

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM
COMPUTACIONAL

Welson de Avelar Soares Filho

**Aprendizado de máquina com dados categóricos na modelagem chuva-vazão
para previsão de vazão em bacias hidrográficas de Minas Gerais**

Juiz de Fora

2024

Welson de Avelar Soares Filho

**Aprendizado de máquina com dados categóricos na modelagem chuva-vazão
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 dados categóricos na modelagem chuva-vazão para previsão de vazão em bacias hidrográficas de Minas Gerais / Welson de Avelar Soares Filho. – 2024.

72 f. : il.

Orientador: Leonardo Goliatt

Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Modelagem Computacional, 2024.

1. recursos hídricos. 2. redes neurais. 3. previsão de vazão. I. Goliatt, Leonardo, orient. II. Doutor.

Welson de Avelar Soares Filho

**Aprendizado de máquina com dados categóricos na modelagem chuva-vazão
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 (dia) de (mês) de (ano)

BANCA EXAMINADORA

Doutor Leonardo Goliatt - Orientador
Universidade Federal de Juiz de Fora

Titulação Nome e sobrenome
Universidade ???

Titulação Nome e sobrenome
Universidade ??

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 apoio tive neste momento da vida.

Minha noiva, 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 amigos 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. Tivemos conversas apenas por aplicativos, distantes, e anseio poder rever seus olhos.

Aos meus colegas do PGMGC: 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.

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.

Ao professor Celso Bandeira e à colega de projeto Paula pelas conversas e 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 e à Agência Nacional de Águas e Saneamento Básico (ANA) pelo hercúleo trabalho desenvolvido na gestão e conhecimento de nossas águas.

Meu muito obrigado.

RESUMO

Resumo do trabalho

Palavras-chave: Aprendizado de máquina. Recursos hídricos. Previsão de vazão.

ABSTRACT

Project summary

Keywords: Machine learning. Water resources. Runoff forecasting.

LISTA DE ILUSTRAÇÕES

Figura 3.1–Série temporal incompleta da estação t_vz_54790000 (fonte: o autor)	27
Figura 3.2–Detalhe da série temporal da estação t_vz_54790000, ainda sem dados imputados, de 2013 a 2016 (fonte: o autor)	27
Figura 3.3–Detalhe da série temporal da estação t_vz_54790000, com dados imputados, de 2013 a 2016 (fonte: o autor)	27
Figura 3.4–Série temporal incompleta da estação t_vz_54790000 no detalhe entre 2021 e 2022 (fonte: o autor)	28
Figura 3.5–Série temporal completa da estação t_vz_54790000 no detalhe entre 2021 e 2022 (fonte: o autor)	28
Figura 3.6–Série temporal completa da estação t_vz_54790000 (fonte: o autor)	28
Figura 3.7–Série temporal incompleta da estação t_cv_54790000 (fonte: o autor)	29
Figura 3.8–Série temporal completa da estação t_cv_54790000 (fonte: o autor)	29
Figura 3.9–Série temporal completa da estação t_cv_01640000 (fonte: o autor)	29
Figura 3.10–Série temporal completa da estação c_vz_56994500 (fonte: o autor)	30
Figura 3.11–Série temporal da estação t_cv_56990850 - não utilizada (fonte: o autor)	31
Figura 3.12–Série temporal da estação t_cv_56994500 - não utilizada (fonte: o autor)	31
Figura 3.13–Série temporal completa da estação c_cv_01941010 (fonte: o autor)	31
Figura 3.14–Série temporal completa da estação c_cv_01941004 (fonte: o autor)	32
Figura 3.15–Série temporal completa da estação c_cv_01941006 (fonte: o autor)	32
Figura 3.16–Série temporal completa da estação t_cv_56990005 (fonte: o autor)	32
Figura 3.17–Série temporal incompleta da estação t_vz_62020080 (fonte: o autor)	33
Figura 3.18–Série temporal completa da estação t_vz_62020080 (fonte: o autor)	34

Figura 3.19–Série completa da estação t_cv_61998080	
(fonte: o autor)	34
Figura 3.20–Detalhe do trecho com dados nulos da estação c_vz_44290002	
(fonte: o autor)	35
Figura 3.21–Série temporal completa da estação c_vz_44290002	
(fonte: o autor)	35
Figura 3.22–Série temporal completa da estação c_cv_01544017	
(fonte: o autor)	36
Figura 3.23–Série temporal completa da estação c_cv_01544032	
(fonte: o autor)	36
Figura 3.24–Série temporal completa da estação c_cv_01544036	
(fonte: o autor)	36
Figura 3.25–Autocorrelação para a vazão do rio Jequitinhonha	
(fonte: o autor)	40
Figura 3.26–Componente sazonal da série de vazão do rio Jequitinhonha	
(fonte: o autor)	40
Figura 3.27–Autocorrelação para a vazão do rio Doce	
(fonte: o autor)	41
Figura 3.28–Componente sazonal da série de vazão do rio Doce	
(fonte: o autor)	41
Figura 3.29–Autocorrelação para a vazão do rio Grande	
(fonte: o autor)	42
Figura 3.30–Componente sazonal da série de vazão do rio Grande	
(fonte: o autor)	42
Figura 3.31–Autocorrelação para a vazão do rio São Francisco	
(fonte: o autor)	43
Figura 3.32–Componente sazonal da série de vazão do rio São Francisco	
(fonte: o autor)	43
Figura 3.33–Dados originais para o rio Jequitinhonha	
(fonte: o autor)	44
Figura 3.34–Dados log-transformados para o rio Jequitinhonha	
(fonte: o autor)	44
Figura 3.35–Dados originais para o rio Doce	
(fonte: o autor)	44
Figura 3.36–Dados log-transformados para o rio Doce	
(fonte: o autor)	45
Figura 3.37–Dados originais para o rio Grande	
(fonte: o autor)	45

Figura 3.38–Dados log-transformados para o rio Grande	
(fonte: o autor)	45
Figura 3.39–Dados originais para o rio São Francisco	
(fonte: o autor)	46
Figura 3.40–Dados log-transformados para o rio São Francisco	
(fonte: o autor)	46
Figura 3.41–Diagrama mostrando a divisão dos dados de treino/teste com <i>refit</i> .	
(fonte: (13))	52
Figura 3.42–WFV com janela expandida e <i>refit</i> - imagem 1.	
(fonte: (13))	52
Figura 3.43–WFV com janela expandida e <i>refit</i> - imagem 2.	
(fonte: (13))	53
Figura 3.44–WFV com janela expandida e <i>refit</i> - imagem 3.	
(fonte: (13))	53
Figura 3.45–Diagrama de como se calcula o intervalo de previsão.	
(fonte: (14))	53
Figura 3.46–Previsão com valor fora dos limites	
(fonte: o autor)	54
Figura 3.47–Previsão completamente nos limites	
(fonte: o autor)	55
Figura 3.48–Fluxo de trabalho	
(fonte: o autor)	57
Figura 4.1–Resultado do SeasonalNaive no teste <i>Walk-Forward Validation</i>	
(fonte: o autor)	59
Figura 4.2– <i>Walk-Forward Validation</i> para o modelo Regressão Linear	
(fonte: o autor)	60
Figura 4.3–Detalhe do trecho final dos dados, em 2022, usados para treinamento.	
(fonte: o autor)	61
Figura 4.4–Resíduos da previsão versus os valores das previsões.	
(fonte: o autor)	61
Figura 4.5–Resíduos da previsão ao longo do tempo.	
(fonte: o autor)	62
Figura 4.6–Histograma dos resíduos.	
(fonte: o autor)	63
Figura 4.7–Resíduos da previsão ao longo do tempo.	
(fonte: o autor)	63
Figura 4.8– <i>Walk-Forward Validation</i> para o modelo Regressão Linear.	
(fonte: o autor)	65

Figura 4.9– Dispersão dos resíduos.
(fonte: o autor) 65

Figura 4.10– Dispersão dos resíduos ao longo do ano.
(fonte: o autor) 66

Figura 4.11– Histograma e curva-normal dos resíduos.
(fonte: o autor) 66

Figura 4.12– Gráfico ACF dos resíduos.
(fonte: o autor) 66

LISTA DE TABELAS

Tabela 3.1 – Estações usadas no rio Jequitinhonha	
(fonte: o autor)	23
Tabela 3.2 – Estações usadas no rio Doce	
(fonte: o autor)	23
Tabela 3.3 – Estações usadas no rio Grande	
(fonte: o autor)	24
Tabela 3.4 – Estações usadas no rio São Francisco	
(fonte: o autor)	24
Tabela 3.5 – Estações de precipitação usadas - final	
(fonte: o autor)	30
Tabela 3.6 – Variáveis utilizadas - rio Jequitinhonha	
(fonte: o autor)	37
Tabela 3.7 – Variáveis utilizadas - rio Doce	
(fonte: o autor)	37
Tabela 3.8 – Variáveis utilizadas - rio Grande	
(fonte: o autor)	37
Tabela 3.9 – Variáveis utilizadas - rio São Francisco	
(fonte: o autor)	37
Tabela 3.10–Variáveis acumuladas para as precipitações	
(fonte: o autor)	38
Tabela 3.11–Hiperparâmetros para o modelo CatBoost	
(fonte: o autor)	49
Tabela 3.12–Hiperparâmetros para o modelo RandomForest	
(fonte: o autor)	50
Tabela 3.13–Configurações do computador usado nos experimentos	
(fonte: o autor)	56

LISTA DE ABREVIATURAS E SIGLAS

ANA	Agência Nacional de Águas e Saneamento Básico
Fil.	Filosofia
IBGE	Instituto Brasileiro de Geografia e Estatística
INMETRO	Instituto Nacional de Metrologia, Normalização e Qualidade Industrial

LISTA DE SÍMBOLOS

\forall	Para todo
\in	Pertence

SUMÁRIO

1	INTRODUÇÃO	17
1.1	Contextualização	18
1.2	Justificativa	18
1.3	Problema de pesquisa	18
1.4	Objetivos	18
1.5	Estrutura da Dissertação	18
2	REVISÃO DA LITERATURA SOBRE O TEMA	19
2.1	Conceitos Teóricos Fundamentais	19
2.2	Estudos Relacionados	19
2.3	Aprendizado de Máquina na Hidrologia	19
2.4	Dados de Precipitação e Vazão	19
2.5	Modelos de Previsão de Vazão	19
3	PROCEDIMENTOS METODOLÓGICOS	20
3.1	Descrição da Área de Estudo	20
3.2	Dados Utilizados	21
3.3	Pré-processamento dos Dados	25
3.3.1	Rio Jequitinhonha	26
3.3.2	Rio Doce	30
3.3.3	Rio Grande	33
3.3.4	Rio São Francisco	35
3.4	Variáveis Utilizadas	37
3.5	Análise exploratória dos dados	39
3.6	Modelos de Aprendizado de Máquina	47
3.6.1	Seasonal Naive - SN	47
3.6.2	Regressão Linear - LR	47
3.6.3	CatBoost - CB	48
3.6.4	Floresta Aleatória - RF	49
3.7	Métricas de Avaliação	50
3.8	Modelo proposto	51
4	RESULTADOS E DISCUSSÃO	58
4.1	Desempenho dos modelos	58
4.1.1	Rio Jequitinhonha	58
4.2	Importância das variáveis	67
4.3	Discussão dos resultados	67
5	CONCLUSÃO E PERSPECTIVAS	68
5.1	Conclusão	68
5.2	Contribuições para a área	68

5.3	Recomendações para Trabalhos Futuros	68
	REFERÊNCIAS	69

1 INTRODUÇÃO

A gestão dos recursos hídricos desempenha um papel crucial nas políticas públicas. Nos âmbitos socioeconômico, cultural e de saúde pública, conhecer a dinâmica dos recursos hídricos e entender como fatores externos impactam seu comportamento é de grande importância para os administradores 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, especialmente em regiões onde a hidroeletricidade desempenha um papel crucial na matriz energética, como é o caso do Brasil. De acordo com 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 (15). Desta forma, a previsão da vazão dos 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 mais 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, mas também a chuva têm impacto significativo nesse cenário. Uma análise criteriosa da previsibilidade da vazão dos rios nos dias seguintes permite ao poder público, por exemplo, reduzir ou até mesmo suspender outorgas para retirada de água, visando o bem-estar populacional. Além disso, essa análise pode auxiliar no planejamento de regimes de racionamento. Basta lembrarmos do ano de 2015, quando noticiava-se 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.(16)

E quando se fala em bem-estar populacional, não podemos deixar de considerar os eventos climáticos extremos.

Basta lembrar dos últimos desastres ocorridos nos estados de Minas Gerais, Rio de Janeiro, Paraná, São Paulo e Bahia que trouxeram não só perda econômica como também perda de vidas humanas. (3) (4) (18) (17)

Especificamente no estado de Minas Gerais há lacunas de conhecimento sobre os processos hidrológicos das bacias hidrográficas

tem
refe-
rencia
para
esta
parte?

seria
bom
colo-
car al-
guma
coisa

referenc

aspas
no la-
tex é
feito
assim
"exem-
plo
de as-
pas",
deixe
as as-
pas
sim-
ples
para
as va-
riáveis
utili

1.1 Contextualização

1.2 Justificativa

1.3 Problema de pesquisa

1.4 Objetivos

Separa objetivo e objetivo específico em seções

O trabalho tem como escopo a modelagem de recursos hídricos baseada em dados históricos de vazão, contrastando com os modelos físicos tradicionalmente aplicados na área, como o modelo SMAP (Soil Moisture Accounting Procedure) e o SWAT (Soil and Water Assessment Tool). Ao invés de utilizar as abordagens físicas que simulam processos hidrológicos baseados em equações diferenciais e características fisiográficas das bacias, a proposta é aplicar métodos de modelagem orientados a dados para prever séries temporais univariadas de vazão. Essas previsões serão realizadas com o suporte de variáveis exógenas, como precipitação, e variáveis categóricas, a serem melhor descritas a posteriori, que ajudam a capturar a sazonalidade e outros padrões importantes do regime de águas.

O objetivo específico deve vir em tópicos e em uma subseção do objetivo

O objetivo mais específico é desenvolver um modelo de previsão eficiente e preciso que possa ser utilizado para auxiliar na gestão dos recursos hídricos em Minas Gerais. Em uma etapa posterior, pretende-se criar uma aplicação web para disponibilizar essas previsões ao Instituto Mineiro de Gestão das Águas (IGAM (27)), permitindo que as informações estejam acessíveis para planejamento e tomada de decisões estratégicas.

Este trabalho se insere como um componente fundamental de um sistema gerencial maior, que visa aprimorar o planejamento e a gestão dos recursos hídricos do estado de Minas Gerais. Ao integrar previsões baseadas em dados históricos, o modelo permitirá uma melhor alocação dos recursos, suporte em períodos críticos, como secas e enchentes, e uma gestão mais eficiente das bacias hidrográficas do estado.

