AQUAVIS: UMA APLICAÇÃO STREAMLIT PARA DETECÇÃO E CORREÇÃO DE ANOMALIAS EM SÉRIES TEMPORAIS HIDROLÓGICAS COM INTELIGÊNCIA ARTIFICIAL

Gustavo de Oliveira Macedo[1], Flavio Augusto Altieri dos Santos[2], Tiago Alves da Fonseca[3], Manuel Nascimento Dias Barcelos Júnior[4], Jorge Andrés Cormane Angarita[5],

[1]Universidade de Brasília, Brasília, Brasil (gustavoomacedo@outlook.com.br)
[2]MD/CENSIPAM/CRBE/, Brasil
[3]Universidade de Brasília, Brasília, Brasil
[4]Universidade de Brasília, Brasília, Brasil
[5]Universidade de Brasília, Brasília, Brasil

Resumo: Este artigo apresenta o AquaVIS, uma aplicação Streamlit voltada para a identificação e correção de valores discrepantes em séries temporais de cotas de rios da bacia amazônica, utilizando modelos de IA Darts. A previsão precisa dos níveis dos rios é vital para agências como o CENSIPAM e a ANA, permitindo alertas antecipados de secas e inundações. A ferramenta aborda problemas de qualidade de dados decorrentes de erros de coleta ou mau funcionamento de sensores. Ademais, permite o upload de dados do usuário ou importação automatizada, fornece insights sobre as séries e permite a detecção de anomalias usando modelos pré-treinados com sensibilidade ajustável. O sistema contempla a corrição de anomalias através de técnicas de periodicidade e interpolação (linear, polinomial, spline), visualizando as alterações antes da aplicação. Este trabalho oferece uma solução inovadora para melhorar a qualidade dos dados de séries temporais, aumentando, em última análise, a previsibilidade de eventos hidrológicos extremos

Palavras-chave: inteligência artificial; aplicação; dados hidrológicos; séries temporais.

INTRODUÇÃO

Historicamente, a inteligência artificial (IA) tem navegado por um cenário de altas expectativas e, por vezes, frustrações. Contudo, rupturas recentes no conhecimento científico impulsionaram o uso exponencial da IA globalmente, tornando-a mais acessível e atraente para a indústria (Fell, 2024). Esse ciclo virtuoso de investimento e desenvolvimento aprimora processos em diversas áreas, resultando em economia de tempo, maior qualidade e impactos positivos no bem-estar geral. Diante desse cenário, torna-se crucial explorar o vasto potencial da IA em campos cada vez mais diversos e estratégicos.

No contexto brasileiro, a Bacia Amazônica, a maior bacia hidrográfica do mundo, exige monitoramento e conservação rigorosos para prevenir desastres ambientais e socioeconômicos. A Agência Nacional de Águas (ANA), por meio do Hidroweb, disponibiliza uma vasta base de dados hidrometeorológicos, incluindo cotas de rios, coletadas pela Rede Hidrometeorológica Nacional (RNH). No entanto, apesar da importância desses dados, a coleta apresenta desafios. Tanto as estações modernas (sensores automáticos) quanto as rudimentares (registros manuais) estão sujeitas a falhas: defeitos em sensores, dados não coletados ou coletados erroneamente, e erros de medição que podem persistir por longos períodos. Essas inconsistências comprometem a qualidade dos dados e, consequentemente, a precisão das análises e previsões necessárias para a gestão hídrica. As previsões dos comportamentos hídricos é extremamente importante para a mitigação de problemas ambientais e socioeconômicos, uma vez que as autoridades responsáveis podem tomar medidas profiláticas antes de comportamentos

naturais extremos, protegendo populações ribeirinhas, a fauna e a flora locais. Dessa forma, faz-se necessário o aumento da eficiência na etapa de pré-processamento de tais dados.

É nesse ponto crítico que o AquaVIS se insere, apresentando uma solução robusta para o pré-processamento de dados hidrológicos. Reconhecendo a capacidade consolidada da IA na identificação de padrões em séries temporais, esta ferramenta foi concebida para identificar e, posteriormente, corrigir dados anômalos (outliers) advindos de erros grosseiros ou sensoriais na coleta.

