Construção de um *Dataset* Padronizado em Python a partir de Dados Públicos do SNIRH

Gabriel R. Silva, Lucca Gabriel H. Ferreira, Roni Rodrigues, Claudia A. Martins, Samuel O. S. Bianch

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
CEP: 78069-900 – Cuiabá – MT – Brasil

{gabriel.silva44,lucca.ferreira}@sou.ufmt.br,roni.rodrigues@live.com,

claudia@ic.ufmt.br, samuelbianch38@gmail.com

Abstract. *This work aims to develop a script in the Python programming language to automate both the collection of public data from the SNIRH platform and the construction of a preprocessed dataset to facilitate the development of statistics and artificial intelligence applications. To this end, data collection was performed using the* `Selenium` *library, while data cleaning was carried out with* `Pandas`*. The work also includes two statistical analyses based on public data obtained with the aid of the developed tool. Both analyses discuss data consistency: the first presents a scatter plot between flow and level, while the second addresses a seasonality analysis of the level from July to early October.*

Resumo. *Este trabalho tem como objetivo o desenvolvimento de um script na linguagem de programação Python para automatizar tanto a coleta de dados públicos na plataforma SNIRH, quanto à construção de um dataset préprocessado para facilitar o desenvolvimento de aplicações de inteligência artificial e estatística. Para tanto, a coleta de dados foi feita com a biblioteca* `Selenium`*, enquanto a limpeza dos dados com* `Pandas`*. O trabalho realizou duas análises estatísticas com base em dados públicos obtidos com o auxílio da script desenvolvido. Ambas mostram a consistência dos dados: a primeira apresenta um gráfico de dispersão entre vazão e o nível, enquanto a segunda aborda uma análise de sazonalidade do nível entre os meses de julho a início de outubro.*

1. Introdução

O Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH) é um dos instrumentos de gestão previsto na Política Nacional de Recursos Hídricos, instituída pela lei nº 9.433, de 08 de Janeiro de 1997, conhecida como Lei das Águas. O SNIRH é coordenado pela Agência Nacional de Águas e Saneamento Básico (ANA) e tem como objetivo reunir, organizar e disponibilizar dados sobre os recursos hídricos no Brasil. A coleta dessas informações é realizada por meio de estações hidrotelemétricas. Esses dados são transmitidos por satélite para os servidores da ANA e de outras instituições parceiras, possibilitando o acesso às informações hidrológicas por meio de plataformas públicas de consulta [SNIRH 2023].

As estações hidrotelemétricas são unidades instaladas em campo para o monitoramento automático de variáveis hidrológicas, como nível do rio, precipitação, temperatura e vazão. Essas estações utilizam sensores conectados a *dataloggers* e sistemas de comunicação via satélite ou rede celular para transmitir os dados em tempo real aos sistemas centrais de monitoramento [ANA 2022]. Elas desempenham papel fundamental na geração contínua de dados para o Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH), subsidiando ações de gestão, previsão de cheias, operação de reservatórios e monitoramento ambiental.

No contexto do monitoramento hidrológico, os dados disponibilizados por diferentes centros podem apresentar formatos variados e estruturas heterogêneas, o que requer atenção especial na etapa de preparação para uso em diversas tipos de aplicações. Processos de coleta, limpeza e normalização são fundamentais para tornar esses dados adequados a análises automatizadas embora, muitas vezes, envolvam procedimentos manuais. Nesse cenário, este trabalho tem como objetivo explorar uma abordagem automatizada para auxiliar na padronização e organização desses dados, contribuindo para sua utilização em experimentos e modelos baseados em aprendizado de máquina [Zhang and et al. 2019].

Este artigo está dividido em quatro seções: a primeira seção é sobre os trabalhos relacionados que possuem propostas semelhantes utilizando dados hídricos públicos tanto do Brasil, quanto do exterior; a segunda seção é sobre a metodologia utilizada para a construção do *script*; a seção três apresenta os resultados e discussões acerca do protótipo; e, por fim, a última seção é a conclusão do trabalho.

2. Trabalhos relacionados

Na literatura, trabalhos relacionados com dados aplicados à hidrologia tem enfatizado a importância das etapas de coleta e pré-processamento de dados antes de serem submetidos aos algoritmos de inteligência artificial. Diversos estudos exploram diferentes estratégias e ferramentas para estruturar dados hidrológicos provenientes de fontes públicas.

No trabalho de [Massulo Neto and Nascimento 2024], foram comparados modelos de inteligência artificial para previsão de eventos hidrológicos no estado do Amazonas, com foco em avaliar o desempenho preditivo entre abordagens distintas. Para isso, utilizaram a API HIDROWEB[1] da ANA, a fim de coletar as séries históricas de dados.

No trabalho de [Rodrigues and Martins 2025] foi discutido a integração entre a coleta de dados, seu tratamento e a aplicação de técnicas computacionais em problemas relacionados à hidrologia, evidenciando a importância da automatização no preparo de dados públicos para análise.

