais criar sub-árvores, diferentemente deste. No entanto, o CB extrai resultados mais rapidamente e com menor consumo de recursos computacionais, o que o torna poderoso e valioso no contexto da previsão de vazões. Este modelo apresentou resultados robustos que, com otimizações adicionais como a afinação de hiperparâmetros ou a remoção de variáveis de entrada por meio de técnicas como RFE, podem ser significativamente melhorados. Isso, inclusive, foi algo explorado, ainda que para outros modelos, em Reis et al. (78) e Duarte et al. (29). Um aspecto favorável do CB é sua elevada capacidade de personalização, com diversos hiperparâmetros que permitem melhorias tanto no processo de aprendizado quanto na complexidade das árvores geradas pelas subdivisões. Isso faz do CB um modelo potencialmente robusto dentre os avaliados, apesar de ainda necessitar de ajustes para alcançar um desempenho superior ao do RF.

Quando se observa as 5 variáveis de mais impacto nas previsões, em todos os modelos, percebe-se a presença de atributos de valor acumulado de precipitação em alguns rios. A partir de um ponto de vista físico, estas variáveis de soma e média podem ser representativas das condições de umidade do solo, provendo informações importantes para a previsão da resposta do escoamento superficial dos rios.(94) Cabe destacar que a permeabilidade do solo impacta na vazão dos rios devido à infiltração, demorada, e seus efeitos são percebidos no escoamento superficial com atraso – escoamento de base. (Collischonn and Dornelles (13), p.81, 129-130)

Os atributos categóricos não apresentaram a importância esperada para o trabalho. Esperava-se que estes atributos ajudassem os modelos a capturar os efeitos da sazonalidade, guardando a informação da estação e do mês em que se apresentassem as maiores vazões nas séries temporais. Mas este não foi o comportamento afinal, e os modelos ficaram concentrados nas variáveis contínuas de precipitação e vazão. De tal forma que é possível concluir que essas informações sobre sazonalidade não importaram para os modelos aplicados quando se trata de regressão e previsão em séries temporais de vazão a partir dos dados utilizados nesta pesquisa.

Refletindo de maneira geral sobre os modelos avaliados, todos demonstraram

resultados bastante satisfatórios, mantendo um comportamento estável ao longo de um extenso período de avaliação. A qualidade e a quantidade dos dados de treinamento influenciaram diretamente a qualidade dos resultados. A utilização da validação *walk-forward* revelou-se um método poderoso para analisar o comportamento e a estabilidade dos modelos em períodos extensos, simulando de forma eficaz a aplicação prática deste estudo e conferindo segurança e confiança nos resultados advindos dos modelos.

Em suma, a regressão linear mostrou-se altamente eficaz em contextos onde a relação entre precipitação e vazão é bem estabelecida e linearizada. Os modelos não-lineares, especialmente o RF, atuaram bem em cenários de dados de menor qualidade e maior ruído, enquanto o CB, apesar de ter apresentado um desempenho inferior, possui um alto potencial de melhoria mediante personalizações adicionais. Esses achados destacam a importância de selecionar e otimizar adequadamente os modelos de aprendizado de máquina conforme as características específicas dos dados e os objetivos da previsão hidrológica.

5 CONCLUSÃO E PERSPECTIVAS

5.1 CONCLUSÃO

Este trabalho teve como objetivo principal desenvolver e aplicar determinados modelos de aprendizado de máquina para previsão de vazão em 4 bacias hidrográficas de Minas Gerais. Com base nos resultados desta pesquisa, este objetivo macro foi conquistado e mostrou que é viável empregar estes modelos por órgãos de gestão competentes na realização de previsões diárias de vazão para tomada de decisões, gestão e distribuição de recursos.

Apesar de básica, a transformação log se mostrou bastante eficaz e tornou possível que um modelo simples como a Regressão Linear apresentasse resultados ótimos e que até mesmo pudesse superar modelos mais complexos como o *CatBoost* e o *Random Forest*. Mostra que é possível extrair valor do que é trivial no aprendizado de máquina. Então, neste sentido, este objetivo foi considerado realizado.

Se o que se busca é melhores previsões, é possível explorar outras transformações, mais profundas, tais como Box-Cox ou Yeo-Johnson. O problema destas transformações – de todas, a bem da verdade – é complicar a interpretação dos dados já que elas alteram a escala dos dados. É neste momento que um método que visa explicar a interação entre as variáveis e as previsões se mostra fundamental no trabalho, como é o caso do método SHAP. E particularmente, a análise de importância de variáveis, realizada aplicando este método, demonstrou a contento quais variáveis foram mais relevantes. A partir do que se apresentou aqui é possível, por exemplo, direcionar esforços no registro, armazenamento e obtenção das informações que sejam mais relevantes para a previsão de vazão.