O presente estudo explora estratégias e técnicas na etapa de pré-processamento de dados hidrológicos da bacia amazônica com uso de inteligência artificial. Tal exploração se deve ao fato de que existe grande consolidação no uso de modelos de IA para identificação de padrões, inclusive, em séries temporais. Dessa forma, pode-se treinar modelos para identificar padrões em séries temporais de cotas de rios e apontar, quando necessário, pontos da série temporal que apresentam valores que se desviem muito do padrão esperado para o rio analisado, sugerindo um erro (grosseiro ou sensorial) na coleta dos dados. Tal apontamento pode facilitar a reparação e/ou limpeza de dados brutos, preparando-os para etapas póstumas de processamento e tomada de decisão, revelando-se, portanto, uma importante ferramenta de impacto econômico e socioambiental. Entretanto, sabe-se que o uso de modelos de IA traz consigo alguns riscos, como baixa performance, overfitting, underfitting, etc.

Na aplicação em pauta, é possível afirmar que os modelos devem minimizar a presença de falsos negativos. Em outras palavras, o modelo deve acusar a maior quantidade de pontos com erros de coleta possível, mesmo que acuse, com certa frequência, pontos sem erros de coleta. Ao minimizar a presença de falsos negativos na detecção de anomalias – ou seja, acusando a maior quantidade possível de pontos com erros, mesmo que alguns pontos sem erros sejam indicados – a aplicação garante que a maior parte dos dados problemáticos seja tratada, contribuindo para a melhoria da qualidade dos dados e, consequentemente, da previsibilidade de eventos hidrológicos extremos, como secas e cheias, essenciais para a proteção de comunidades ribeirinhas e ecossistemas.

A linguagem Python revela-se como uma importante ferramenta para o desenvolvimento de tal aplicação, uma vez que é alicerce de diversos frameworks de renome, como é o caso do Streamlit (framework para criação de aplicações web de alto nível). Além disso, a linguagem Python se destaca por sua grande comunidade de usuários, permitindo a criação de bibliotecas open-source de alta confiabilidade, eficiência e performance. Neste quesito, é possível citar as bibliotecas Darts (para identificação de anomalias em séries temporais) Pandas, Numpy, Scikit-Learn, dentre outras. O presente projeto utiliza a linguagem Python para criar uma aplicação em Streamlit, onde é possível manipular séries históricas de dados através da biblioteca Pandas, e identificar e corrigir anomalias através da biblioteca Darts.

MATERIAIS E MÉTODOS

I. Visão Geral

Visando a otimização da etapa de pré-processamento de séries temporais históricas de cotas de rios da bacia amazônica, apresenta-se, neste projeto, o AquaVIS, uma aplicação inédita que utiliza a biblioteca Darts, uma biblioteca gratuita e open-source (Herzen, 2022) para acessar modelos de inteligência artificial (IA) capazes de identificar anomalias em séries temporais, permitindo a correção pontual de cada inconsistência identificada. A proposta da aplicação é tornar a etapa de pré-processamento das séries temporais de cotas em um processo muito mais eficiente, detalhado e rico, uma vez que o sistema realiza a detecção automática de inconsistências, além de permitir o ajuste granular dos dados da série analisada.

O sistema, construído em linguagem Python, utiliza o framework Streamlit para a construção de uma aplicação web de fácil uso e alto nível. Através deste framework, é possível a criação de páginas web interligadas, cada uma consistindo de um conjunto de elementos, como botões, caixas de texto, gráficos, etc. O framework se conecta a diversas bibliotecas, permitindo a fácil integração da aplicação com componentes externos.