1.5 Estrutura da Dissertação

não é muito usual, alguns membros da banca podem não gostar, verificar com o léo se cabe esta parte na sua dissertação

Resumo breve da organização dos capítulos.

esta parte me parece mais uma descrição do trabalho do que um objetivo

2 REVISÃO DA LITERATURA SOBRE O TEMA

2.1 Conceitos Teóricos Fundamentais

Conceitos e teorias básicas relacionadas ao tema.

2.2 Estudos Relacionados

Revisar literatura existente, destacar pesquisas similares e identificar lacunas.

2.3 Aprendizado de Máquina na Hidrologia

Discutir a aplicação de técnicas de ML na hidrologia, citando estudos relevantes.

2.4 Dados de Precipitação e Vazão

Descrever os tipos de dados utilizados no estudo e respectivas fontes.

2.5 Modelos de Previsão de Vazão

Comparar diferentes modelos de previsão de vazão (vou comparar com Seasonal-Naive e Linear Regression, duas baselines comumente aplicadas).

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Descrição da Área de Estudo

FAZER IMAGEM DAS BACIAS

As bacias de estudo são as bacias dos rios Jequitinhonha, Doce, Grande e São Francisco. Em todas estas, o corpo d'água principal da bacia, seus respectivos rios, passam dentro do estado de Minas Gerais, por isso a escolha. Isso é para distinguir do não uso do rio Paraíba do Sul no estudo, visto que parte de sua bacia está nos limites do estado, mas o rio não passa dentro dos limites administrativos estaduais.

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.(8) (9) (10) (11)

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.(5)

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.

A predominância da Mata Atlântica, com uma pequena fração de Cerrado, confere à bacia uma vasta biodiversidade, ao mesmo tempo em que a coloca 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.

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

1286 km até seu encontro com o Paranaíba. O Cerrado é o bioma predominante em sua bacia, com trechos que se limitam à Mata Atlântica, criando uma rica diversidade de ecossistemas.

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.

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.

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.

3.2 Dados Utilizados

FAZER IMAGEM DE ONDE ESTÃO AS ESTAÇÕES

A coleta dos dados foi realizada com um misto da biblioteca HydroBR (7) e funções próprias de extração de dados. A biblioteca permitiu a listagem de todas as estações hidrométricas disponíveis, como, por exemplo, as estações convencionais de medição de vazão. 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.

O período de dados analisado compreende **de 1º de janeiro de 2013 a 31 de dezembro de 2023**, totalizando 11 anos completos.

Foram utilizadas **séries temporais diárias** de precipitação e vazão. As colunas correspondentes às datas foram formatadas como “*datetime*”, enquanto os dados de precipitação e vazão foram representados como valores de ponto flutuante (“*float*”). 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. Estes aspectos serão discutidos em detalhes em seções subsequentes.

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 que cai sobre uma unidade de área 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 consistentes pela ANA, ou seja, foram medidos e validados pela agência. O presente trabalho não questionou a veracidade dos dados; eles foram utilizados conforme disponibilizados pela ANA.

É relevante destacar que a consulta prévia ao sistema *on-line* da ANA foi essencial, pois frequentemente selecionavam-se códigos de estação que, ao final, não possuíam dados para o período especificado ou apresentavam códigos alterados na base de dados, sendo retornados como “inexistentes”. Quando um código de estação não retornava resultados na consulta ao sistema, foi necessário utilizar o sistema gentilmente cedido pela Rhama Analysis para verificar se o código da estação havia sido modificado. Nos casos em que se constatava a alteração, o novo código foi adotado, enquanto o código anteriormente informado como inexistente foi descartado.

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.

Por fim, é importante destacar a existência de estações híbridas, classificadas como “pluviométricas/fluviométricas”. Em alguns casos, o código da estação pode indicar que

se trata de uma estação de vazão (com códigos iniciados em 5 ou 6, por exemplo), mas que também possui informações de precipitação. O inverso também ocorre, onde códigos indicam estações pluviométricas (com códigos iniciados em 016 ou 019, por exemplo) que, no entanto, contêm dados de vazão. Para garantir a consistência com a nomenclatura utilizada pela ANA, manteve-se a classificação original das estações, mesmo que estas contenham apenas dados de precipitação ou vazão.

Para facilitar a visualização, as estações de vazão e precipitação utilizadas no trabalho são apresentadas abaixo.

Tabela 3.1 – Estações usadas no rio Jequitinhonha
(fonte: o autor)

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

Tabela 3.2 – Estações usadas no rio Doce
(fonte: o autor)

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	<i>Colatina ponte</i>	<i>Colatina</i>	<i>-19,5333</i>	<i>-40,6297</i>

Tabela 3.3 – Estações usadas no rio Grande
(fonte: o autor)

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

Tabela 3.4 – Estações usadas no rio São Francisco
(fonte: o autor)

Convencionais				
Fluviométricas				
Código	Nome	Município	Latitude	Longitude
44290002	<i>Pedras de Maria da Cruz</i>	<i>Pedras de Maria da Cruz</i>	-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

É importante 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.

3.3 Pré-processamento dos Dados

Com os dados disponíveis localmente, o primeiro passo antes de qualquer análise foi garantir a continuidade temporal dos mesmos. Existiam dias faltantes, e, para garantir uma linha do tempo contínua, foi necessário preencher essas lacunas. Os 11 anos de dados diários resultaram em um total de 4017 linhas de dados após essa etapa.

A sazonalidade é um fenômeno bem conhecido e estabelecido na análise hidrológica das bacias hidrográficas da América do Sul. 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 nos rios. (32)

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 realizou-se uma nova contagem dos dados que ainda permaneciam faltantes. Para esses casos nulos, foi aplicada a imputação de dados utilizando o modelo kNN (k-Nearest Neighbors - k-vizinhos mais próximos), com o objetivo de 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. 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.

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, de tendência e estacionariedade. 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.

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’, conseqüentemente, 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 de vazão, código 54790000’.

3.3.1 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 uma clara sazonalidade na série, com picos de vazão ocorrendo predominantemente no final e início de cada ano (figura 3.1). As figuras (3.1), (3.6), (3.2), (3.3), (3.4) e (3.5) 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. Cabe lembrar que, inicialmente, foi aplicada a média dos últimos três anos para replicar a sazonalidade. Para os dados que permaneceram ausentes, utilizou-se o modelo kNN para completar a série. 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.

A seguir, destaca-se o trecho da série com a maior quantidade de dados faltantes (figura 3.2), que abrange o período de janeiro de 2013 a janeiro de 2016. Em sequência, é apresentada a série após a imputação dos dados (figura 3.3). 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 finais da análise, já que se referem a um período distante do foco principal deste estudo. Uma

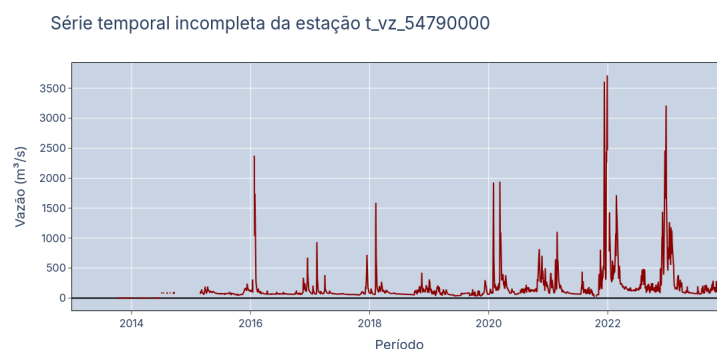


Figura 3.1 – Série temporal incompleta da estação t_vz_54790000
(fonte: o autor)

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 nos resultados.

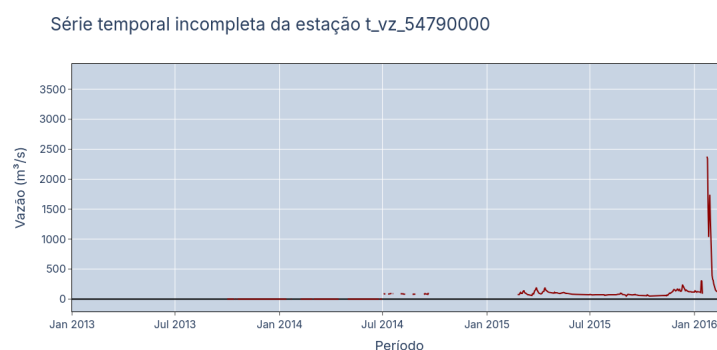


Figura 3.2 – Detalhe da série temporal da estação t_vz_54790000, ainda sem dados imputados, de 2013 a 2016 (fonte: o autor)



Figura 3.3 – Detalhe da série temporal da estação t_vz_54790000, com dados imputados, de 2013 a 2016 (fonte: o autor)

Observe também o trecho de dados faltantes mais próximo ao final dos anos analisados, em 2021 e 2022 (figura 3.4). Esta porção da série ficou boa visto que havia

informação prévia suficiente, a inserção de dados respeitou coerentemente a sazonalidade (figura 3.5).



Figura 3.4 – Série temporal incompleta da estação t_vz_54790000 no detalhe entre 2021 e 2022 (fonte: o autor)

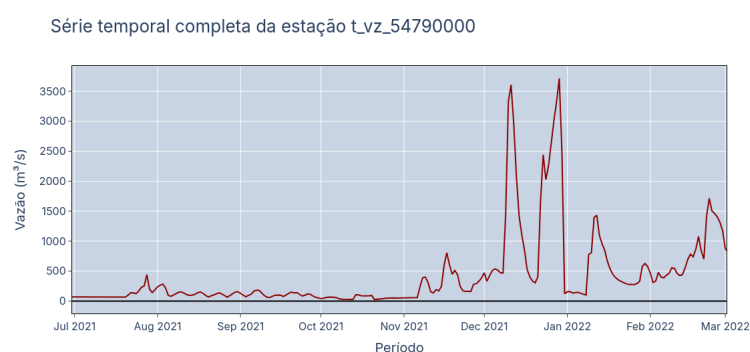


Figura 3.5 – Série temporal completa da estação t_vz_54790000 no detalhe entre 2021 e 2022 (fonte: o autor)

Por fim, uma visão ampla de como ficou a série temporal após os procedimentos de imputar os dados. (figura 3.6)

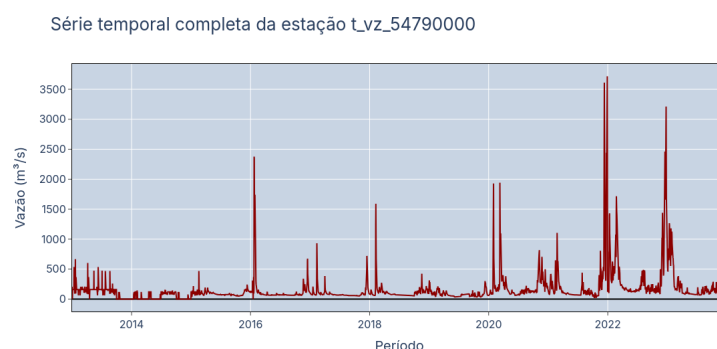


Figura 3.6 – Série temporal completa da estação t_vz_54790000 (fonte: o autor)

A mesma análise foi realizada para as estações de chuva. Na estação t_cv_54790000 (figura 3.7) faltavam 273 dias de dados (6,79%). Já a estação t_cv_01640000 estava totalmente preenchida, sem valores nulos. (figura 3.9)

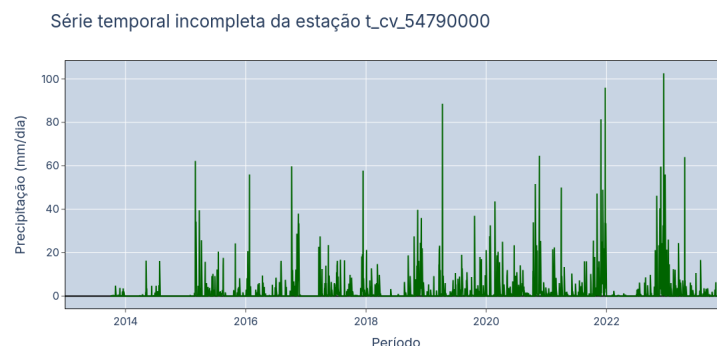


Figura 3.7 – Série temporal incompleta da estação t_cv_54790000
(fonte: o autor)

Note que no início desta série de precipitação, o ano de 2013, não possuem dados. As séries de chuva completas ficaram desta forma (figuras 3.8 e 3.9)

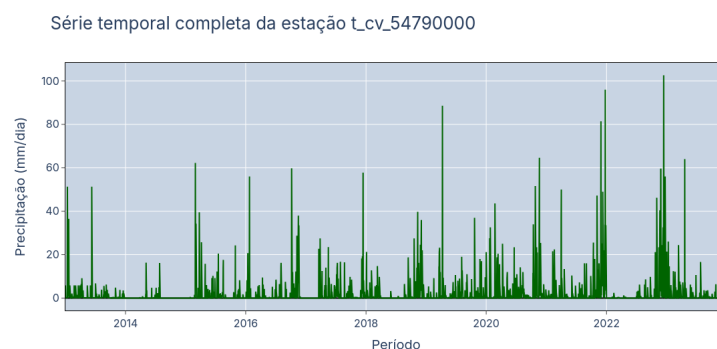


Figura 3.8 – Série temporal completa da estação t_cv_54790000
(fonte: o autor)

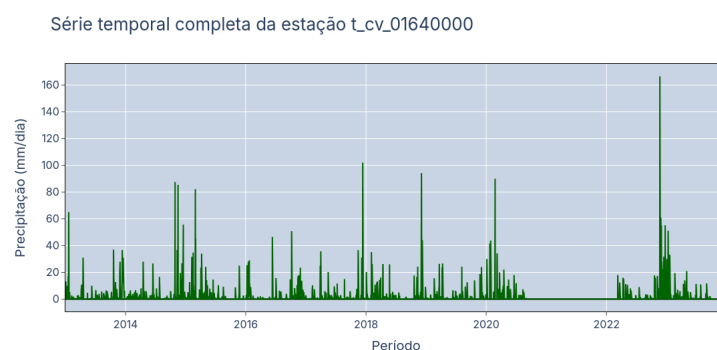


Figura 3.9 – Série temporal completa da estação t_cv_01640000
(fonte: o autor)

3.3.2 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 3.10)

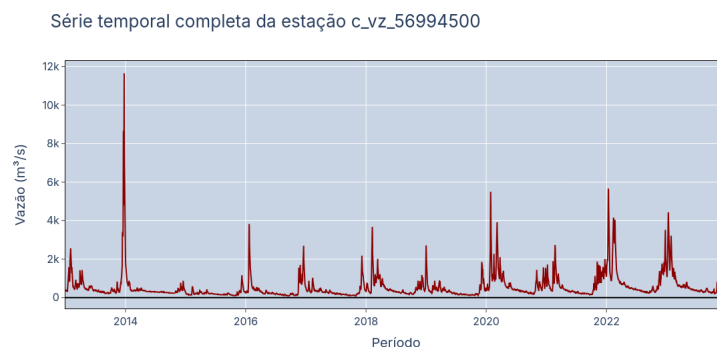


Figura 3.10 – Série temporal completa da estação c_vz_56994500
(fonte: o autor)