5.2 CONTRIBUIÇÕES PARA A ÁREA

Esta pesquisa se insere em um nicho muito específico, uma lacuna apontada neste estudo a partir da revisão sistemática. Aplicar aprendizado de máquina para estas bacias de Minas Gerais mostra que é possível avançar mais sobre a compreensão das dinâmicas destas bacias a partir de uma ótica diferente, a partir de modelos de aprendizado de máquina. Os modelos físicos, contudo, não se tornam menos importantes. Ao contrário! O crescimento para a ciência dos recursos hídricos só se beneficia desta possível interação. Modelos de aprendizado de máquina podem prover de modo bastante eficiente, dados em regiões onde há uma anemia de informações e isso ser utilizado por modelos físicos, que realizam grandes simulações e muitas vezes de longo prazo – longo mesmo, anos, décadas e até mesmo séculos. E dados simulados obtidos a partir de modelos físicos podem alimentar modelos de aprendizado de máquina para análise de cenários mais contidos, de curto prazo, que visem o emprego, por exemplo, em um sistema de alertas. Existe nesta interação um

ciclo virtuoso de alimentação e retroalimentação entre estas distintas técnicas.

5.3 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

As recomendações a seguir incluem, mas não se restringem, aos seguintes apontamentos. Alguns itens servem também como melhoria a este próprio trabalho.

- Explorar outras técnicas de transformação dos dados, como Box-Cox e Yeo-Johnson;
- Aplicar outras variáveis físicas para treinamento dos modelos, como por exemplo a temperatura, conforme menção nos estudos (94) e (80), tipo de solo, tipo de cobertura do solo, ou mesmo índices climáticos, tal qual em Fernando De Toledo et al. (35);
- Explorar outros modelos mais complexos, como os diversos modelos de *deep learning* – redes LSTM, TFT, PatchTST e tantos outros;
- Outras estratégias de modelagem, que apliquem mais ou menos *lags* de vazão, que incluam mais variáveis de entrada a partir de técnicas de processamento de sinais, tais como *Wavelet* ou *Variational Mode Decomposition*, por exemplo, ou outras estatísticas;
- Realizar otimização de hiperparâmetros nos modelos *CatBoost* e *Random Forest* presentes neste trabalho;
- criar modelos híbridos específicos para as bacias analisadas, semelhante ao realizado no estudo Saraiva et al. (84).

As possibilidades são imensas. Existe um oceano de pesquisa e conhecimento a ser explorado. Há muito que se avançar.

Aqui se encerra uma misera gota.

REFERÊNCIAS

1 Portal de periódicos capes, 2024. URL https://www.periodicos.capes.gov.br/. Acessado em: maio de 2024.

2 Jhones da Silva Amorim, Marcelo Ribeiro Viola, Rubens Junqueira, Carlos Rogério de Mello, Nicole Lopes Bento, and Junior Cesar Avanzi. Quantifying the climate change-driven impacts on the hydrology of a data-scarce watershed located in the brazilian tropical savanna. *Hydrological Processes*, 36(7):e14638, 2022. doi: 10.1002/hyp.14638. URL https://doi.org/10.1002/hyp.14638.

3 Robert J. Belloto Jr. and Theodore D. Sokoloski. Residual analysis in regression. *American Journal of Pharmaceutical Education*, 49(Fall):295–303.

4 Chiara Bordin, Hans Ivar Skjelbred, Jiehong Kong, and Zhirong Yang. Machine learning for hydropower scheduling: State of the art and future research directions. *Procedia Computer Science*, 176:1659–1668, 2020. doi: 10.1016/j.procs.2020.09.190. URL https://doi.org/10.1016/j.procs.2020.09.190.

5 Agência Brasil. Um mês de calamidade: A cronologia dos alertas da tragédia no rs, junho 2024. URL https://agenciabrasil.ebc.com.br/geral/noticia/2024-06/um-mes-de-calamidade-a-cronologia-dos-alertas-da-tragedia-no-rs. Acessado em: agosto de 2024.

6 Agência Brasil. Mais de 78maio 2024. URL https://agenciabrasil.ebc.com.br/geral/noticia/2024-05/mais-de-78-dos-municipios-gauchos-foram-impactados-pelas-chuvas. Acesso em: junho de 2024.

7 BBC News Brasil. Chuvas na bahia: os fenômenos extremos que causam a tragédia no estado, julho 2024. URL https://www.bbc.com/portuguese/brasil-59804297. Acessado em: julho de 2024.

8 CNN Brasil. Temporais causam estragos em minas gerais e deixam desabrigados e desalojados, julho 2024. URL https://www.cnnbrasil.com.br/nacional/temporais-causam-estragos-em-minas-gerais-e-deixam-desabrigados-e-desalojados/. Acessado em: julho de 2024.