A aplicação permite, inicialmente, que o usuário insira a série temporal via upload ou via consulta por código de estação. Quando o usuário opta por consultar por código de estação, uma vez que o código é inserido, inicia-se uma cascata de automações, sendo estas responsáveis por acessar o portal HidroWEB, pesquisar pela estação digitada, realizar o download do arquivo que comporta a série temporal (espera-se um arquivo no formato Microsoft Database - MDB), tratar o arquivo através de consultas SQL e, finalmente, obter a base de dados tratável pelo sistema. A automação descrita é realizada através do Selenium, uma poderosa ferramenta open-source disponível, dentre outras, para a linguagem Python. Ao final da automação a base de dados é convertida em um objeto manipulável da biblioteca Pandas (dataframe) e, portanto, também tratável pelo Streamlit. Uma vez que a série temporal está carregada e operacional, o sistema entra no núcleo da aplicação. A figura 1, disponível abaixo, ilustra o fluxo de trabalho da aplicação, servindo para referência dos mecanismos expostos em sequência.

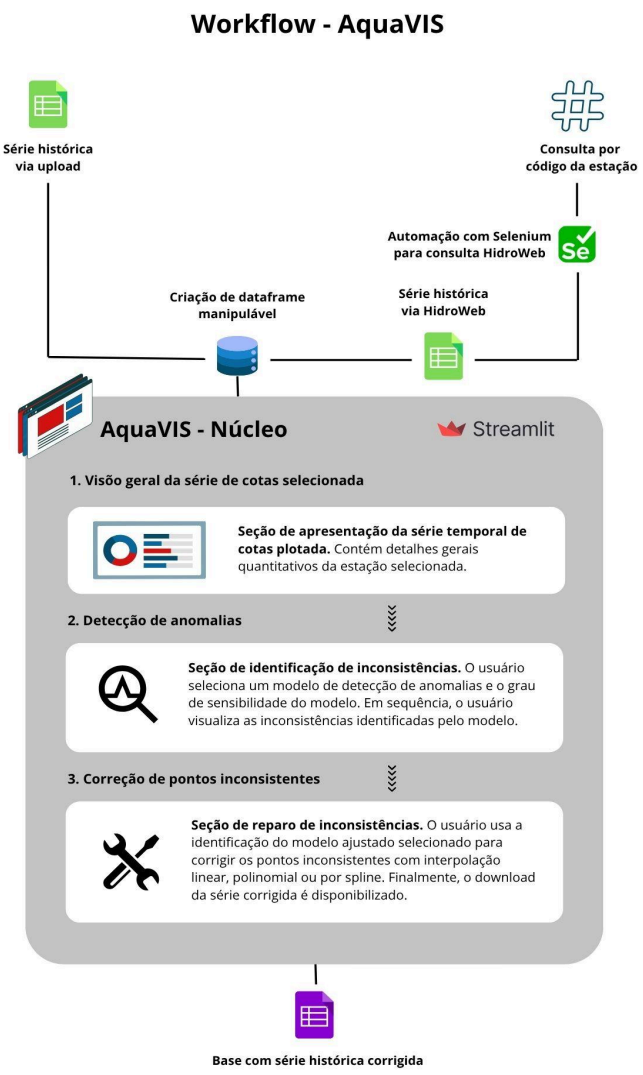

Figura 1 - Fluxo de trabalho do AquaVIS

II. Visão Técnica

O núcleo da aplicação é composto de 3 seções, conforme exposto abaixo:

Visão geral da série de cotas selecionada (SEÇÃO 1): nesta seção, o sistema apresenta a série temporal plotada em um gráfico interativo (gerado através da conexão entre o Streamlit e a biblioteca Plotly, biblioteca de análise de dados disponível no Python), além da exibição de dados gerais referentes à distribuição presente na série, como o número de registros diários na série, o número de dados faltantes e o valor médio da cota. É possível observar, na Figura 2, a série temporal original da estação de Manacapuru, exibida na tela da primeira seção da

aplicação. Cabe ressaltar que as figuras abaixo retratam apenas partes das telas da aplicação. Para acessar a aplicação por inteiro, consulte os anexos deste projeto.