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 3.11)

A outra estação removida foi a t_cv_56994500. Conforme pode ser observado na figura 3.12, 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 na tabela e, adiante, o gráfico da série temporal de cada uma delas. (figuras 3.13, 3.14, 3.15 e 3.16)

Tabela 3.5 – Estações de precipitação usadas - final
(fonte: o autor)

Estação	# dados faltantes	% dados faltantes
c_cv_01941010	153	3,81
c_cv_01941004	31	0,77
c_cv_01941006	0	0,00
t_cv_56990005	1395	34,73

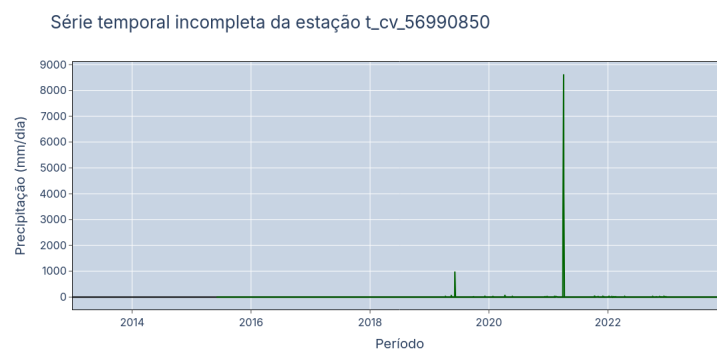


Figura 3.11 – Série temporal da estação t_cv_56990850 - não utilizada
(fonte: o autor)



Figura 3.12 – Série temporal da estação t_cv_56994500 - não utilizada
(fonte: o autor)

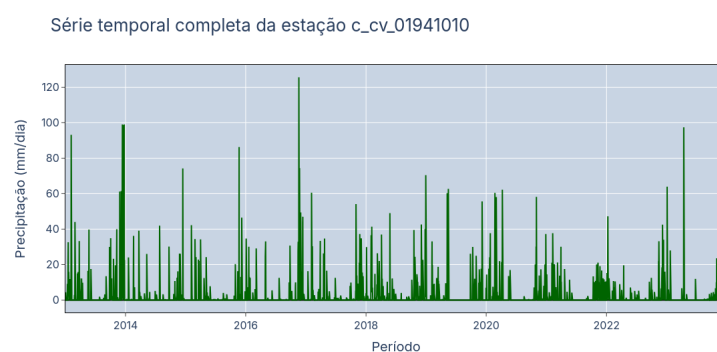


Figura 3.13 – Série temporal completa da estação c_cv_01941010
(fonte: o autor)

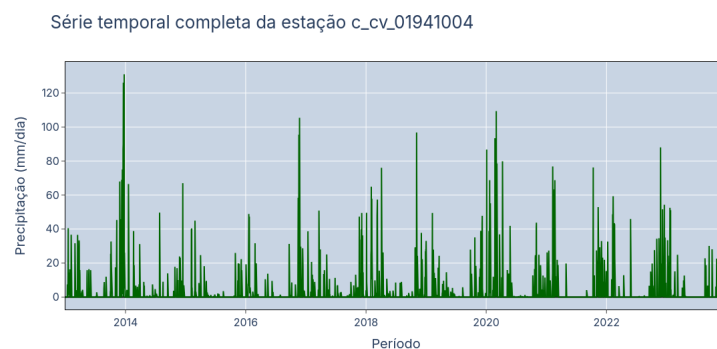


Figura 3.14 – Série temporal completa da estação c_cv_01941004
(fonte: o autor)

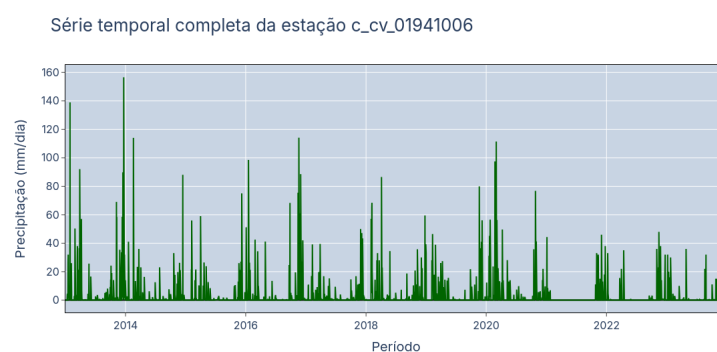


Figura 3.15 – Série temporal completa da estação c_cv_01941006
(fonte: o autor)

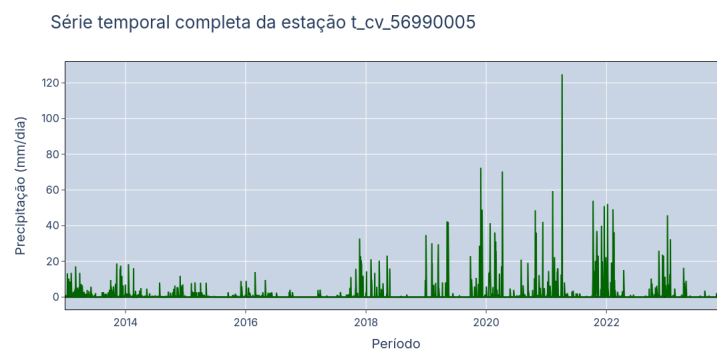


Figura 3.16 – Série temporal completa da estação t_cv_56990005
(fonte: o autor)

3.3.3 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, conforme mencionado anteriormente. 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 3.17) Os dados disponíveis mais antigos datavam de 2020. Contudo, em conformidade com o escopo estabelecido para este estudo, foi realizado o preenchimento dos dados faltantes, aplicando-se o mesmo protocolo utilizado para os demais rios analisados. Este procedimento foi necessário para garantir a consistência, integridade e comparabilidade das análises subsequentes.



Figura 3.17 – Série temporal incompleta da estação t_vz_62020080
(fonte: o autor)

Infelizmente, o caráter ruidoso da série permaneceu mesmo após a aplicação do protocolo de preenchimento dos dados ausentes, conforme pode ser observado na imagem final gerada. (figura 3.18) A série em questão apresentava 2099 dias faltantes, correspondendo a aproximadamente 64% de dados nulos. Outro aspecto relevante para essa estação é que, diferentemente das outras, não foram utilizados os 4.017 registros previstos inicialmente. As informações mais antigas disponíveis datavam de 2015, resultando, assim, em um total de 3289 registros diários utilizados especificamente para o rio Grande.

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 exclusivamente em valores zero. Por essa razão, a estação t_cv_62020080 foi completamente excluída do estudo.



Figura 3.18 – Série temporal completa da estação t_vz_62020080
(fonte: o autor)

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. (figura 3.19) Trata-se de uma série com uma quantidade expressiva de dados, que efetivamente pôde contribuir de maneira significativa para as análises realizadas.

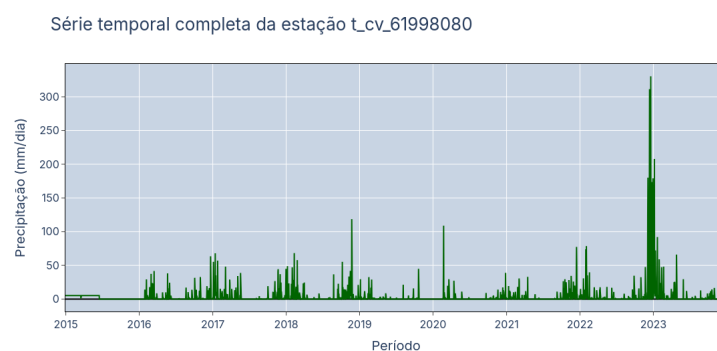


Figura 3.19 – Série completa da estação t_cv_61998080
(fonte: o autor)

Ressalta-se que o trecho de dados faltantes para a estação t_cv_61998080 concentrava-se no início da série temporal, especificamente no ano de 2015. No gráfico os dados já estão imputados.

3.3.4 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 (3.20).

Para esta estação, o preenchimento sazonal foi suficiente para suprir as lacunas existentes, não sendo necessário aplicar procedimentos adicionais de imputação de dados. (figura 3.21)



Figura 3.20 – Detalhe do trecho com dados nulos da estação c_vz_44290002 (fonte: o autor)



Figura 3.21 – 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 3.22, 3.23, 3.24)

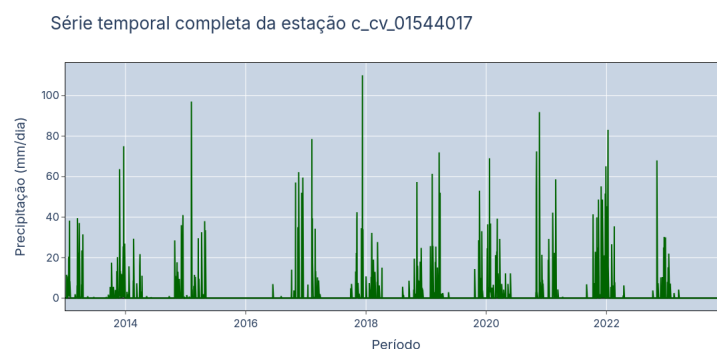


Figura 3.22 – Série temporal completa da estação c_cv_01544017
(fonte: o autor)

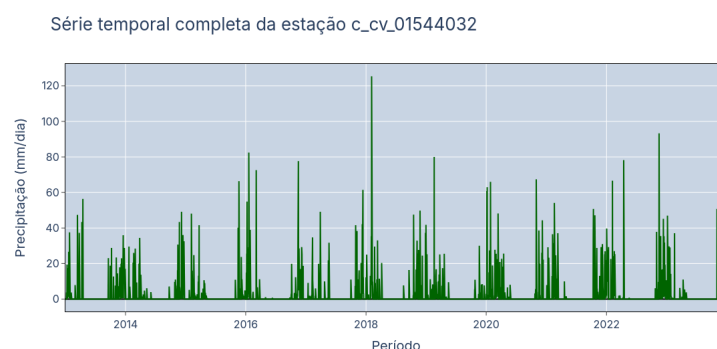


Figura 3.23 – Série temporal completa da estação c_cv_01544032
(fonte: o autor)

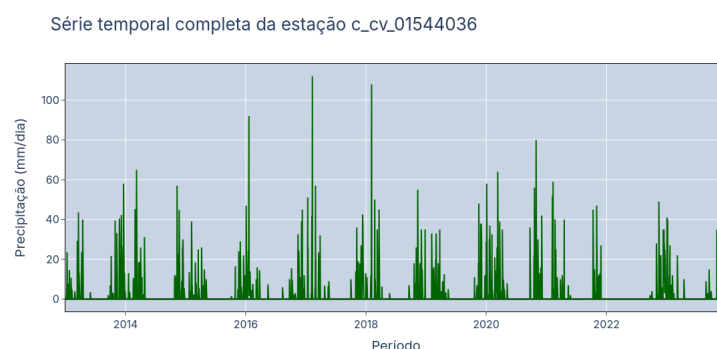


Figura 3.24 – Série temporal completa da estação c_cv_01544036
(fonte: o autor)

3.4 Variáveis Utilizadas

As variáveis do trabalho, exceto as categóricas, obviamente, são todas contínuas. Todas as séries temporais foram ajustadas para estarem completas dentro do período trabalhado, totalizando 4017 registros diários. A exceção ficou por conta dos dados do rio Grande, em que o dado mais antigo foi o dia 30 de dezembro de 2014.

De início, os valores passados (*lags*) de vazão da estação alvo foram utilizados como variáveis preditoras. Utilizou-se 7 dias passados, ou seja, o valor de uma semana passada para capturar o comportamento mais imediatamente próximo da vazão medida na estação.

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
(fonte: o autor)

Variável	#	Média	Desvio-padrão	Mín	< 50%	Máx
t_cv_01640000	4017	1,58	6,85	0,00	0,00	166,60
t_cv_54790000	4017	1,68	6,31	0,00	0,00	102,60
t_vz_54790000 (y)	4017	160,95	267,90	0,00	95,95	3716,65

Tabela 3.7 – Variáveis utilizadas - rio Doce
(fonte: o autor)

Variável	#	Média	Desvio-padrão	Mín	< 50%	Máx
c_cv_01941010	4017	2,21	8,41	0,00	0,00	125,60
c_cv_01941004	4017	2,59	9,83	0,00	0,00	131,00
c_cv_01941006	4017	2,44	9,76	0,00	0,00	156,60
t_cv_56990005	4017	1,15	5,19	0,00	0,00	124,80
c_vz_56994500 (y)	4017	542,05	656,99	75,15	341,26	11655,20

Tabela 3.8 – Variáveis utilizadas - rio Grande
(fonte: o autor)

Variável	#	Média	Desvio-padrão	Mín	< 50%	Máx
t_cv_61998080	3289	3,13	13,92	0,00	0,00	330,40
t_vz_62020080 (y)	3289	3405,27	873,88	1603,58	3170,08	11939,49

Tabela 3.9 – Variáveis utilizadas - rio São Francisco
(fonte: o autor)

Variável	#	Média	Desvio-padrão	Mín	< 50%	Máx
c_cv_01544017	4017	1,60	7,39	0,00	0,00	110,00
c_cv_01544032	4017	2,30	8,13	0,00	0,00	125,40
c_cv_01544036	4017	1,99	7,59	0,00	0,00	112,00
c_vz_44290002 (y)	4017	1115,88	998,40	254,75	812,26	7338,65

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 $0,00 \text{ m}^3/\text{s}$, 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, 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 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. Contudo, com este rio especificamente, os dados de vazão tiveram alguns problemas e dificuldades e é provável que estes números não estejam coerentes com a realidade. Mas o rio Doce é realmente considerável. (tabela 3.7) Vale destacar, no entanto, que esta amplitude não especifica se foi dentro de 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 “< 50%”, que significa metade de toda massa de dados de precipitação estavam abaixo deste valor, ou seja, metade de todos os dados de precipitação estavam em $0,00 \text{ mm}/\text{dia}$. 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 à elas no treinamento dos modelos. As novas variáveis foram:

Tabela 3.10 – Variáveis acumuladas para as precipitações
(fonte: o autor)

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

Quanto aos dados categóricos utilizados neste estudo, inicialmente foi 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.(30) 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, aplica-se transformações apenas nas variáveis contínuas.

Pois bem, havia uma dificuldade aqui, uma vez que essas variáveis categóricas seriam incorporadas com o objetivo de capturar o comportamento sazonal da série temporal. 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 daquelas variáveis, mês do ano 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 variáveis foram codificadas no esquema “*label encoding*” e desta forma as faixas de valores foram 1 a 12 para ‘*month*’ e de 1 a 4 para ‘*estacao*’.

3.5 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.

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) 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.(37) A série temporal do rio Grande apresentou o pior desempenho em termos de autocorrelação.(figura 3.29) A decomposição da série também revelou um comportamento mais ruidoso, o que pode ser atribuído ao fato de esta série conter mais lacunas e apresentar maiores desafios no preenchimento dos dados ausentes.(figura 3.30) Nos gráficos de autocorrelação, o *lag* de 365 dias – correspondente a

um ano – foi destacado com uma linha preta vertical.