9 MapBiomas Brasil. Plataforma mapbiomas uso e cobertura, 2024. URL https://plataforma.brasil.mapbiomas.org/cobertura. Acessado em: junho de 2024.

10 Jason Brownlee. *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2018.

11 Taís Maria Nunes Carvalho, Francisco de Assis de Souza Filho, and Victor Costa Porto. Urban water demand modeling using machine learning techniques: Case study of fortaleza, brazil. *Journal of Water Resources Planning and Management*, 147(1): 05020026, 2020. doi: 10.1061/(ASCE)WR.1943-5452.0001310. URL https://doi.org/10.1061/(ASCE)WR.1943-5452.0001310.

12 Walter Collischonn and Fernando Dornelles. *Hidrologia para Engenharia e Ciências Ambientais*.

13 Walter Collischonn and Fernando Dornelles. *Hidrologia Para Engenharia e Ciências Ambientais*. ABRH, 2015. URL https://www.abrhidro.org.br/SGCv3/publicacao.php?PUB=5.

14 Comitê da Bacia Hidrográfica do Rio Doce. Comitê da bacia hidrográfica do rio doce (cbh-doce), 2024. URL https://www.cbhdoce.org.br. Acessado em: junho de 2024.

15 Comitê da Bacia Hidrográfica do Rio Grande. Comitê da bacia hidrográfica do rio grande (cbh-grande), 2024. URL https://www.cbhgrande.org.br. Acessado em: 10 de novembro de 2024.

16 Comitê da Bacia Hidrográfica do Rio São Francisco. Comitê da bacia hidrográfica do rio são francisco (cbhsf), 2024. URL https://cbhsaofrancisco.org.br. Acessado em: 10 de novembro de 2024.

17 David de Andrade Costa, Yared Bayissa, Mariana Dias Villas-Boas, Shreedhar Maskey, Jader Lugon Junior, Antônio José da Silva Neto, and Raghavan Srinivasan. Water availability and extreme events under climate change scenarios in an experimental watershed of the brazilian atlantic forest. *Science of The Total Environment*, 946: 174417, 2024. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2024.174417. URL https://www.sciencedirect.com/science/article/pii/S0048969724045650.

18 Wallisson Moreira de Carvalho. Hydrobr: A python package to work with brazilian hydrometeorological time series, julho 2020. URL http://doi.org/10.5281/zenodo.3931027. Version 0.1.1.

19 Instituto Brasileiro de Geografia e Estatística. Brasil | cidades e estados | ibge brasil |, 2024. URL https://cidades.ibge.gov.br/. Acessado em: junho de 2024.

20 Instituto Brasileiro de Geografia e Estatística (IBGE). Portal de mapas do ibge, 2024. URL https://portaldemapas.ibge.gov.br/portal.php#homepage. Acessado em: maio de 2024.

21 Instituto Mineiro de Gestão das Águas. Comitê da bacia hidrográfica do rio jequitinhonha - alto rio jequitinhonha, 2024. URL https://comites.igam.mg.gov.br/comites-estaduais-mg/jq1-cbh-do-alto-rio-jequitinhonha. Acessado em: junho de 2024.

22 Instituto Mineiro de Gestão das Águas. Comitê da bacia hidrográfica do rio araçuai, 2024. URL https://comites.igam.mg.gov.br/comites-estaduais-mg/jq2-cbh-do-rio-aracuai. Acessado em: junho de 2024.

23 Instituto Mineiro de Gestão das Águas. Comitê da bacia hidrográfica do rio jequitinhonha - médio e baixo rio jequitinhonha, 2024. URL https://comites.igam.mg.gov.br/comites-estaduais-mg/jq3-cbh-do-medio-e-baixo-rio-jequitinhonha. Acessado em: junho de 2024.

24 Pieter de Jong, Tarssio B. Barreto, Clemente A. S. Tanajura, Karla P. Oliveira-Esquerre, Asher Kiperstok, and Ednildo Andrade Torres. The impact of regional climate change on hydroelectric resources in south america. *Renewable Energy*, 173:76–91, 2021. doi: 10.1016/j.renene.2021.03.077. URL https://doi.org/10.1016/j.renene.2021.03.077.

25 M.F. de Sousa, E.M. Uliana, R.V.U. Aires, L.M.C. Rápalo, D.D. da Silva, M.C. Moreira, L. Lisboa, and D. da Silva Rondon. Streamflow prediction based on machine learning models and rainfall estimated by remote sensing in the brazilian savanna and amazon biomes transition. *Modeling Earth Systems and Environment*, 10(1):1191–1202, 2024. ISSN 23636203 (ISSN). doi: 10.1007/s40808-023-01837-9. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164936830&doi=10.1007%2fs40808-023-01837-9&partnerID=40&md5=5e197705f0663315a8e6f9304b4303d7.