De forma semelhante, [Calegario et al. 2024] abordam tanto a coleta quanto o pré-processamento de dados utilizando o pacote `hydrobr`, desenvolvido na linguagem R. O trabalho destaca a importância de ferramentas que integram acesso automatizado a dados com funcionalidades específicas para o tratamento de séries hidrológicas.

No trabalho de [Eng and Wolock 2022] foi utilizado o conjunto de dados GAGES-II[2], disponibilizado pelo United States Geological Survey (USGS), como base para a

[1]Disponível em: `https://www.snirh.gov.br/dadoshidrometereologicos/manual`

[2]Geospatial Attributes of Gages for Evaluating Streamflow (GAGES-II) é um conjunto de dados man-

avaliação do desempenho de algoritmos de aprendizado de máquina na previsão de regimes de vazão. Em uma abordagem mais abrangente, [Hasan et al. 2024] exploraram diversas bases de dados hidrológicos públicas, tais como o CAMELS, Caravan, GRDC e CHIRPS, com o objetivo de investigar a eficácia de diferentes modelos de previsão de vazão, considerando aspectos de generalização e robustez.

3. Metodologia

A metodologia adotada baseou-se na utilização da linguagem de programação `Python` e uso das bibliotecas `Selenium`[3] e `Pandas`. A biblioteca `Selenium` foi utilizada para automatizar o acesso ao site do SNIRH, simulando a interação com a interface web para seleção de estações hidrométricas e intervalos de datas. Com isso, foi possível extrair os dados apresentados na página e armazená-los localmente em arquivos no formato `.csv`.

Após a coleta, os dados foram tratados com o auxílio da biblioteca `Pandas`, organizando-se as informações em uma estrutura tabular. Durante essa etapa, registros com valores inválidos ou ausentes foram identificados e removidos. As colunas foram reestruturadas de forma padronizada, seguindo o esquema: `estacao`, `data_hora`, `nivel`, `chuva` e `vazao`. Por fim, foi construído um *DataFrame* unificado contendo os dados consolidados de cada estação, pronto para aplicações futuras, como interpolação de dados faltantes ou uso em modelos baseados em inteligência artificial.

Neste trabalho, ferramentas baseadas em inteligência artificial generativa, como assistentes de texto, foram utilizadas ao longo do projeto para apoiar na correção e clareza dos parágrafos do artigo, além de auxiliar na refatoração e organização do código-fonte. Essa abordagem contribuiu para melhorar a legibilidade dos textos e otimizar o desenvolvimento das soluções em `Python`.

4. Resultados e discussões

Como resultado deste trabalho, foram desenvolvidos dois *scripts* em `Python`, cada um com finalidades distintas. O primeiro script realiza a coleta automatizada dos dados públicos no site do SNIRH, além da limpeza e padronização dos registros, estruturação em um *DataFrame* e exportação para o formato `.csv`. O segundo script complementa o processo ao realizar a leitura dos arquivos gerados, aplicar etapas adicionais de limpeza e organizar todas as informações em um único dataset consolidado.

Tanto o script desenvolvido utilizando o `Selenium` quanto o pacote `hydrobr` [Calegario et al. 2024] possuem vantagens e desvantagens um em relação ao outro. A biblioteca `Selenium`, por exemplo, possui uma curva de aprendizado mais acentuada do que a do pacote `hydrobr`, além de o código precisar de manutenção caso haja mudanças no HTML da página, o que não acontece utilizando o `hydrobr`, porque obtém os dados por meio de requisições via API. No entanto, o `Selenium` pode ser utilizada em quaisquer plataformas, aumentando a versatilidade da ferramenta, enquanto o `hydrobr` é limitado às fontes de dados da ANA.

---

tido pelo United States Geological Survey (USGS) que reúne atributos geoespaciais e hidrológicos de estações fluviométricas nos EUA. Disponível em: `https://www.usgs.gov/mission-areas/water-resources`

[3]Selenium é um framework de código aberto amplamente utilizado para automação de interações com navegadores web. Documentação oficial disponível em: `https://www.selenium.dev/`

A partir do primeiro *script* foram selecionadas dez estações do Manso para a coleta dos dados. Com a relação à limpeza dos dados, foram excluídos 20 registros considerando todas as estações. Portanto, ao considerar um período de sete dias, foram coletados 1329 registros antes da limpeza e restaram 1309 após a limpeza dos dados.

Com o objetivo de validar preliminarmente os algoritmos de coleta e processamento desenvolvidos, foram realizadas duas análises estatísticas exploratórias utilizando a biblioteca `Pandas`. Essas análises tiveram caráter de teste inicial, buscando avaliar se os dados gerados estavam consistentes com o comportamento esperado de séries hidrológicas reais. A primeira análise consistiu na construção de um gráfico de dispersão (Figura 1), também conhecido como curva-chave, com o objetivo de investigar a correlação entre nível e vazão. O coeficiente de correlação calculado foi de 0,9989, indicando uma forte relação linear positiva. Este resultado é condizente com o esperado do ponto de vista físico e sugere consistência nos dados obtidos.