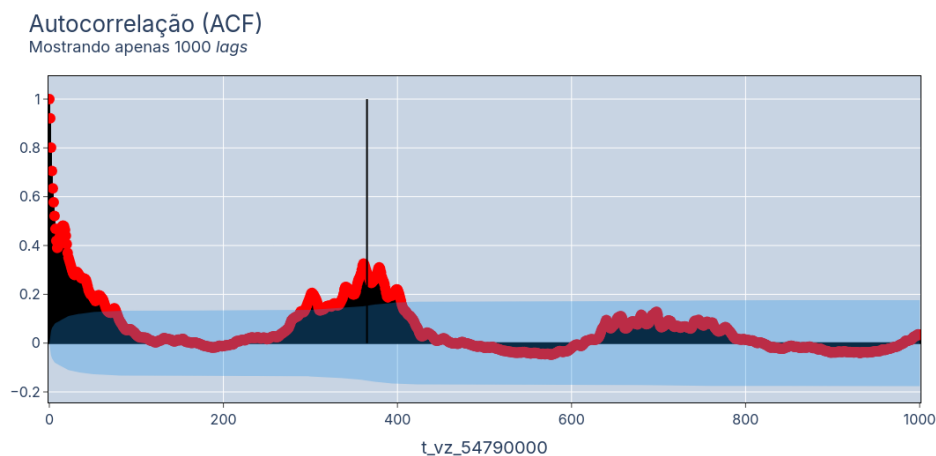


Figura 3.25 – Autocorrelação para a vazão do rio Jequitinhonha
(fonte: o autor)

Decomposição da série temporal: t_vz_54790000

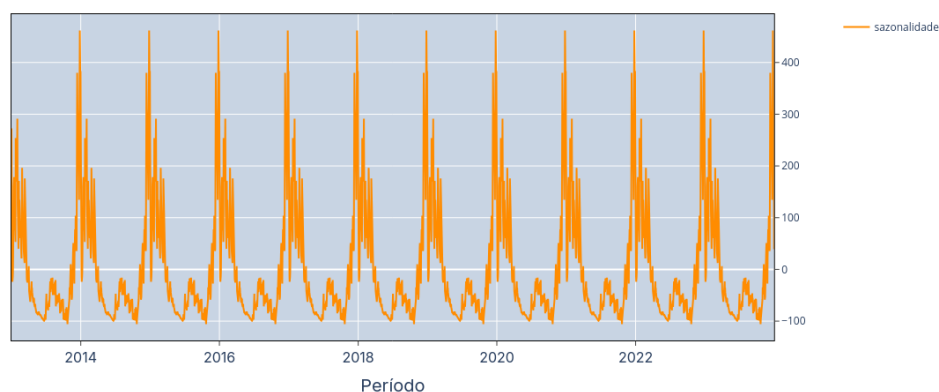


Figura 3.26 – Componente sazonal da série de vazão do rio Jequitinhonha
(fonte: o autor)

As séries temporais consideradas, digamos, mais bem comportadas foram as dos rios Doce e São Francisco.(figuras (3.27), (3.31)) 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.(29)

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 distribuição. Em uma distribuição normal, a maioria dos eventos se concentra em torno da média,

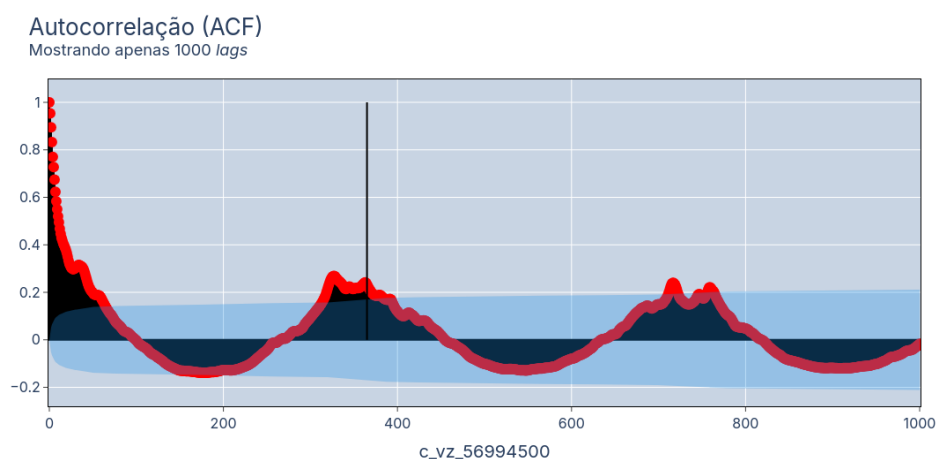


Figura 3.27 – Autocorrelação para a vazão do rio Doce
(fonte: o autor)

Decomposição da série temporal: c_vz_56994500

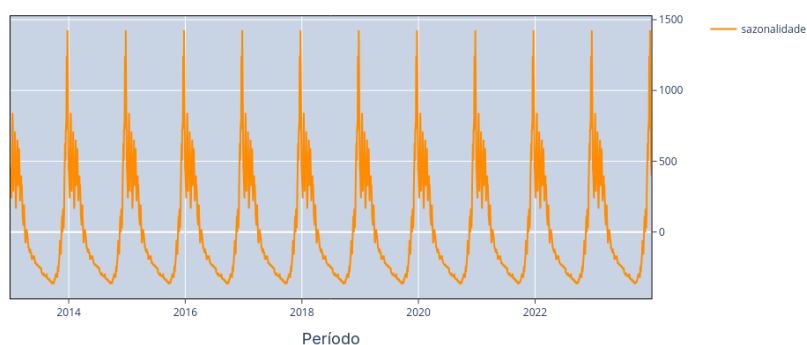


Figura 3.28 – Componente sazonal da série de vazão do rio Doce
(fonte: o autor)

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

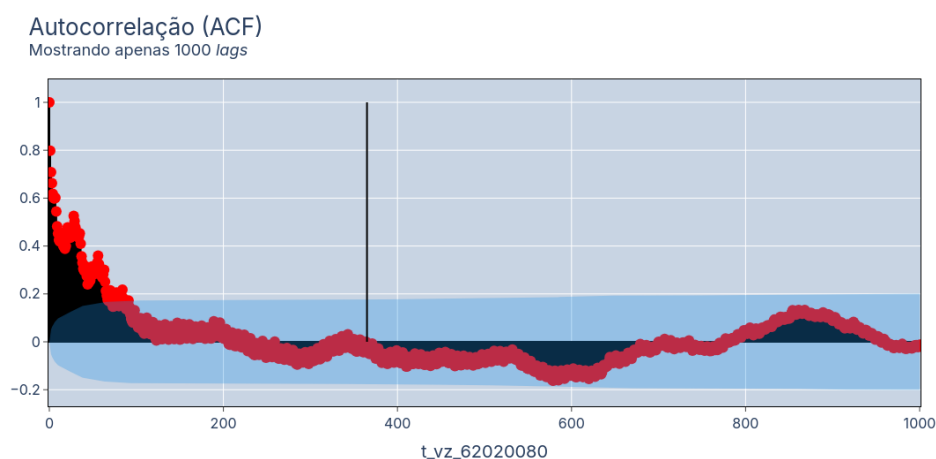


Figura 3.29 – Autocorrelação para a vazão do rio Grande
(fonte: o autor)

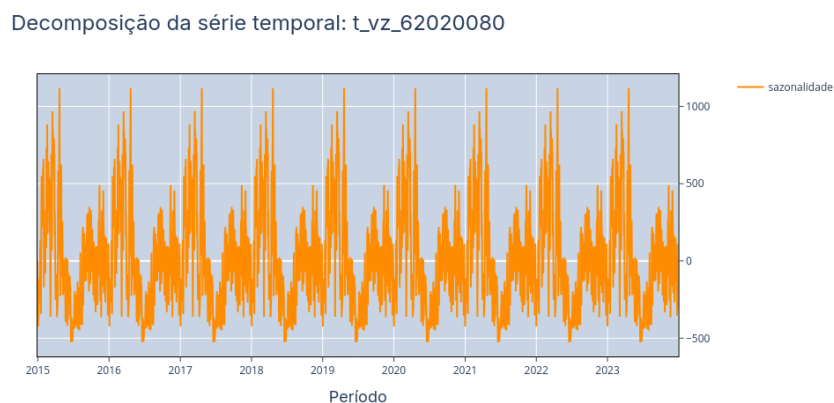


Figura 3.30 – Componente sazonal da série de vazão do rio Grande
(fonte: o autor)

Posto que os dados contêm valores iguais a zero, a transformação pelo logaritmo natural ($\ln(\cdot)$) não foi aplicada, pois o cálculo de logaritmo não é definido para valores zero. Ao invés disso, foi utilizada a transformação ‘ $\log1p(\cdot)$ ’ da biblioteca NumPy (20), que adiciona 1 ao valor antes da transformação, evitando erros relacionados ao logaritmo de zero.

Existem outras formas de ajustar a distribuição dos dados, tais como Box-Cox, Yeo-Johnson e transformação por quantis [REF REF], mas era necessário que houvesse fácil reversibilidade nos valores finais preditos e que garantisse interpretabilidade, por isso optou-se por uma transformação simples mas que atendeu satisfatoriamente ao propósito.

Após a transformação a distribuição dos eventos ficou menos assimétrica, como pode ser visto nas figuras 3.33, 3.34, 3.35, 3.36, 3.37, 3.38, 3.39, 3.40. Para o rio Grande, visualmente, parece não ter havido tanta diferença, mas quando se analisa os valores

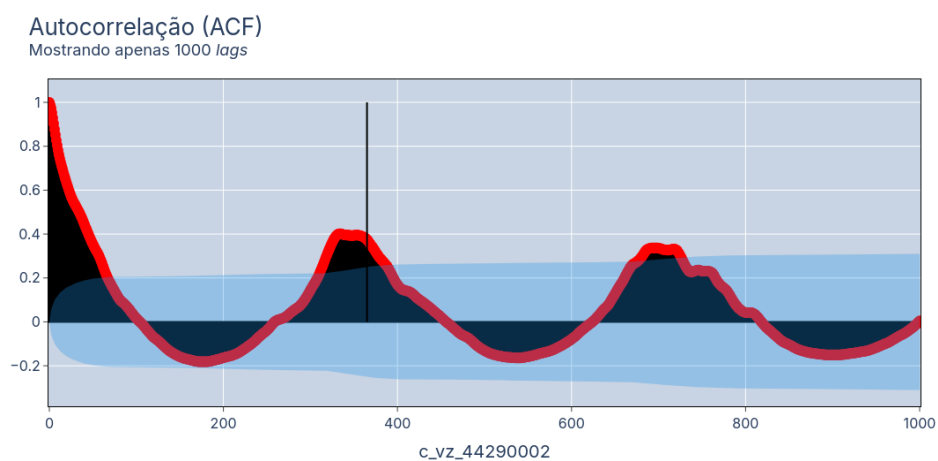


Figura 3.31 – Autocorrelação para a vazão do rio São Francisco
(fonte: o autor)

Decomposição da série temporal: c_vz_44290002

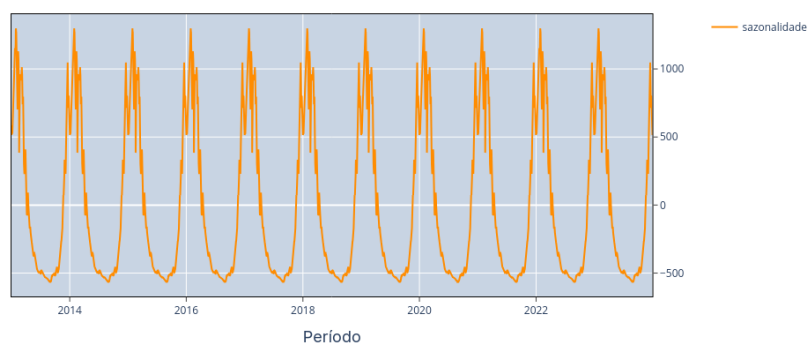


Figura 3.32 – Componente sazonal da série de vazão do rio São Francisco
(fonte: o autor)

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

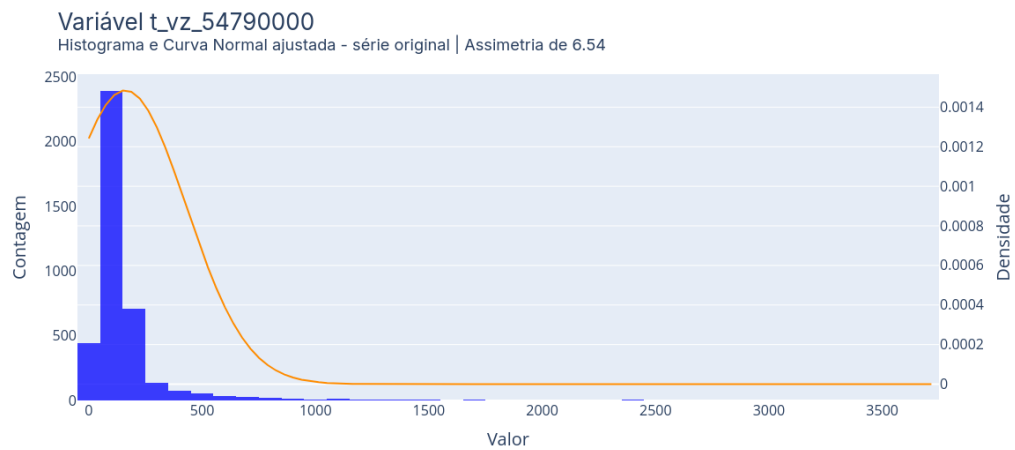


Figura 3.33 – Dados originais para o rio Jequitinhonha
(fonte: o autor)

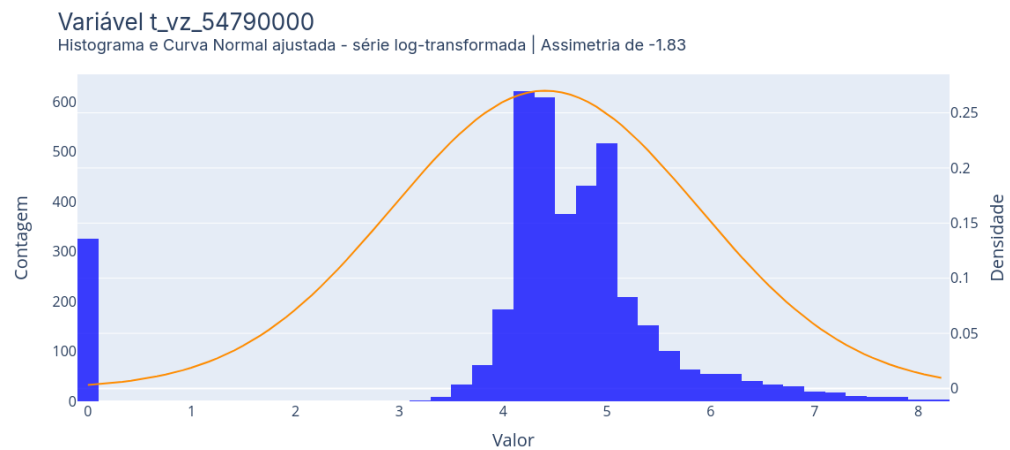


Figura 3.34 – Dados log-transformados para o rio Jequitinhonha
(fonte: o autor)