26 Agência Nacional de Águas e Saneamento Básico (ANA). Catálogo de metadados - sistema nacional de informações sobre recursos hídricos (snirh), 2024. URL https://metadados.snirh.gov.br/geonetwork/srv/por/catalog.search. Acessado em: maio de 2024.

27 Helena Donato and Mariana Donato. Stages for undertaking a systematic review. *Acta Médica Portuguesa*, 32(3):227–235, 2019. doi: 10.20344/amp.11923. URL https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/11923.

28 Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

29 V.B.R. Duarte, M.R. Viola, M. Giongo, E.M. Uliana, and C.R. de Mello. Streamflow forecasting in tocantins river basins using machine learning. *Water Supply*, 22(7): 6230–6244, 2022. ISSN 16069749 (ISSN). doi: 10.2166/ws.2022.155. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135962633&doi=10.2166%2fws.2022.155&partnerID=40&md5=830850fe1ecf675dd24f3d7cd0b0204d.

30 Joaquin Amat Rodrigo e Javier Escobar Ortiz. skforecast, 8 2024a. URL https://skforecast.org/.

31 Joaquin Amat Rodrigo e Javier Escobar Ortiz. skforecast - probabilistic forecasting: prediction intervals and prediction distribution, 8 2024b. URL https://skforecast.org/0.13.0/user_guides/probabilistic-forecasting.

32 Empresa de Pesquisa Energética (Brasil). *Balanço Energético Nacional 2023: Ano base 2022 / Brazilian Energy Balance 2023 Year 2022*. Empresa de Pesquisa Energética (EPE), Rio de Janeiro, 2023. 274 p., 182 ill., 23 cm.

33 A.H. Essenfelder, F. Larosa, P. Mazzoli, S. Bagli, D. Broccoli, V. Luzzi, J. Mysiak, P. Mercogliano, and F. Dalla Valle. Smart climate hydropower tool: A machine-learning seasonal forecasting climate service to support cost–benefit analysis of reservoir management. *Atmosphere*, 11(12):1–21, 2020. ISSN 20734433 (ISSN). doi: 10.3390/atmos11121305. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097513615&doi=10.3390%2fatmos11121305&partnerID=40&md5=09568a53d1f6c99d30b09b7e09b64a58.

34 Nicolas Fernandez, Luis A. Camacho, and A. Pouyan Nejadhashemi. Modeling streamflow in headwater catchments: A data-based mechanistic grounded framework. *Journal of Hydrology: Regional Studies*, 44:101243, 2022. doi: 10.1016/j.ejrh.2022.101243. URL https://doi.org/10.1016/j.ejrh.2022.101243.

35 J. Fernando De Toledo, H. Valadares Siqueira, L.H. Biuk, R. Sacchi, R.D.R. Azambuja, R.A. Junior, and P.T.L. Asano. Climate indices impact in monthly streamflow series forecasting. *IEEE Access*, 11:21451–21464, 2023. ISSN 21693536 (ISSN). doi: 10.1109/ACCESS.2023.3237982. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147262025&doi=10.1109%2fACCESS.2023.3237982&partnerID=40&md5=e9ba6d0190de29d0d4a94dbc397097c5.

36 H.A.R. Filho, E.M. Uliana, U.R.V. Aires, I.F. da Cruz, L. Lisboa, D.D. da Silva, M.R. Viola, and V.B.R. Duarte. Nowcast flood predictions in the amazon watershed based on the remotely sensed rainfall product pdirnow and artificial neural networks. *Environmental Monitoring and Assessment*, 196(3), 2024. ISSN 01676369 (ISSN). doi: 10.1007/s10661-024-12396-6. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85184683766&doi=10.1007%2fs10661-024-12396-6&partnerID=40&md5=0f4e0611c1a6e493d366f10c85f3392c.

37 G1. Há 1 ano no volume morto, cantareira precisará de reserva até final de 2015, maio 2015. URL https://g1.globo.com/sao-paulo/noticia/2015/05/ha-1-ano-no-volume-morto-cantareira-precisara-de-reserva-ate-final-de-2015.html. Acessado em: julho de 2024.

38 G1. Temporal em petrópolis: entenda o que provocou as chuvas intensas que causaram destruição na cidade, fevereiro 2022. URL https://g1.globo.com/meio-ambiente/noticia/2022/02/15/temporal-em-petropolis-entenda-o-que-provocou-as-chuvas-intensas-que-causaram-destruicao-na-cidade.ghtml. Acessado em: julho de 2024.