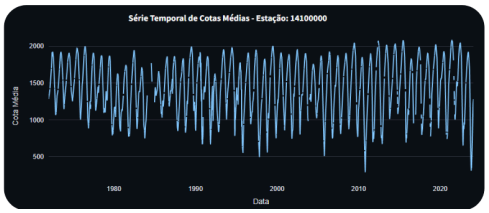

Figura 2 - Série temporal de Manacapuru na seção 1

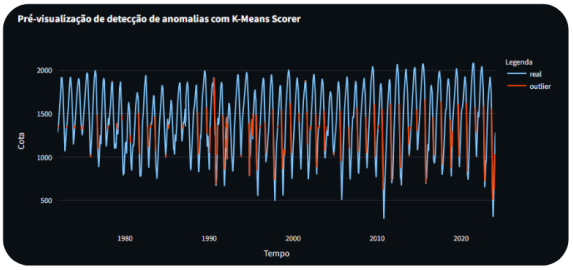

Figura 3 - Identificação de anomalias na série temporal de Manacapuru com uso do K-Means Scorer.

Detecção de anomalias (SEÇÃO 2): nesta seção, o sistema acessa um conjunto de modelos pré-definidos da biblioteca Darts. Para a versão alpha (apresentada no presente artigo), foram selecionados e treinados dois modelos: K-Means e Auto Encoder, seguindo as diretrizes e achados experimentais para o treinamento de modelos voltados à identificação de anomalias em séries temporais de cotas de rios da bacia amazônica, evidenciados por Macedo et al. (2025). O usuário pode, então, selecionar um dos modelos e ajustar sua sensibilidade a anomalias, através de um slider que modifica o quantil do detector de anomalias. O quantil do detector de anomalias é um hiperparâmetro dos modelos pontuadores (scorers) da biblioteca Darts (Herzen, 2022). Quanto maior o quantil, mais rigoroso é o modelo, apenas evidenciando anomalias às quais estão associados valores mais altos de certeza. Em outras palavras, o modelo apontará menos anomalias, pois aponta apenas os pontos em que exista mais confiança de que tal ponto é uma anomalia real. Uma vez que o usuário define um modelo e um nível para o quantil de detecção de anomalias, o sistema plota o gráfico que evidencia a detecção de anomalias do modelo e quantil selecionados, além de mostrar a quantidade total de anomalias detectadas, a quantidade de anomalias em pontos extremos da série temporal e o desvio padrão das anomalias detectadas. Portanto, o usuário visualiza a identificação de pontos inconsistentes na série por meio do modelo selecionado e utiliza dos dados adicionais para ajustar o valor do quantil e a seleção do modelo. Ao se encerrar a seleção do modelo e ajuste do quantil para o nível desejado, o usuário pode avançar para a terceira seção.

Correção de pontos inconsistentes (SEÇÃO 3): nesta última seção, o sistema utiliza a identificação de anomalias do modelo selecionado na seção 2 para organizar um processo que permite a correção de tais inconsistências de forma granular. Inicialmente, é solicitado ao usuário que selecione uma periodicidade para realização das correções, podendo esta ser diária, mensal ou anual. O sistema cria, em seguida, um conjunto de janelas temporais com respeito à periodicidade selecionada (por exemplo, se o usuário seleciona a periodicidade mensal, o sistema cria uma lista de janelas de 30 dias em que existam inconsistências, com base na identificação de anomalias obtida na seção 2). Em seguida, o sistema apresenta ao usuário a primeira janela do conjunto de janelas, oferecendo a correção da janela atual através de métodos pré-definidos. São eles: interpolação linear, interpolação polinomial e interpolação por spline. No caso das interpolações polinomiais e por spline, é solicitado ao usuário a ordem dos polinômios (por exemplo, se a ordem for igual a 3, a interpolação é cúbica). Após selecionar um método de interpolação, o sistema realiza a interpolação ao deletar os dados inconsistentes e realizar a interpolação na série restante. O usuário pode, enfim, visualizar a interpolação e compará-la à mesma janela da série original. Caso deseje manter a interpolação, o usuário pode clicar em um botão que realiza a substituição da série original pela série interpolada, mas apenas para a janela em análise. Caso o usuário deseje manter a série original, basta avançar para a próxima janela. O processo é iterativo e se encerra quando o usuário desejar, dando-lhe liberdade para acessar e corrigir janelas não sequencialmente. Além disso, a qualquer momento do processo de correção, o usuário pode realizar o

download da série corrigida até então, obtendo uma nova base.