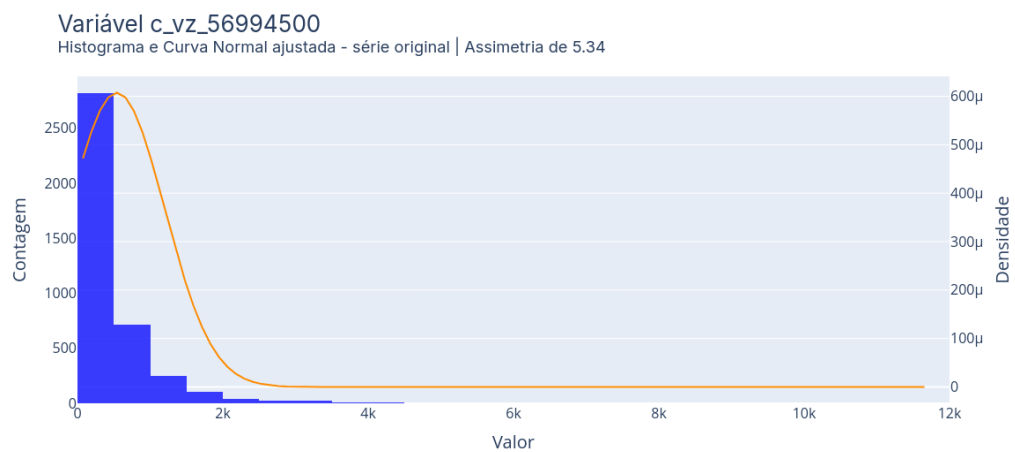


Figura 3.35 – Dados originais para o rio Doce
(fonte: o autor)

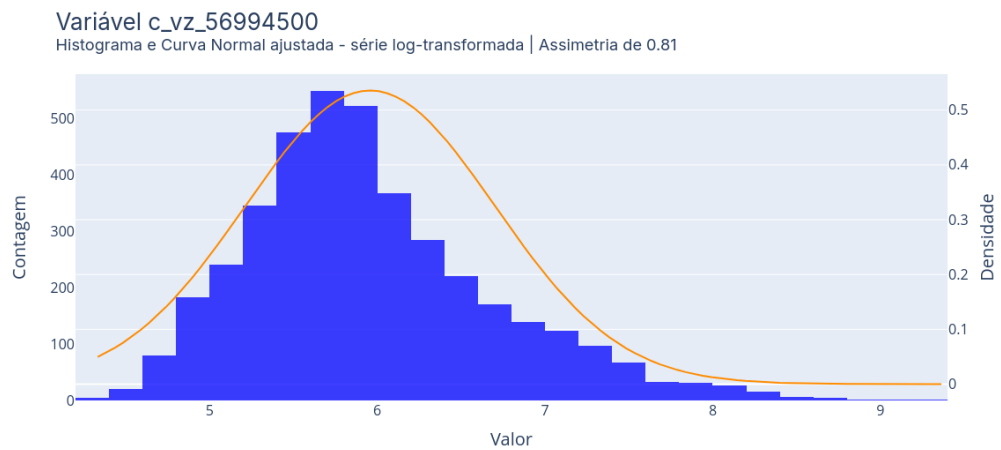


Figura 3.36 – Dados log-transformados para o rio Doce
(fonte: o autor)

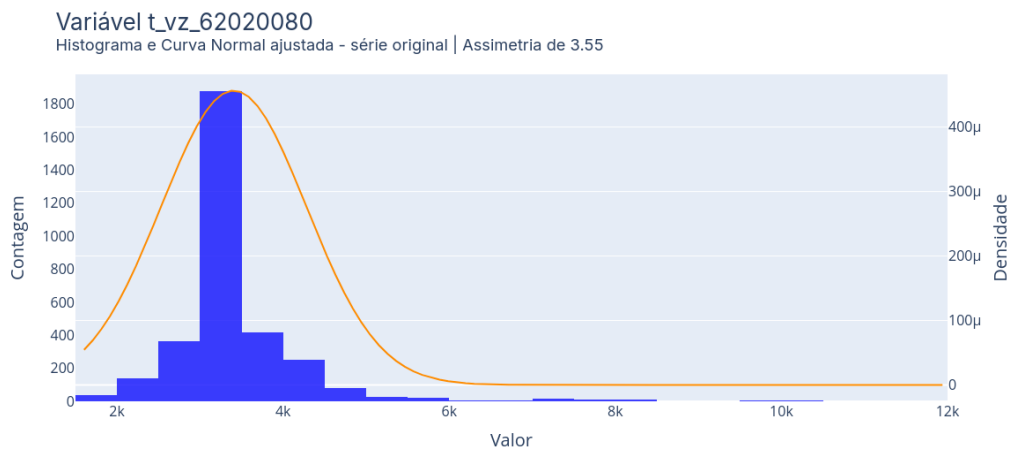


Figura 3.37 – Dados originais para o rio Grande
(fonte: o autor)

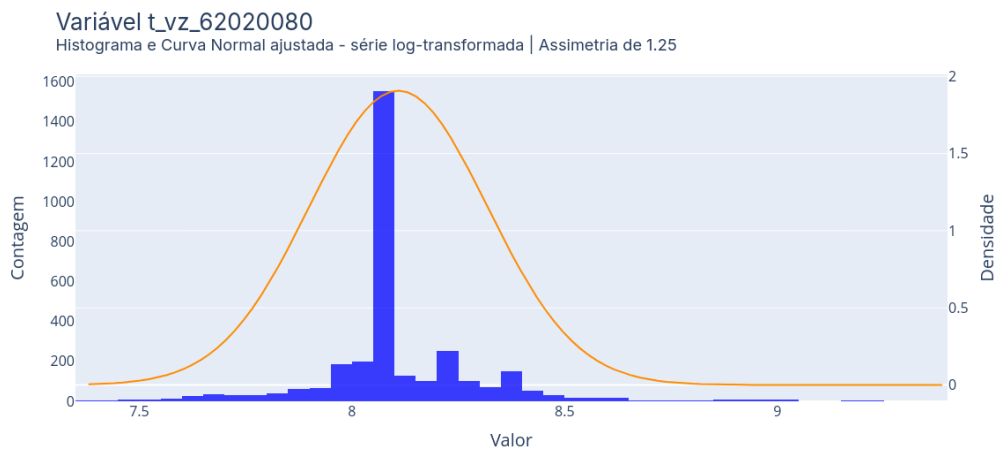


Figura 3.38 – Dados log-transformados para o rio Grande
(fonte: o autor)

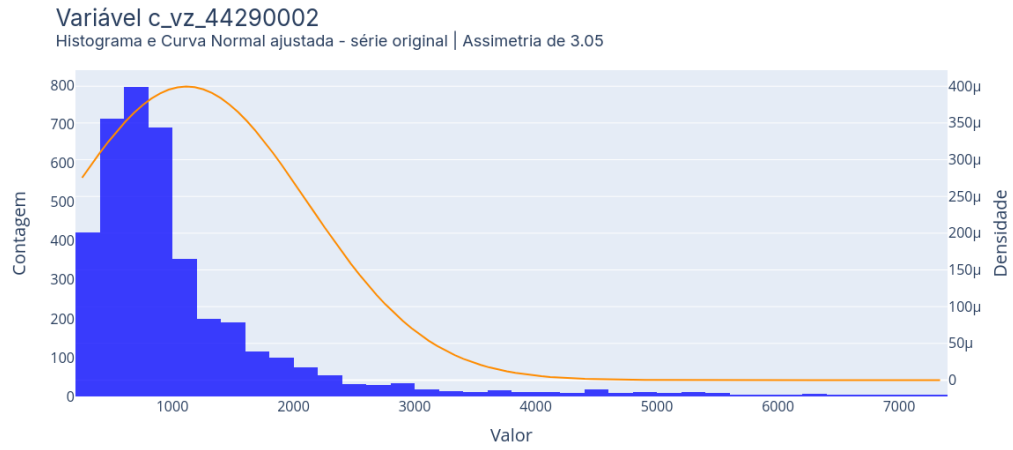


Figura 3.39 – Dados originais para o rio São Francisco
(fonte: o autor)

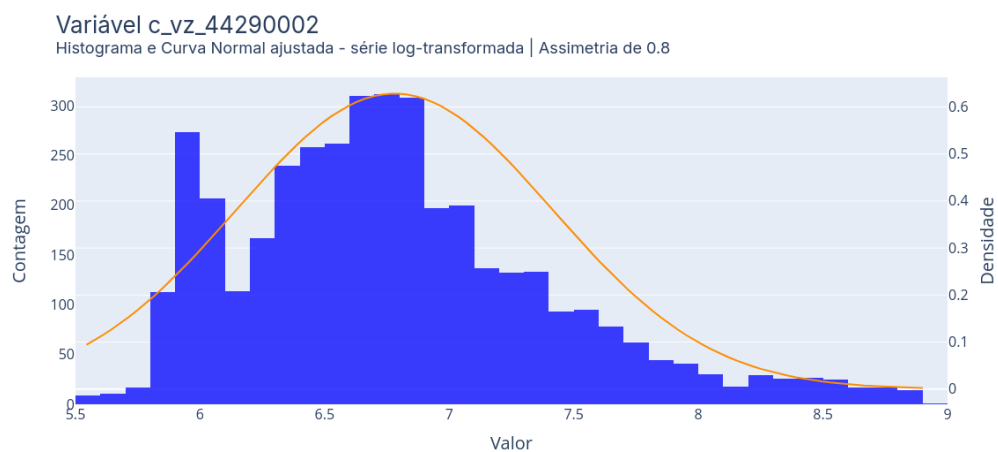


Figura 3.40 – Dados log-transformados para o rio São Francisco
(fonte: o autor)

Para finalizar, uma análise comumente realizada em séries temporais é a verificação de sua estacionariedade. A estacionariedade pressupõe que características estatísticas da série, como média e variância, permaneçam constantes ao longo do tempo. No entanto, neste estudo, essa avaliação não foi realizada, pois séries temporais de dados ambientais, como precipitação e vazão, geralmente apresentam sazonalidade e tendências que violam o conceito de estacionariedade, conforme explicado em Hyndman and Athanasopoulos (26). Outros desafios a esta análise também incluem a rápida mudança do uso do solo e remoção de cobertura vegetal original, alterando significativamente a dinâmica hidrológica da região.(6)

Essas características, longe de serem consideradas como ruído ou anomalias, fazem parte da própria natureza dos dados ambientais e são cruciais para a modelagem e previsão. Assim, em oposição a tentar forçar a estacionariedade, este trabalho optou por lidar com a sazonalidade e as tendências diretamente, utilizando técnicas que captam essas dinâmicas, visando garantir previsões mais realistas e representativas dos processos hidrológicos.

3.6 Modelos de Aprendizado de Máquina

3.6.1 Seasonal Naive - SN

Um método que é similar ao *Naive*, contudo, usa a última observação conhecida do mesmo período de tempo (por exemplo, mesmo dia do mês anterior, mesma semana do ano anterior) numa tentativa de capturar o comportamento sazonal.

O *Seasonal Naive* não pode ser considerado um modelo de previsão sofisticado. Em vez disso, ele funciona como uma linha de base (*baseline*), servindo como ponto de partida para avaliar o desempenho de outros modelos de previsão.(25)

Por ser um modelo simples, ele não captura tendências ou variações complexas, mas estabelece um *benchmark* mínimo com o qual outros métodos mais elaborados e complexos devem ser comparados.

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

3.6.2 Regressão Linear - LR

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

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \dots + \beta_k x_{k,t} + \varepsilon_t \quad (3.1)$$

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.(24)

$$\sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n (y_t - \beta_0 - \beta_1 x_{1,t} - \beta_2 x_{2,t} - \cdots - \beta_k x_{k,t})^2. \quad (3.2)$$

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

Os resultados obtidos com este modelo mostraram-se muito bons, 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.

3.6.3 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 pesquisador tem controle sobre eles.(39)(12)(33)

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, pra 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, pra acelerar o treinamento do modelo. O hiperparâmetro *‘verbose’* é pra não escrever cada passo de treinamento na saída de texto do sistema e *‘random_seed’* é pra garantir reprodutibilidade dos resultados em todas as execuções do modelo. Apenas estes hiperparâmetros foram selecionados, tudo mais ficou de acordo com o padrão da biblioteca.

Tabela 3.11 – Hiperparâmetros para o modelo CatBoost
(fonte: o autor)

Hiperparâmetro	Valor
random_seed	1984
cat_features	['estacao', 'month']
verbose	False
allow_writing_files	False
thread_count	os.cpu_count()//2

3.6.4 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*).⁽³⁴⁾⁽³⁶⁾

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. O modelo de Regressão Linear serve como uma *baseline* confiável devido à sua eficácia e simplicidade. Já os modelos CatBoost e Floresta Aleatória foram opções mais avançadas, capazes de lidar com a complexidade dos dados e oferecer previsões precisas e eficientes. A comparação entre todos estes modelos permitiu que se escolhesse a abordagem que melhor atendesse às necessidades específicas da previsão de vazões.

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

Tabela 3.12 – Hiperparâmetros para o modelo RandomForest
(fonte: o autor)

Hiperparâmetro	Valor
random_state	1984
verbose	0
n_jobs	os.cpu_count()/2

3.7 Métricas de Avaliação

Para avaliar o desempenho dos modelos de previsão utilizados neste estudo, foram adotadas quatro métricas: MAPE (*Mean Absolute Percentage Error*), RMSE (*Root Mean Square Error*), PBIAS (*Percent Bias*) e KGE (*Kling-Gupta Efficiency*). 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.

- **MAPE (*Mean Absolute Percentage Error*):** 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.(23)

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{O_i - P_i}{O_i} \right| \quad (3.3)$$

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

- **PBIAS (*Percent Bias*):** 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.(28)

$$PBIAS = 100 \times \frac{\sum_{i=1}^n (P_i - O_i)}{\sum_{i=1}^n O_i} \quad (3.4)$$

- ◇ O_i valores observados
 - ◇ P_i valores previstos
 - ◇ n o número total de observações
- **KGE (*Kling-Gupta Efficiency*)**: 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.(19)

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (3.5)$$

- ◇ 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.

3.8 Modelo proposto

Tudo detalhado até aqui, agora é preciso descrever o fluxo de trabalho. Por fim, será apresentado o fluxograma do modelo proposto para uma rápida depreensão visual.