39 G1. Temporal devastador no litoral norte de sp completa um mês; confira um resumo da tragédia, março 2023. URL https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2023/03/19/temporal-devastador-no-litoral-norte-de-sp-completa-um-mes-confira-um-resumo-da-tragedia.ghtml. Acessado em: agosto de 2024.

40 G1. A cronologia da tragédia no rio grande do sul, maio 2024. URL https://g1.globo.com/rs/rio-grande-do-sul/noticia/2024/05/12/a-cronologia-da-tragedia-no-rio-grande-do-sul.ghtml. Acessado em: agosto de 2024.

41 G1. Entenda o que causou temporal na região sul do es e o que pode ser feito para evitar novas tragédias, março 2024. URL https://g1.globo.com/es/espirito-santo/noticia/2024/03/27/entenda-o-que-causou-temporal-na-regiao-sul-do-es-e-o-que-pode-ser-feito-para-evitar-novas-tragedias.ghtml. Acessado em: julho de 2024.

42 Maria Cristiane Barbosa Galvão and Ivan Luiz Marques Ricarte. Revisão sistemática da literatura: Conceituação, produção e publicação. *Logeion: Filosofia da Informação*, 6(1):57–73, 2019. doi: 10.21728/logeion.2019v6n1.p57-73. URL https://revista.ibict.br/fiinf/article/view/4835.

43 Ehab Gomaa, Bilel Zerouali, Salah Difi, Khaled A. El-Nagdy, Celso Augusto Guimarães Santos, Zaki Abda, Sherif S.M. Ghoneim, Nadjem Bailek, Richarde Marques da Silva, Jitendra Rajput, and Enas Ali. Assessment of hybrid machine learning algorithms using trmm rainfall data for daily inflow forecasting in três marias reservoir, eastern brazil. *Heliyon*, 9(8):e18819, 8 2023. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2023.e18819. URL https://www.sciencedirect.com/science/article/pii/S2405844023060279.

44 Hoshin V. Gupta, Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91, 2009. doi: 10.1016/j.jhydrol.2009.08.003. URL https://doi.org/10.1016/j.jhydrol.2009.08.003.

45 Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

46 Hydrologic Engineering Center (HEC). Hec-hms: Hydrologic modeling system, 2024. URL https://www.hec.usace.army.mil/software/hec-hms/. Acessado em: agosto de 2024.

47 Rob J. Hyndman. Prediction intervals too narrow, 2014. URL https://robjhyndman.com/hyndsight/narrow-pi/. Acessado em: agosto de 2024.

48 Rob J. Hyndman and George Athanasopoulos. Forecasting: Principles and practice (3rd ed.) - 5.5 distributional forecasts and prediction intervals, 2024a. URL https://otexts.com/fpp3/prediction-intervals.html. Acessado em: junho de 2024.

49 Rob J. Hyndman and George Athanasopoulos. Forecasting: Principles and practice (3rd ed.) - 5.8 evaluating point forecast accuracy, 2024b. URL https://otexts.com/fpp3/accuracy.html. Acessado em: junho de 2024.

50 Rob J. Hyndman and George Athanasopoulos. Forecasting: Principles and practice (3rd ed.) - 7.2 least squares estimation, 2024c. URL https://otexts.com/fpp3/least-squares.html. Acessado em: junho de 2024.

51 Rob J. Hyndman and George Athanasopoulos. Forecasting: Principles and practice (3rd ed.) - 5.2 some simple forecasting methods, 2024d. URL https://otexts.com/fpp3/simple-methods.html. Acessado em: junho de 2024.

52 Rob J. Hyndman and George Athanasopoulos. Forecasting: Principles and practice (3rd ed.) - 2.8 autocorrelation models, 2024g. URL https://otexts.com/fpp3/acf.html. Acessado em: junho de 2024.

53 INMET. Gráficos climatológicos - inmet, 2024. URL https://clima.inmet.gov.br/GraficosClimatologicos/DF/83377. Acessado em: agosto de 2024.

54 IPCC. Climate change 2023: Synthesis report - summary for policymakers, 2023. URL https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_SPM.pdf. Acessado em: agosto de 2024.

55 Ramakar Jha, Vijay P. Singh, Vivekanand Singh, L. B. Roy, and Roshni Thendiyath. *Climate Change Impacts on Water Resources: Hydraulics, Water Resources and Coastal Engineering*, volume 98 of *Water Science and Technology Library*. Springer Nature Switzerland AG, Cham, Switzerland, 2021. ISBN 978-3-030-64201-3. doi: 10.1007/978-3-030-64202-0. URL https://doi.org/10.1007/978-3-030-64202-0.