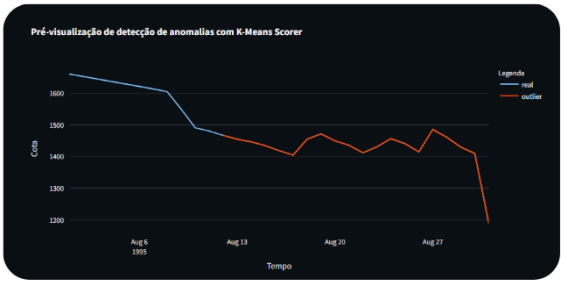

Figura 4 - Uma das janelas mensais da série de Manacapuru durante a correção (seção 3).

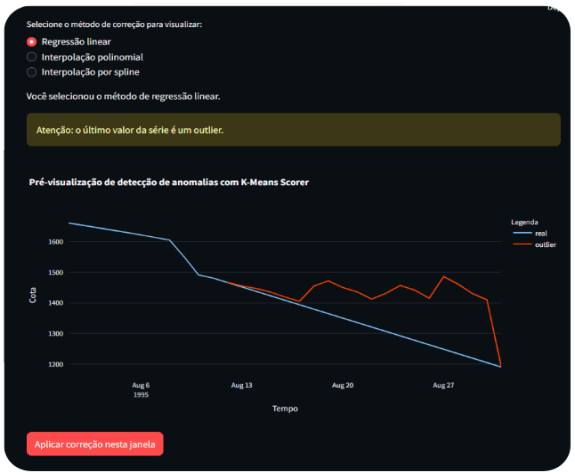

Figura 5 - Visualização da aplicação da regressão linear durante correção da janela da Figura 4.

Dessa forma, o AquaVIS consiste em uma poderosa ferramenta de pré-processamento de séries temporais, possibilitando a quantificação, localização e correção de dados anômalos em séries temporais. Entretanto, ressalta-se que os modelos empregados foram treinados com dados de séries temporais de cotas de rios localizados na bacia amazônica. Isso implica na limitação do contexto da aplicação, configurando-se como uma ferramenta especialista em dados hidrológicos dessa região.

RESULTADOS E DISCUSSÃO

Apesar de utilizar ferramentas simples, a composição dos elementos da aplicação a tornam complexa. É importante notar a relevância das aplicações open-source no que se refere à acessibilidade, facilidade de manutenção e oferta de suporte pela comunidade. De acordo com Rayhan (2023), a linguagem Python é extremamente popular por conta de 3 principais fatores: a facilidade de se aprender, a versatilidade da linguagem e o tamanho da comunidade, permitindo a criação de diversas ferramentas e centros de suporte. Este projeto reflete essas características em todos os aspectos. A manutenção é simples e não exige conhecimento rebuscado de técnicas de programação. Além disso, as ferramentas utilizadas são altamente populares e possuem documentações extensas e ricamente escritas. Por fim, o estilo de programação adotado para esta aplicação foi o estilo pipeline, conforme proposto por Lopes (2014), o que configura alta compatibilidade com o mercado, visto que esse estilo é amplamente utilizado e divulgado mundo afora.

A aplicação proposta surge como uma solução inovadora para a ineficiência do pré-processamento de dados de séries temporais de cotas de rios. Métodos manuais para a identificação de dados anômalos (ou outliers) podem ser demorados, suscetíveis a erros humanos e impraticáveis em conjuntos massivos de dados. Abordagens puramente estatísticas, por sua vez, embora muito úteis, carecem da capacidade de capturar padrões complexos e sutis, os quais ferramentas de IA podem ajudar a discernir.

O AquaVIS ataca tais dificuldades ao oferecer um fluxo de trabalho semi-automatizado e interativo, dando ao usuário uma gama de possibilidades no que se refere à variabilidade de ajustes, técnicas de análise de dados e correção granular de anomalias.

O fluxo de trabalho semi-automatizado justifica-se na automação na aquisição de dados, eliminando a necessidade de download e manipulação complexa de arquivos, uma vez que a aquisição de bases de dados com granularidade diária via HidroWeb é altamente dificultosa, pois os dados disponíveis, à primeira vista, possuem periodicidade mensal. Para se obter o conjunto correto de dados com periodicidade diária (fator essencial para a análise de anomalias de natureza hídrica, em se tratando de cotas de rios), é necessário realizar consultas complexas em tabelas específicas. Em outras palavras, o HidroWeb não permite a fácil obtenção de tais dados, sendo este um ponto de vulnerabilidade que o AquaVIS busca eliminar através da automação com a ferramenta Selenium.