Primeiramente é necessário especificar o período histórico de análise. Depois de 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.(38) 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, conseqüentemente, consumiria muita memória, ou o que foi adotado neste trabalho que foi realizar reajustes (*'refit'*) ao modelo inicial num processo recursivo (*'recursive forecasting'*). Esse processo de 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 imagens a seguir (observação: a fonte de onde as imagens foram retiradas nomeia o processo de Walk-Forward Validation como *'backtesting'* (13))

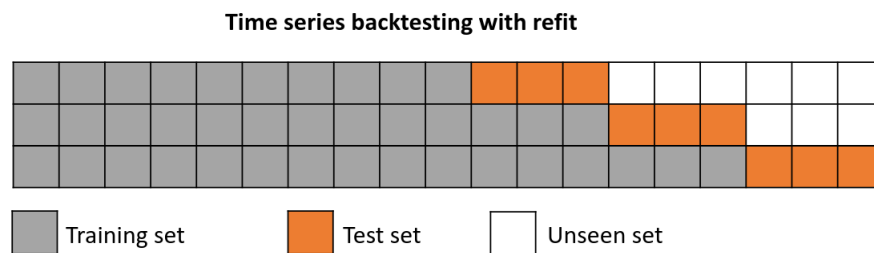


Figura 3.41 – Diagrama mostrando a divisão dos dados de treino/teste com *refit*.
(fonte: (13))

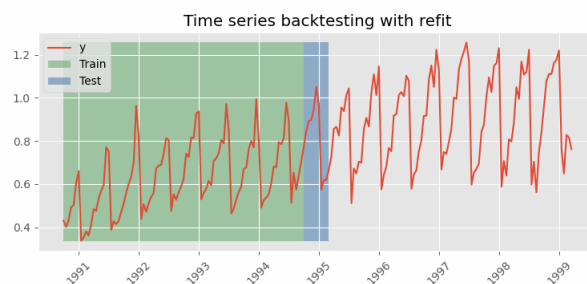


Figura 3.42 – WFV com janela expandida e *refit* - imagem 1.
(fonte: (13))

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.(22) Apesar de estar sendo usado intervalo de previsão de 95% na análise, nem sempre o valor real é capturado. Existe uma incerteza quanto à

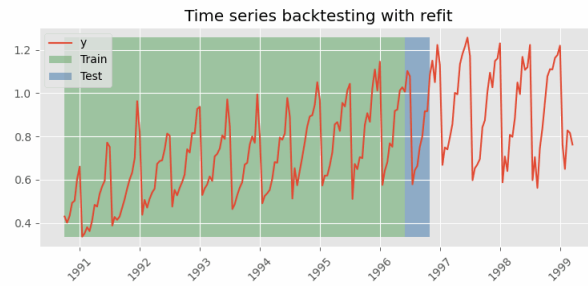


Figura 3.43 – WFV com janela expandida e *refit* - imagem 2.
(fonte: (13))

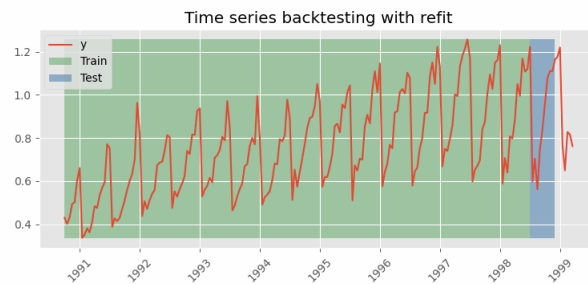


Figura 3.44 – WFV com janela expandida e *refit* - imagem 3.
(fonte: (13))

previsão. Esta 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 3.45)

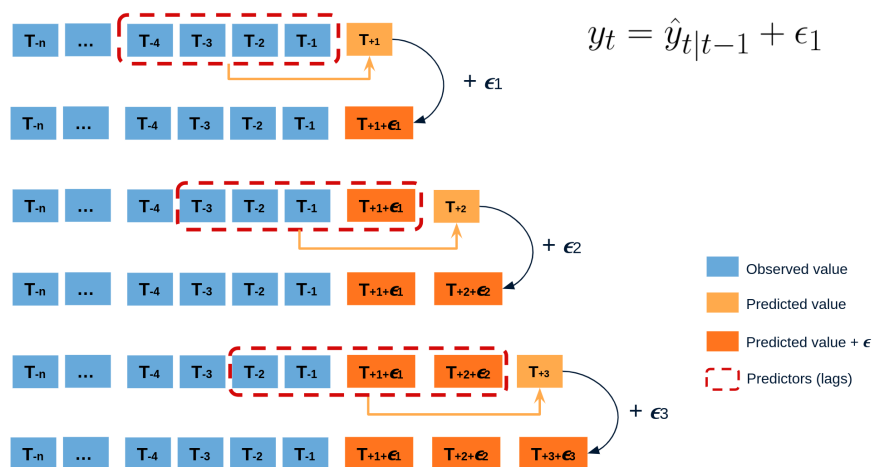


Figura 3.45 – Diagrama de como se calcula o intervalo de previsão.
(fonte: (14))

Um fenômeno que pode acontecer é os intervalos de previsão serem estreitos.(21) Isso porque os modelos não conseguem capturar todos as fontes de incerteza à partir dos dados, a saber:

- 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 confiou-se 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.

```

1 def cobertura_empirica(y, lower_bound, upper_bound):
2     media = np.mean(np.logical_and(y >= lower_bound, y <= upper_bound))
3     return f"{round(100*media, 2)} %"

```

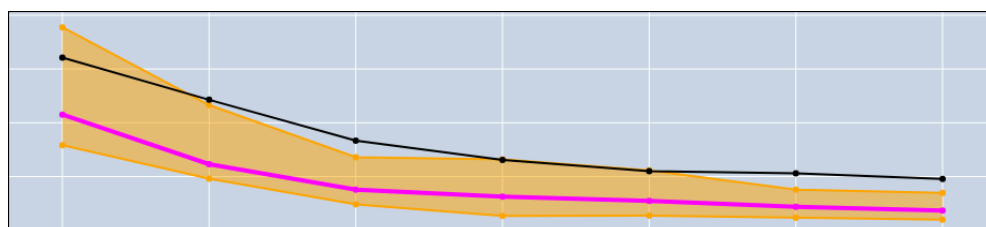


Figura 3.46 – Previsão com valor fora dos limites
(fonte: o autor)

Aplicando a função à figura 3.46 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

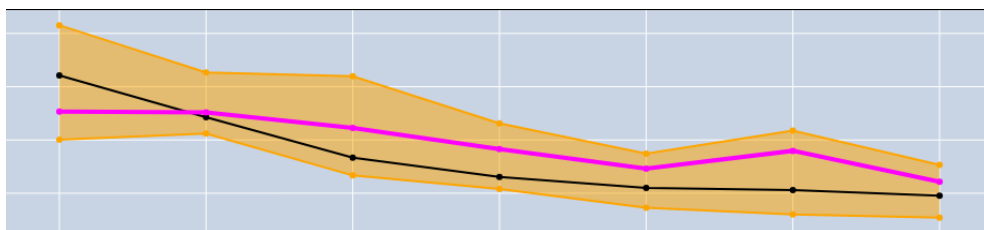


Figura 3.47 – Previsão completamente nos limites
(fonte: o autor)

pretendido de 95% só conseguiu, empiricamente, capturar cerca de 43% dos valores reais do horizonte de previsão em questão. Na figura 3.47, por sua vez, o intervalo capturou 100% dos valores observados. Ambas figuras mostram previsão de 7 dias, e sem considerar a linha rosa, que foi a previsão que os modelos geraram, o modelo da figura 3.47 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 $0 \text{ m}^3/\text{s}$ a 200 mil m^3/s , então qualquer valor observado seria capturado, de fato, mas isso não ajudaria em nada na tomada de decisão.

Em cada resultado será discutido o atraso (*delay*) da série prevista em relação à série original, quando houver, a partir da aplicação do algoritmo FastDTW.(35) *Dynamic Time Warping* (DTW) é uma forma de avaliar a similaridade entre duas sequências quaisquer, neste caso, duas séries temporais: a série temporal dos dados observados e a série temporal dos dados previstos. O algoritmo calcula as distâncias entre os pontos de duas séries esticando e comprimindo-as no eixo do tempo para descobrir o melhor caminho de alinhamento (*warping path*) entre ambas.(2) No caso aqui utilizado, o algoritmo FastDTW é uma implementação do conceito original de DTW de ordem de complexidade $O(n)$ que resulta numa aproximação, visto que o algoritmo de DTW é de ordem $O(n^2)$. Ele escala com a quantidade de dados de entrada, podendo facilmente demandar *terabytes* de memória.(35) Esta análise ajudará a compreender a dinâmica das vazões, fornecendo interpretação sobre quando os eventos passados irão ser percebidos pelos modelos no futuro. E como as análises são na base diária, o resultado do *delay* é a expressão em dias do atraso (*lag*) ou adiantamento (*lead*) de quando um evento (vazão) será percebido pelo modelo e expresso na série gerada por este. Apenas os melhores resultados terão o *delay* discutido.

Encerrada a discussão do *delay*, procede-se com a análise dos resíduos. Os modelos não-lineares CatBoost e RandomForest não necessitam que haja normalidade, nem para os dados de entrada, nem para os resíduos. Diferentemente do modelo de *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.(1)

A última análise é avaliação da importância das variáveis. Como na análise de *delay*, 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.

SHAP values (*SHapley Additive exPlanations*) constituem uma metodologia 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 um impacta o resultado, seja de forma positiva ou negativa, promovendo assim a interpretabilidade e transparência.(31)

Não consta no fluxograma do modelo proposto, mas para todos os modelos teve também uma rodada de execução com os dados originais sem transformação logarítmica, apenas havendo aplicação de escalonamento de dados usando MinMax para o LinearRegression. 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 utilizar os modelos pensando em termos de gestão dos recursos hídricos.

O ambiente computacional em que os experimentos foram executados é:

Tabela 3.13 – Configurações do computador usado nos experimentos
(fonte: o autor)

Configuração	Descrição
CPU	Ryzen 7 5700 (8 núcleos/16 <i>threads</i>)
GPU	RTX 3060
Memória RAM	32 GB 3200MHz
Disco	SSD NVMe
Kernel	Linux 6.8.0-40-generic
Python	versão 3.9.18

Os resultados serão apresentados no capítulo 4.

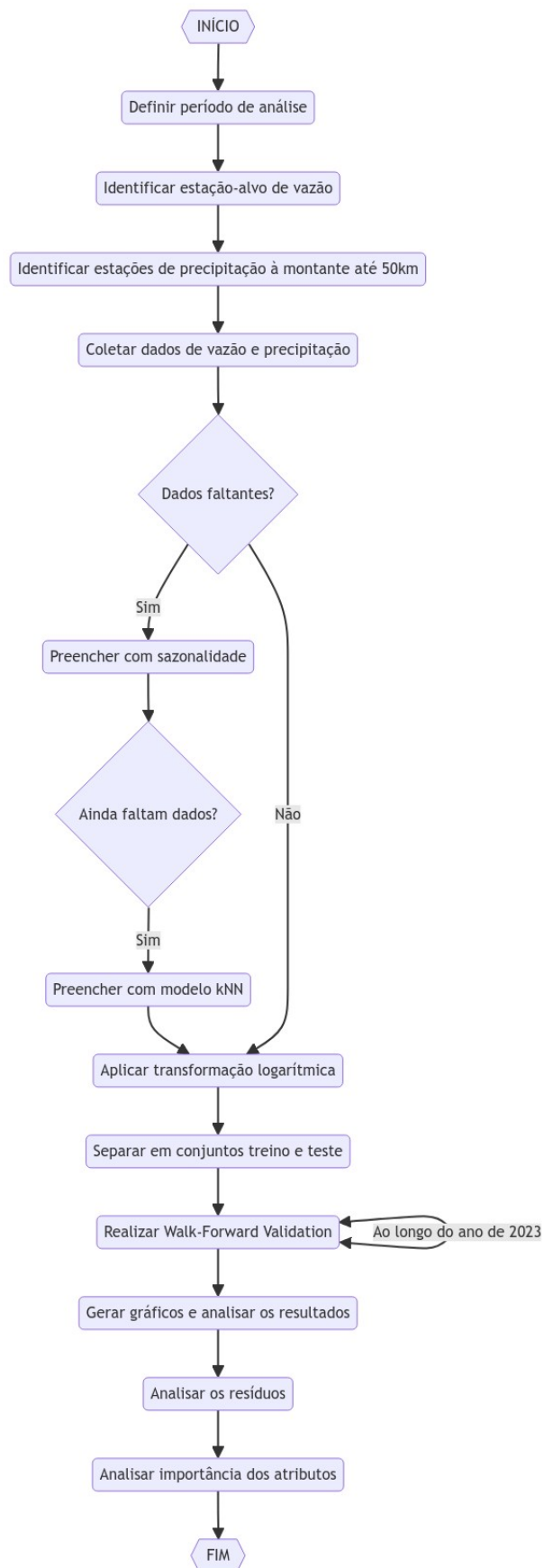


Figura 3.48 – Fluxo de trabalho
(fonte: o autor)

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 pela menor bacia hidrográfica estudada, os resultados revelaram-se bastante satisfatórios, especialmente no que se refere ao modelo de Regressão Linear, 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) SeasonalNaive apresentou resultados abaixo das expectativas em todas as situações avaliadas. (figura 4.1)

Com a MAPE extremamente elevada, de 150%, os resultados indicaram um viés significativo de superestimação, conforme evidenciado pela métrica PBIAS. Verificando a KGE, ficou negativa. 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 zero, o que não faz sentido para o rio em questão, pois implicaria na ausência total de vazão, algo inviável para as condições medidas pela estação. Além disso, o atraso (*delay*) da série prevista foi considerável, atingindo 56,88 dias, com um desvio-padrão de 61,3 dias, sugerindo que um evento pode demorar mais de 60 dias para ser refletido na previsão do modelo.

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 de Regressão Linear mostraram-se bastante promissores. (figura 4.2) Nesta primeira avaliação, os dados foram log-transformados. Isso é verificável no título, com o uso de “destransformados”. Os dados foram log-transformados e retornados para a escala original para desenhar o gráfico e ficarem

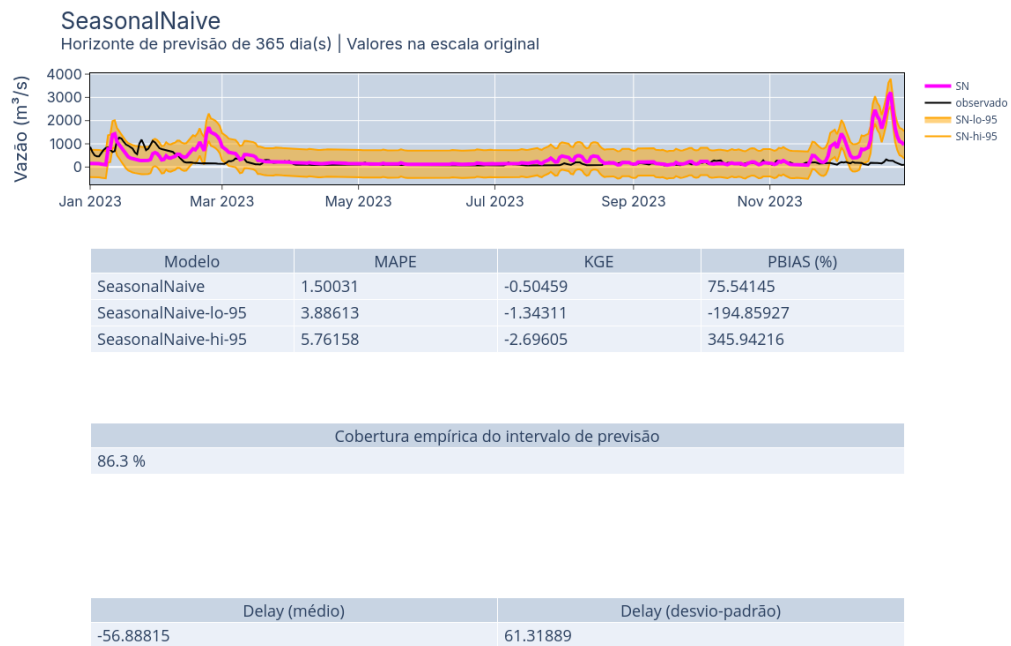


Figura 4.1 – Resultado do SeasonalNaive no teste *Walk-Forward Validation* (fonte: o autor)

acessíveis para quem lê. Essa dinâmica no título se manterá por todo trabalho.