56 Arthur Kolling Neto, Vinícius Alencar Siqueira, Cléber Henrique de Araújo Gama, Rodrigo Cauduro Dias de Paiva, Fernando Mainardi Fan, Walter Collischonn, Reinaldo Silveira, Cássia Silmara Aver Paranhos, and Camila Freitas. Advancing medium-range streamflow forecasting for large hydropower reservoirs in brazil by means of continental-scale hydrological modeling. *Water*, 15(9):1693, 2023. doi: 10.3390/w15091693. URL https://doi.org/10.3390/w15091693.

57 S.A.T. Korsic, C. Notarnicola, M.U. Quirno, and L. Cara. Assessing a data-driven approach for monthly runoff prediction in a mountain basin of the central andes of argentina. *Environmental Challenges*, 10, 2023. ISSN 26670100 (ISSN). doi: 10.1016/j.envc.2023.100680. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146455451&doi=10.1016%2fj.envc.2023.100680&partnerID=40&md5=6b9929d2fbd0400d251e433c6115e624.

58 Francesca Lazzeri. *Machine Learning for Time Series Forecasting with Python - CHAPTER 1: Overview of Time Series Forecasting*. John Wiley & Sons, Inc., 2021. ISBN 978-1-119-68237-0.

59 Wenbin Liu, Tao Yang, Fubao Sun, Hong Wang, Yao Feng, and Muye Du. Observation-constrained projection of global flood magnitudes with anthropogenic warming. *Water Resources Research*, 57(3):e2020WR028830, 2021. doi: 10.1029/2020WR028830. URL https://doi.org/10.1029/2020WR028830.

60 JEG Lopes, BPF Braga, and JGL Conejo. Smap–a simplified hydrological model. *Applied Modelling in Catchment Hydrology*, pages 167–176, 1982.

61 Zhu Lu and Sun Ligang. Predictive analysis of equipment health state based on time series model. pages 322–325, 2024. doi: 10.1109/SSAIC61213.2024.00067.

62 E. Lujano, R. Lujano, J.C. Huamani, and A. Lujano. Hydrological modeling based on the knn algorithm: An application for the forecast of daily flows of the ramis river, peru. *Tecnologia y Ciencias del Agua*, 14(2):169–203, 2023. ISSN 01878336 (ISSN). doi: 10.24850/j-tyca-14-02-05. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152258597&doi=10.24850%2fj-tyca-14-02-05&partnerID=40&md5=744953cc1a53a593a90d56358f6deb81.

63 Elena Macdonald, Bruno Merz, Björn Guse, Viet D. Nguyen, Xiaoxiang Guan, and Sergiy Vorogushyn. What controls the tail behaviour of flood series: Rainfall or runoff generation?, 2023.

64 Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

65 M. Meis, M.P. Llano, and D. Rodriguez. A statistical tool for a hydrometeorological forecast in the lower la plata basin. *International Journal of River Basin Management*, 21(4):685–696, 2023. ISSN 15715124 (ISSN). doi: 10.1080/15715124.2022.2079657. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134035138&doi=10.1080%2f15715124.2022.2079657&partnerID=40&md5=138136b5dc07dbe55b28cb4579ba98f4.

66 Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction*, pages 17–38, 2020. ISBN 978-3-030-57321-8. URL https://link.springer.com/chapter/10.1007/978-3-030-57321-8_2.

67 Christoph Molnar. Interpretability: Importance - interpretable machine learning, 2024. URL https://christophm.github.io/interpretable-ml-book/interpretability-importance.html. Acessado em: agosto de 2024.

68 Christoph Molnar. Shapley values - interpretable machine learning, 2024. URL https://christophm.github.io/interpretable-ml-book/shapley.html. Acessado em: agosto de 2024.

69 S.L. Neitsch, J.G. Arnold, J.R. Kiniry, J.R. Williams, and K.W. King. *Soil and Water Assessment Tool - Theoretical Documentation.* Texas Water Resources Institute, College Station, Texas, 2002. URL https://swat.tamu.edu/media/1290/swat2000theory.pdf. Acessado em: julho de 2024.

70 F.J.M. Nogueira Filho, F.A. Souza Filho, V.C. Porto, R.V. Rocha, Á.B.S. Estácio, and E.S.P.R. Martins. Deep learning for streamflow regionalization for ungauged basins: Application of long-short-term-memory cells in semiarid regions. *Water (Switzerland)*, 14(9), 2022. ISSN 20734441 (ISSN). doi: 10.3390/w14091318. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129356156&doi=10.3390%2fw14091318&partnerID=40&md5=50ec2b8b3a37ec7c4d6b1ff402572729.