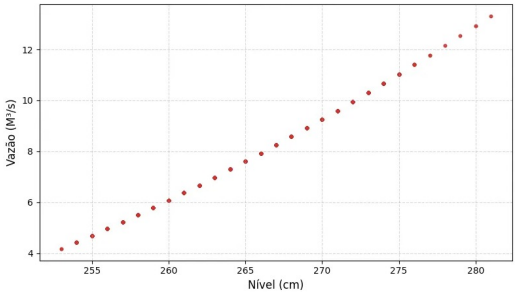

Figura 1. Gráfico de dispersão entre nível e vazão.

A segunda análise teve como foco a variação sazonal dos níveis registrados. Na Figura 2 é apresentada a distribuição dos dados entre os meses de julho a outubro de 2025. Observa-se que o mês de julho apresenta níveis médios mais elevados e menor variação, caracterizando um período mais úmido. Em contraste, setembro apresenta valores medianos menores e maior presença de *outliers* refletindo, provavelmente, um período mais seco e irregular. Esses padrões reforçam a expectativa hidrológica regional e evidenciam que os dados extraídos e tratados pelo *pipeline* automatizado mantêm coerência com o comportamento natural dos sistemas monitorados.

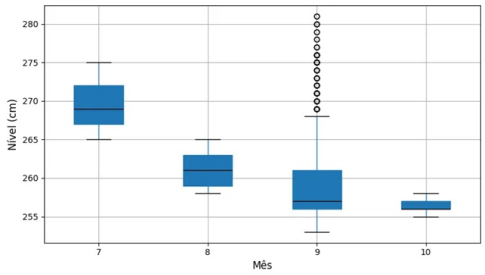

Figura 2. Sazonalidade do nível por mês.

5. Conclusão

Este trabalho apresentou o desenvolvimento de um *script* em `Python` com o objetivo de automatizar a coleta, limpeza e padronização de dados hidrológicos públicos provenientes

do SNIRH. Como resultado, foi gerado um *dataset* estruturado e pronto para análises exploratórias e aplicações futuras, especialmente no contexto de modelagem preditiva em aprendizado de máquina.

Foram realizadas duas análises estatísticas com o objetivo de conduzir testes preliminares sobre os algoritmos desenvolvidos para coleta e processamento. Os resultados obtidos mostraram consistência e coerência com o comportamento esperado dos dados hidrológicos, indicando que os *scripts* cumpriram adequadamente seu propósito técnico. A forte correlação entre nível e vazão, bem como a identificação de padrões sazonais nos níveis observados, reforçam a validade do pipeline proposto.

Como trabalhos futuros, pretende-se ampliar a funcionalidade do sistema para incluir múltiplas variáveis hidrológicas, integrar métodos de interpolação de falhas e valores ausentes e adaptar os dados coletados para uso direto em modelos de aprendizado de máquina. Adicionalmente, propõe-se explorar visualizações interativas e a publicação do *dataset* em plataformas públicas com metadados completos, visando fomentar a reutilização por outros pesquisadores.

Referências

ANA (2022). Manual técnico de operação de estações hidrometeorológicas. `https://www.gov.br/ana/pt-br/assuntos/`. Acesso em: out. 2025.

Calegario, A. T., Assis, C. M. d. A., Brumano, C. L., Damasceno, C. R. A., Aragao, J., Amorim, R. S. S., Moreira, M. C., da Silva, D. D., Vaz Junior, G., and Lourenco, A. M. G. (2024). Download e pre-processamento de dados hidroclimaticos do Hidroweb/ANA com o pacote Hydrobr. Acesso em: out. 2025.

Eng, K. and Wolock, D. (2022). Evaluation of machine learning approaches for predicting streamflow metrics across the conterminous United States. Technical Report 2022-5058, U.S. Geological Survey, Reston, VA.

Hasan, F., Medley, P., Drake, J., and Chen, G. (2024). Advancing hydrology through machine learning: Insights, challenges, and future directions using the camels, caravan, grdc, chirps, persiann, nldas, gldas, and grace datasets. *Water*, 16(13).

Massulo Neto, G. and Nascimento, E. L. B. d. (2024). Mudanças climáticas: inteligência artificial na previsão de eventos hidrológicos extremos no estado do amazonas. *Cuadernos de Educación y Desarrollo*, 16(13):e6849.

Rodrigues, R. and Martins, C. (2025). Modelos de aprendizado na predição de nível de rios para detecção de falhas em estações hidrotelemétricas. In *Anais do Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA)*. Disponível em: `https://sol.sbc.org.br/index.php/wcama/article/view/36108`. Acesso em: out. 2025.

SNIRH, A. (2023). Sistema nacional de informações sobre recursos hídricos (snirh). `https://www.snirh.gov.br/`. Acesso em: out. 2025.

Zhang, Z. and et al. (2019). Data preprocessing in machine learning: A review. In *Proceedings of the International Conference on Artificial Intelligence and Data Processing (IDAP)*, pages 1–5. IEEE.