Partindo pela KGE calculada, o resultado mostrou-se excelente. Recordando: quanto mais próxima de 1, melhor. Este valor sugere que o modelo é eficaz na previsão do comportamento hidrológico do sistema em análise, oferecendo previsões que estão bem alinhadas com os dados observados. A MAPE de 14% sugere que o modelo tem uma precisão razoável e é bastante confiável. O modelo apresentou um viés sistemático de subestimar os resultados, conforme aponta a PBIAS de -1,76%. Considerando todas as métricas, o resultado indica que o modelo teve um desempenho global muito bom. A KGE alta 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 97,81% 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 (*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).

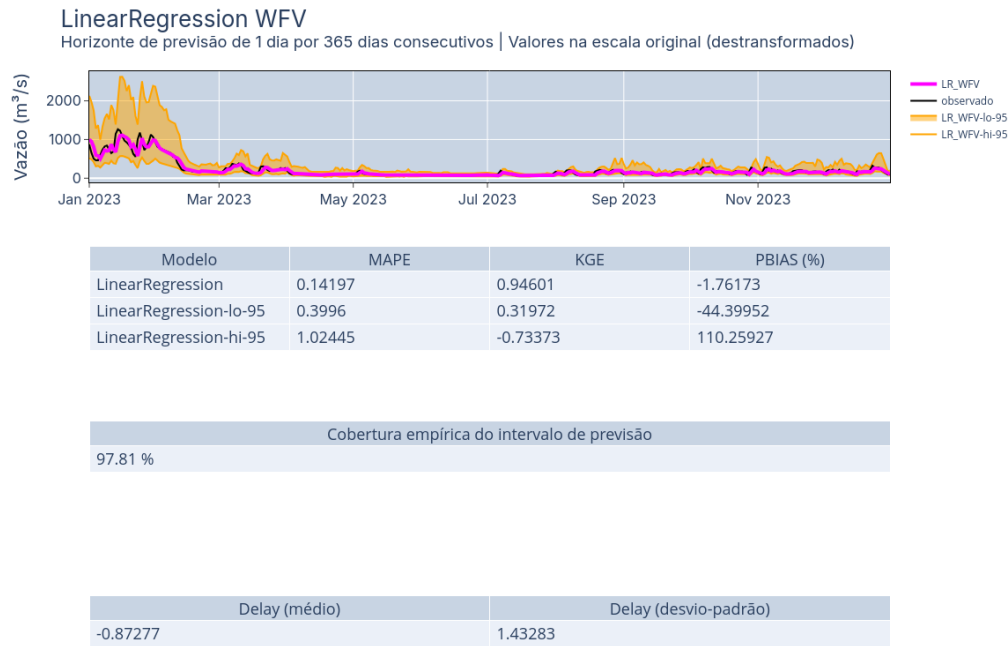


Figura 4.2 – *Walk-Forward Validation* para o modelo Regressão Linear (fonte: o autor)

A análise de *delay* mostrou um resultado médio de $-0,87$. Se pegar a série prevista pelo modelo e comprimir linearmente em $-0,87$, ambas as sequências serão idênticas, ou seja, a série prevista será a série observada. Considerando que os dados estão numa frequência diária, isso pode ser interpretado como o modelo atrasando a previsão em $0,87$ dias, o que seria menos de 24 horas entre o evento ocorrer e ele ser percebido na previsão. Afirmar que o modelo está atrasado 20,88 horas, talvez, não convirja com a realidade, por isso que se considerar o desvio-padrão no resultado, confere mais coerência, pois indicaria que o atraso percebido no fenômeno é de cerca de um dia e meio, dois dias.

Na figura 4.4 podemos observar a concentração dos resíduos em torno de zero. A linha vermelha tracejada é onde o valor previsto e observado são iguais, designando a previsão perfeita, com resíduo 0. Este é o comportamento que se espera, idealmente, os resíduos estarem distribuídos aleatoriamente em torno de zero, sem padrões evidentes

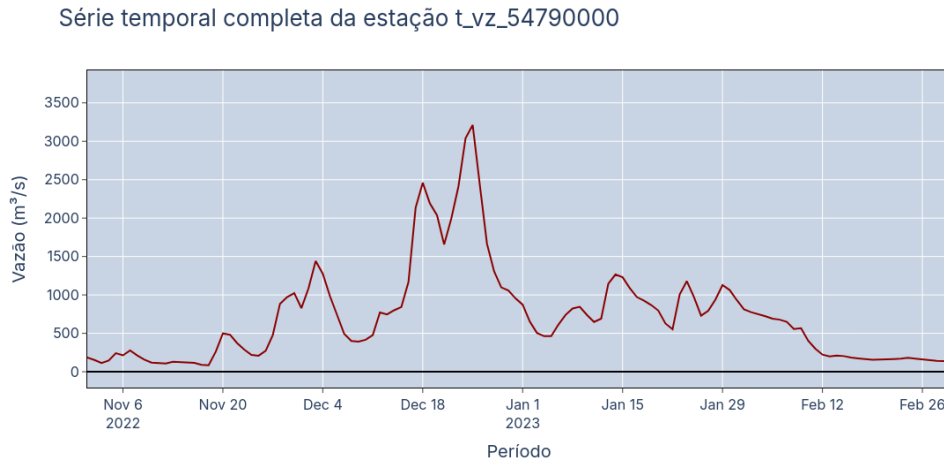


Figura 4.3 – Detalhe do trecho final dos dados, em 2022, usados para treinamento.
(fonte: o autor)

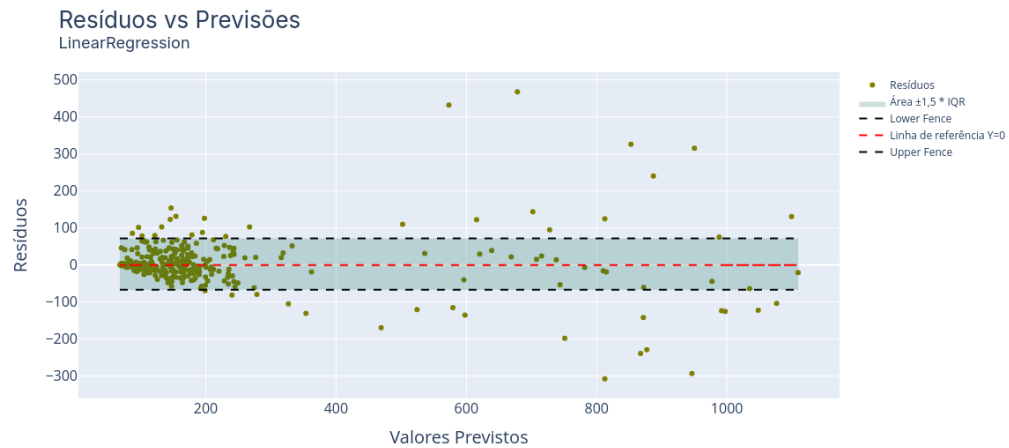


Figura 4.4 – Resíduos da previsão versus os valores das previsões.
(fonte: o autor)

(nenhuma tendência clara de curvatura ou cone). Contudo, à medida que se caminha sobre a linha vermelha tracejada, aumentando o valor no eixo x, há um aumento na dispersão dos resíduos. A área sombreada demarcada são os valores calculados para *lower fence* e *upper fence* a partir do primeiro (−14, 68) e terceiro (19, 93) quartil, que aqui resultaram em −66, 60 e 71, 85, respectivamente. Entre estes valores é onde se espera que os resíduos estejam distribuídos (houve uma prevalência de 87, 95% dos resíduos nesta área), o que fica de fora dessas faixas pode ser interpretado como um *outlier*. 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 significar números elevados.

Veja na figura 4.5 como os resíduos estão dispersos ao longo do tempo. No início do ano e final do ano, exatamente quando os eventos de chuva mais ocorrem, o modelo tende a se dispersar. 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-se 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 4.6 apresenta que os resíduos estão próximos da normalidade quanto à distribuição destes em torno de 0, com uma assimetria de 1,46, o que pode ser considerada baixa, mas não descartável. O valor positivo para assimetria sugere que houve casos em que o modelo subestimou os valores (existe uma cauda à direita do centro dos dados), o que combina com os gráficos anteriores de dispersão. Mas cabe destacar que os resíduos não estão significativamente desviados em nenhuma direção, apresentam distribuição aleatória e não sistemática e este é um comportamento desejado. A análise da função de autocorrelação (ACF) está na figura 4.7 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.

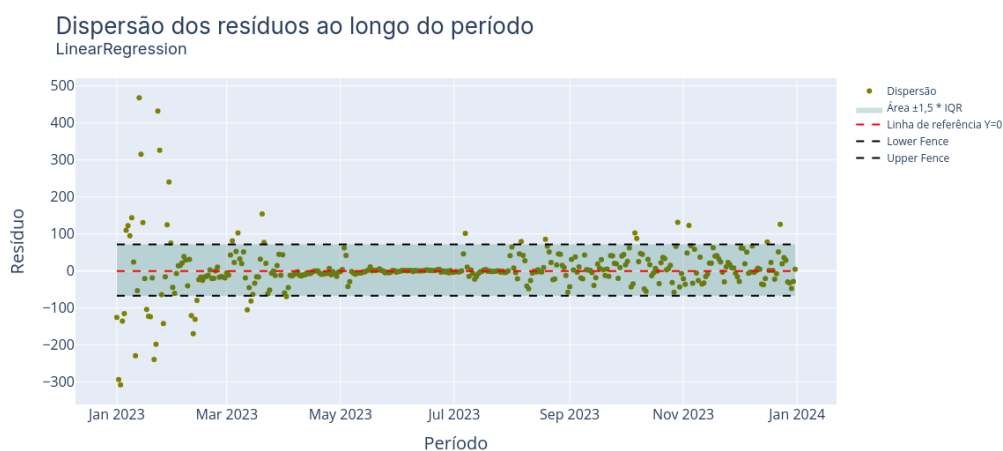


Figura 4.5 – Resíduos da previsão ao longo do tempo.
(fonte: o autor)

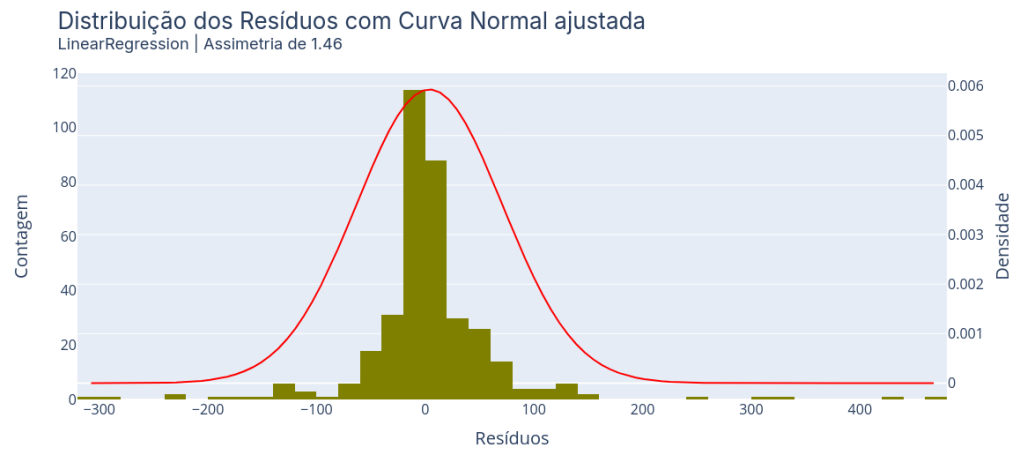


Figura 4.6 – Histograma dos resíduos.
(fonte: o autor)

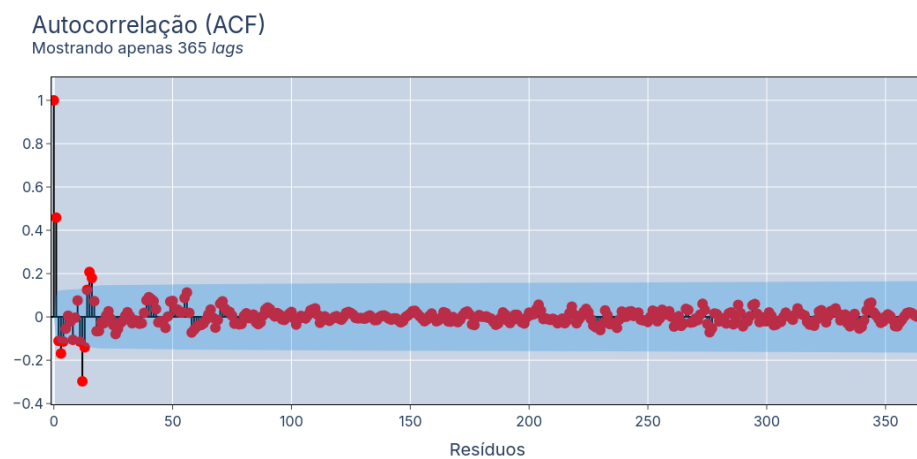


Figura 4.7 – Resíduos da previsão ao longo do tempo.
(fonte: o autor)

Agora os resultados sem emprego da 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 precisar realizar normalização antes empregando algoritmo MinMax. Este algoritmo coloca todos os dados em valores entre 0 e 1, sendo o valor mínimo correspondente a cada série temporal passado para 0 e o valor máximo 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 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 prevalência de faixas elevadas no início do ano, neste resultado o início do ano esteve, digamos, bem comportado. Ao decorrer do ano que os intervalos incrementaram, ali 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 incertezas no período.