71 Ana Paez-Trujilo, Jeffer Cañon, Beatriz Hernandez, Gerald Corzo, and Dimitri Solomatine. Multivariate regression trees as an explainable machine learning approach to explore relationships between hydroclimatic characteristics and agricultural and hydrological drought severity: Case of study cesar river basin. *Natural Hazards and Earth System Sciences*, 23:3863–3883, 2023. doi: 10.5194/nhess-23-3863-2023. URL https://doi.org/10.5194/nhess-23-3863-2023.

72 J.D. Pastrana-Cortés, D.A. Cardenas-Peña, M. Holguín-Londoño, G. Castellanos-Dominguez, and Á.A. Orozco-Gutiérrez. Multi-output variational gaussian process for daily forecasting of hydrological resources †. *Engineering Proceedings*, 39(1), 2023. ISSN 26734591 (ISSN). doi: 10.3390/engproc2023039083. URL https:

//www.scopus.com/inward/record.uri?eid=2-s2.0-85172783480&doi=10.3390%2fengproc2023039083&partnerID=40&md5=4ffbffe838bfc840e7d99c00dc61c96c.

73 José Marcelo Rodrigues Pereira, Clebson do Carmo Raimundo, Jr. Reis, Dirceu Silveira, Jr. Vasconcelos, Francisco das Chagas, and Eduardo Sávio Passos Rodrigues Martins. Use of climate information in water allocation: A case of study in a semiarid region. *Water*, 15(2460):1–21, 2023. doi: 10.3390/w15132460. URL https://doi.org/10.3390/w15132460.

74 Marcelo Portuguez-Maurtua, José Luis Arumi, Octavio Lagos, Alejandra Stehr, and Nestor Montalvo Arquiñigo. Filling gaps in daily precipitation series using regression and machine learning in inter-andean watersheds. *Water*, 14(11):1799, 2022. doi: 10.3390/w14111799. URL https://doi.org/10.3390/w14111799.

75 Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

76 Pablo Páliz Larrea, Xavier Zapata-Ríos, and Lenin Campozano Parra. Application of neural network models and anfis for water level forecasting of the salve faccha dam in the andean zone in northern ecuador. *Water*, 13(15):2011, 2021. doi: 10.3390/w13152011. URL https://doi.org/10.3390/w13152011.

77 QGIS Development Team. QGIS: A free and open source geographic information system, 2024. URL https://qgis.org/. Acessado em: setembro de 2024.

78 G.B. Reis, D.D. da Silva, E.I. Fernandes Filho, M.C. Moreira, G.V. Veloso, M.D.S. Fraga, and S.A.R. Pinheiro. Effect of environmental covariable selection in the hydrological modeling using machine learning models to predict daily streamflow. *Journal of Environmental Management*, 290, 2021. ISSN 03014797 (ISSN). doi: 10.1016/j.jenvman.2021.112625. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104810882&doi=10.1016%2fj.jenvman.2021.112625&partnerID=40&md5=7b2db7421ae0b6cdb6e680fcd95ab441.

79 V.H.A. Ribeiro, G. Reynoso-Meza, and H.V. Siqueira. Multi-objective ensembles of echo state networks and extreme learning machines for streamflow series forecasting. *Engineering Applications of Artificial Intelligence*, 95, 2020. ISSN 09521976 (ISSN). doi: 10.1016/j.engappai.2020.103910. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089817082&doi=10.1016%2fj.engappai.2020.103910&partnerID=40&md5=11b8155551004b9274cdc01f8416966a.

80 J.A.M. Rodrigues, A.C.O. Andrade, M.R. Viola, D.D. Ferreira, C.R. de Mello, and M.S. Thebaldi. Hydrological modeling in a basin of the brazilian cerrado biome. *Revista Ambiente e Agua*, 16(1):1–18, 2021. doi: 10.4136/ambi-agua.2639. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100339826&doi=10.4136%2fambi-agua.2639&partnerID=40&md5=0a32fa83271c05954cfa82e50b76f9f3.

81 José-Víctor Rodríguez, Ignacio Rodríguez-Rodríguez, and Wai Lok Woo. Machine learning-based prediction of sunspots using fourier transform analysis of the time series. *Publications of the Astronomical Society of the Pacific*, 134(1042):124201, dec 2022. doi: 10.1088/1538-3873/aca4a3. URL https://dx.doi.org/10.1088/1538-3873/aca4a3.

82 Rafael Rodríguez, Marcos Pastorini, Lorena Etcheverry, Christian Chreties, Mónica Fossati, Alberto Castro, and Angela Gorgoglione. Water-quality data imputation with a high percentage of missing values: A machine learning approach. *Sustainability*, 13 (11):6318, 2021. doi: 10.3390/su13116318. URL https://www.mdpi.com/2071-1050/13/11/6318.

83 Hasan Ahmed Salman, Ali Kalakech, and Amani Steiti. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024:69–79, 2024. doi: 10.58496/bjml/2024/007.

84 S.V. Saraiva, F.D.O. Carvalho, C.A.G. Santos, L.C. Barreto, and P.K.D.M.M. Freire. Daily streamflow forecasting in sobradinho reservoir using machine learning models coupled with wavelet transform and bootstrapping. *Applied Soft Computing*, 102, 2021. ISSN 15684946 (ISSN). doi: 10.1016/j.asoc.2021.107081. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85099235295&doi=10.1016%2fj.asoc.2021.107081&partnerID=40&md5=dd7e12f4e30a2f51bc00bd0db172c37b.

85 Scikit-learn. Randomforestregressor - scikit-learn documentation, 2024. URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Acessado em: abril de 2024.

86 Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python, 2010.

87 Inovação e Insumos Estratégicos em Saúde Secretaria de Ciência, Tecnologia. *Diretrizes metodológicas: elaboração de revisão sistemática e meta-análise de ensaios clínicos randomizados*. Ministério da Saúde, Brasília, 2021. ISBN 978-65-5993-021-0. URL http://bvsms.saude.gov.br/bvs/publicacoes/diretrizes_elaboracao_revisao_sistematica_meta-analise.pdf.

88 Shivam Singh and Manish Kumar Goyal. Enhancing climate resilience in businesses: The role of artificial intelligence. *Journal of Cleaner Production*, 418:138228, 2023. doi: 10.1016/j.jclepro.2023.138228. URL https://doi.org/10.1016/j.jclepro.2023.138228.

89 Vinícius Alencar Siqueira, Fernando Mainardi Fan, Rodrigo Cauduro Dias de Paiva, Maria-Helena Ramos, and Walter Collischonn. Potential skill of continental-scale, medium-range ensemble streamflow forecasts for flood prediction in south america. *Journal of Hydrology*, 590:125430, 2020. doi: 10.1016/j.jhydrol.2020.125430. URL https://doi.org/10.1016/j.jhydrol.2020.125430.

90 Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH). Portal de metadados snirh - pesquisa por ottobacia, 2024. URL https://metadados.snirh.gov.br/geonetwork/srv/search?keyword=Ottobacia. Acessado em: abril de 2024.

91 Stefano Frizzo Stefenon, Laio Oriel Seman, Luiza Scapinello Aquino da Silva, Viviana Cocco Mariani, and Leandro dos Santos Coelho. Hypertuned temporal fusion transformer for multi-horizon time series forecasting of dam level in hydroelectric power plants. *International Journal of Electrical Power and Energy Systems*, 157: 109876, 2024. doi: 10.1016/j.ijepes.2024.109876. URL https://doi.org/10.1016/j.ijepes.2024.109876.

92 F.L.R. Torres, L.M.M. Lima, M.S. Reboita, A.R. de Queiroz, and J.W.M. Lima. Integrating hydrological and machine learning models for enhanced streamflow forecasting via bayesian model averaging in a hydro-dominant power system. *Water (Switzerland)*, 16(4), 2024. ISSN 20734441 (ISSN). doi: 10.3390/w16040586. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85185717688&doi=10.3390%2fw16040586&partnerID=40&md5=a00e3a48688225a735f0271d57da725e.

93 R.S. Vilanova, S.S. Zanetti, and R.A. Cecílio. Artificial neural networks-based model parameter transfer in streamflow simulation of brazilian atlantic rainforest watersheds. *Journal of Hydrologic Engineering*, 25(7), 2020. ISSN 10840699 (ISSN). doi: 10.1061/(ASCE)HE.1943-5584.0001947. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084632633&doi=10.1061%2f%28ASCE%29HE.1943-5584.0001947&partnerID=40&md5=811109b65a634772ca7366f5db87c1fb.

94 F. Vilaseca, A. Castro, C. Chreties, and A. Gorgoglione. Assessing influential rainfall–runoff variables to simulate daily streamflow using random forest. *Hydrological Sciences Journal*, 68(12):1738–1753, 2023. ISSN 02626667 (ISSN). doi: 10.1080/02626667.2023.2232356. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165877308&doi=10.1080%2f02626667.2023.2232356&partnerID=40&md5=07f390b3facd2b82cd1a3c1cda2c5f7d.

95 Yan Yan, Huan Wu, Guojun Gu, Philip J. Ward, Lifeng Luo, Xiaomeng Li, Zhijun Huang, and Jing Tao. Exploring the enso impact on basin-scale floods using hydrological simulations and trmm precipitation. *Geophysical Research Letters*, 47(22): e2020GL089476, 2020. doi: 10.1029/2020GL089476. URL https://doi.org/10.1029/2020GL089476.

96 Yandex. *CatBoost Documentation*, 2024. URL https://catboost.ai/en/docs/. Acessado em: junho de 2024.