Ainda que tenha apresentado uma KGE inferior, bem como nas demais métricas, quando se observa o comportamento dos resíduos, houve uma prevalência de 90,96% dos resíduos na área sombreada no gráfico de dispersão.(figura 4.9) Isso significa que houve menos *outliers* nas previsões. Pode-se verificar também este resultado na figura 4.10, no início do ano, em que menos *outliers* estão presentes, ainda que na porção final do ano tenha aparecido alguns a mais que não foram vistos na análise anterior (figura 4.5). Para os dados originais, ainda que nas métricas pareça piorado, pela análise dos resíduos o modelo mostrou boa estabilidade nas previsões pontuais. Quando se considera os intervalos de previsão, precisa considerar com parcimônia, visto que o crescimento dos valores superiores (hi-95) e valores de vazão $0m^3/s$ na faixa inferior (lo-95) indicam instabilidade de longo prazo.

Para a análise de *delay*, o mesmo feito anteriormente serve para este: aplicar o valor do desvio-padrão parece ser mais coerente com o comportamento real, sendo que aqui houve uma piora significativa no atraso, passando para mais de 3 dias (3,78).

Caminhando para o fim, o histograma e curva-normal apresentou uma assimetria consideravelmente superior (3,38) ao resultado de antes, com presença de cauda longa à direita. O modelo teve tendência de subestimar os resultados, o que é verificável pelo PBIAS. Na figura 4.12 é possível perceber picos fora do intervalo de confiança. Depois de cerca de 10 lags, os pontos de autocorrelação ficam dentro desta faixa, indicando que a maioria dos resíduos após esse ponto não está significativamente correlacionada com valores anteriores. Este pontos vermelhos no início do gráfico indicam que pode haver correlações significativas entre os resíduos com pequenos *lags*, o que sugere, nos primeiros períodos, que os resíduos estão correlacionados com os valores anteriores. Isso é indicativo

de que o modelo pode não estar capturando totalmente a estrutura temporal dos dados. Pode ser necessário ajustar o modelo, adicionar variáveis que expliquem essa dependência temporal, ainda que variáveis de valor acumulado tenham sido inseridas exatamente na intenção de capturar tais comportamentos. Mas claramente precisaria aprimorar.

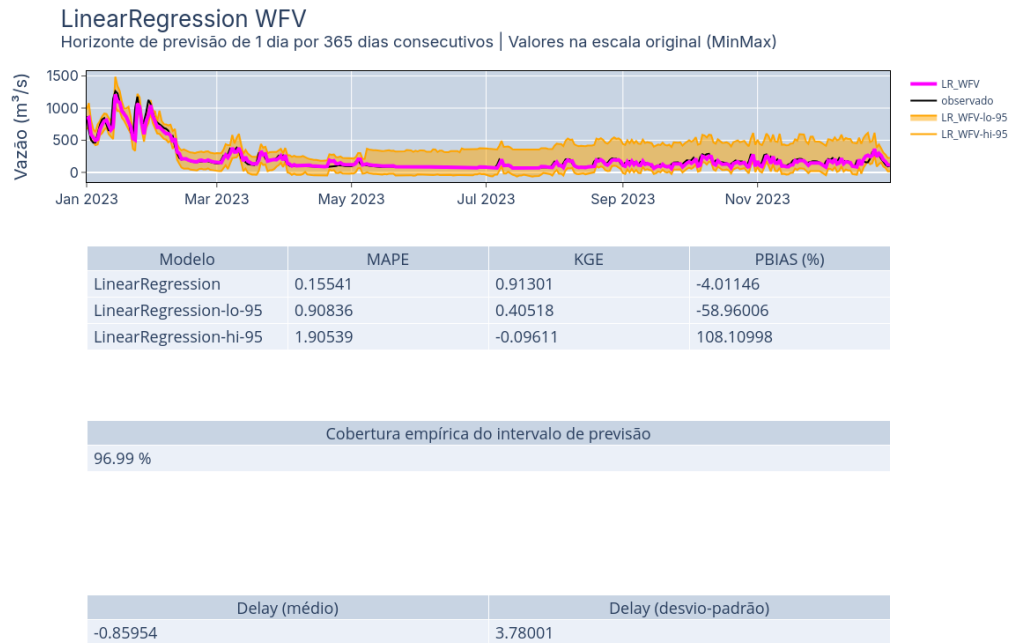


Figura 4.8 – *Walk-Forward Validation* para o modelo Regressão Linear.
(fonte: o autor)

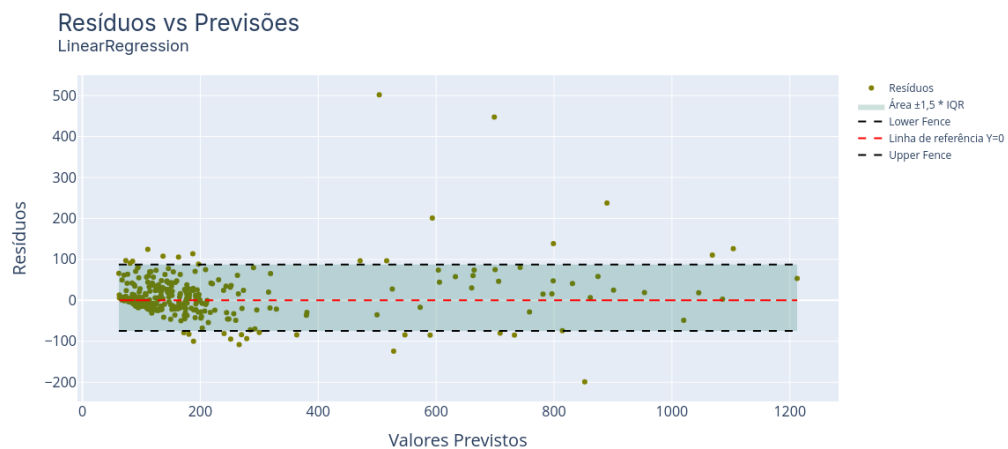


Figura 4.9 – Dispersão dos resíduos.
(fonte: o autor)

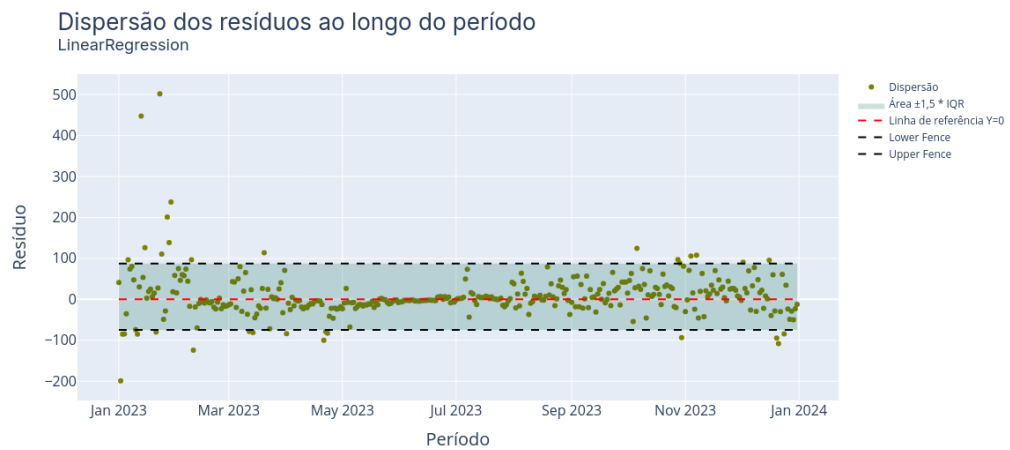


Figura 4.10 – Dispersão dos resíduos ao longo do ano.
(fonte: o autor)

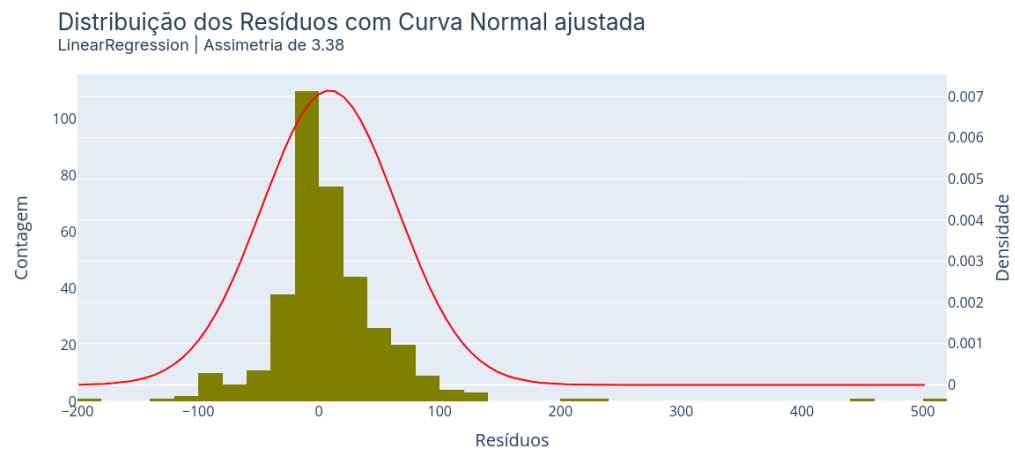


Figura 4.11 – Histograma e curva-normal dos resíduos.
(fonte: o autor)

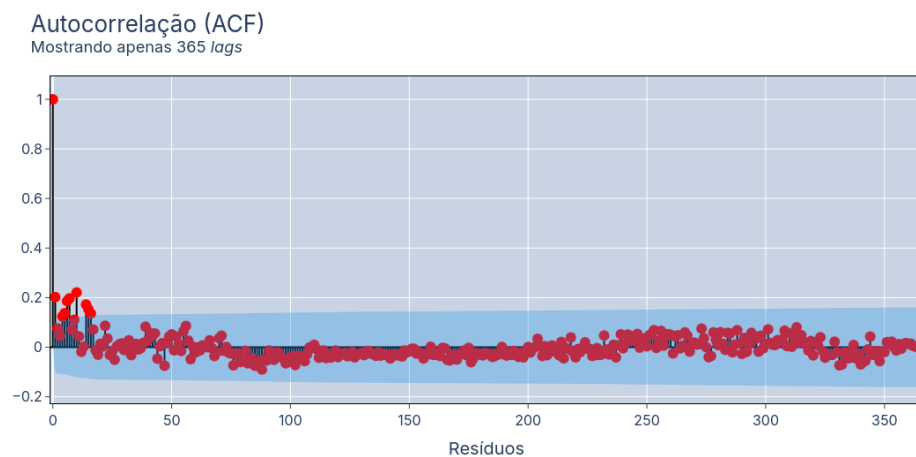


Figura 4.12 – Gráfico ACF dos resíduos.
(fonte: o autor)

Partindo para os modelos principais proposto no trabalho, CatBoost e RandomForest, as análises serão feitas juntas pois ambos apresentaram comportamento assemelhado.

4.2 Importância das variáveis

4.3 Discussão dos resultados

5 CONCLUSÃO E PERSPECTIVAS

5.1 Conclusão

Resumo das principais descobertas da pesquisa

5.2 Contribuições para a área

Destacar as contribuições do estudo para a área de hidrologia (aperfeiçoamento da previsão com uso de atributos categóricos, além das variáveis contínuas; dados de chuva fazendo "ajuste fino" na previsão de vazão).

5.3 Recomendações para Trabalhos Futuros

Sugerir direções para futuras pesquisas baseadas nas descobertas e limitações do meu estudo.

REFERÊNCIAS

- 1 Robert J. Belloto Jr. and Theodore D. Sokoloski. Residual analysis in regression. *American Journal of Pharmaceutical Education*, 49(Fall):295–303.
- 2 Clifford J. Berndt DJ. Using dynamic time warping to find patterns in time series. *AAAI-94 Workshop on Knowledge Discovery in Databases*, pages 359–370, 1994. URL <https://cdn.aaai.org/Workshops/1994/WS-94-03/WS94-03-031.pdf>.
- 3 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.
- 4 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.
- 5 MapBiomias Brasil. Plataforma mapbiomas uso e cobertura, 2024. URL <https://plataforma.brasil.mapbiomas.org/cobertura>. Acessado em: junho de 2024.
- 6 R.T. Clarke. Hydrological prediction in a non-stationary world. *Hydrology and Earth System Sciences*, 11(1):408–414, 2007.
- 7 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.
- 8 Instituto Brasileiro de Geografia e Estatística. Brasil | cidades e estados | ibge brasil |, 2024. URL <https://cidades.ibge.gov.br/>. Acessado em: junho de 2024.
- 9 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.
- 10 Instituto Mineiro de Gestão das Águas. Comitê da bacia hidrográfica do rio araquai, 2024. URL <https://comites.igam.mg.gov.br/comites-estaduais-mg/jq2-cbh-do-rio-aracuai>. Acessado em: junho de 2024.
- 11 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>.
- 12 Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.
- 13 Joaquin Amat Rodrigo e Javier Escobar Ortiz. skforecast, 8 2024a. URL <https://skforecast.org/>.

- 14 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.
- 15 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.
- 16 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.
- 17 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.gh.html>. Acessado em: julho de 2024.
- 18 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/espírito-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.gh.html>. Acessado em: julho de 2024.
- 19 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>.
- 20 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>.
- 21 Rob J. Hyndman. Prediction intervals too narrow, 2014. URL <https://robjhyndman.com/hyndsight/narrow-pi/>. Acessado em: agosto de 2024.
- 22 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.
- 23 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.

- 24 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.
- 25 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.
- 26 Rob J. Hyndman and George Athanasopoulos. Forecasting: Principles and practice (3rd ed.) - 9.1 stationarity and differencing, 2024e. URL <https://otexts.com/fpp3/stationarity.html>. Acessado em: junho de 2024.
- 27 Instituto Mineiro de Gestão das Águas (IGAM). Instituto mineiro de gestão das Águas, 2024. URL <http://www.igam.mg.gov.br/>. Acessado em: agosto de 2024.
- 28 A. Kolling Neto, V.A. Siqueira, C.H.D.A. Gama, R.C.D.D. Paiva, F.M. Fan, W. Collischonn, R. Silveira, C.S.A. Paranhos, and C. Freitas. Advancing medium-range streamflow forecasting for large hydropower reservoirs in brazil by means of continental-scale hydrological modeling. *Water (Switzerland)*, 15(9), 2023.
- 29 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.
- 30 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.
- 31 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. URL https://link.springer.com/chapter/10.1007/978-3-030-57321-8_2.
- 32 Ingrid Petry, Fernando Mainardi Fan, Vinicius Alencar Siqueira, Walter Collishonn, Rodrigo Cauduro Dias de Paiva, Erik Quedi, Cléber Henrique de Araújo Gama, Reinaldo Silveira, Camila Freitas, and Cassia Silmara Aver Paranhos. Seasonal streamflow forecasting in south america’s largest rivers. *Journal of Hydrology: Regional Studies*, 49:101487, 10 2023. ISSN 2214-5818.
- 33 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.
- 34 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.
- 35 Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.
- 36 Scikit-learn. Randomforestregressor - scikit-learn documentation, 2024. URL <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>. Acessado em: abril de 2024.

- 37 Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python, 2010.
- 38 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.
- 39 Yandex. *CatBoost Documentation*, 2024. URL <https://catboost.ai/en/docs/>. Acessado em: junho de 2024.