A alta variabilidade de ajustes encontra-se na possibilidade de selecionar modelos, alterar a sensibilidade de tais modelos, selecionar diferentes métodos de correção e, finalmente, optar por corrigir ou não uma janela da série. Isso leva a um novo nível no que se refere a correção personalizada de tais dados. Combinado com a especialidade do analista responsável pela correção, a etapa de pré-processamento de tais dados se torna potencialmente mais rápida, eficiente e fácil.

Adicionalmente, a correção granular empodera a união entre análises visuais e automáticas, dado que torna possível a correção de dados inconsistentes que não foram identificados pelo modelo selecionado.

CONCLUSÃO

Conclui-se, portanto, que o AquaVIS é potencialmente capaz de impulsionar a eficiência da etapa de pré-processamento de dados no tocante a séries temporais de cotas de rios da bacia amazônica, configurando-se como uma ferramenta gratuita, de código aberto, de fácil manutenção e interpretabilidade; oferecendo aos usuários a possibilidade de selecionar entre diferentes modelos de IA, ajustes de sensibilidade dos modelos, técnicas de correção de anomalias e correção granular de séries temporais.

LIMITAÇÕES E TRABALHO FUTURO

Apesar do AquaVIS ser uma poderosa ferramenta no auxílio ao pré-processamento de dados, ainda existem limitações que devem ser trabalhadas em atualizações futuras.

As principais limitações do AquaVIS são:

- Susceptibilidade da ferramenta a erros provenientes de treinamentos com dados não suficientemente abrangentes. Uma possível solução é o treinamento em conjuntos maiores de dados ou abrir a possibilidade do usuário realizar os treinamentos com as séries que desejar.

- Deploy da ferramenta pode não ser gratuito, principalmente no que se refere a uma aplicação a ser inserida em ambiente de produção escalável e de acesso público. Uma possível solução é optar por sistemas de deploy baratos e seguros, como a própria nuvem do Streamlit (Streamlit Cloud) e, com a popularização da ferramenta, migrar para um sistema mais robusto em termos de escalabilidade e eficiência.

- Limitação no número de modelos de detecção de anomalias, se restringindo, até o momento, a apenas dois. Tal problema ocorre pela complexidade de se realizar testes exaustivos e na necessidade de maior poder computacional para o treinamento e otimização de hiperparâmetros de novos modelos. Uma possível solução seria a implementação lenta e gradual de novos modelos com base em testes feitos com modelos já implementados, aproveitando o conhecimento adquirido nos testes de treinamento.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (ANA). Região Hidrográfica Amazônica. ANA, [s.d.]. Disponível em: https://www.gov.br/ana/pt-br/assuntos/gestao-das-aguas/panorama-das-aguas/regioes-hidrograficas/regiao-hidrografica-amazonica.

BRASIL. Agência Nacional de Águas e Saneamento Básico (ANA). Serviços e Informações do Brasil. 26 de maio de 2022. Disponível em: https://www.gov.br/pt-br/orgaos/agencia-nacional-de-aguas. Acesso em: 30 de janeiro de 2025.

FELL, James. The current state of AI, according to Stanford's AI Index. World Economic Forum, 26 abr. 2024. Emerging Technologies. Disponível em: https://www.weforum.org/stories/2024/04/stanford-university-ai-index-report/. Acesso em: 31 jan. 2025.

HERZEN, Julien. et al. Darts: User-Friendly Modern Machine Learning for Time Series. Journal of Machine Learning Research, Suíça, v.23, n.124, p.1-6, mar. 2022. Disponível em: https://www.jmlr.org/papers/v23/21-1177.html. Acesso em: 15 jan. 2025.

LOPES, Cristine. Exercises in Programming Style. New York. Chapman and Hall/CRC, 2020.

ZHAO, Yue. NASRULLAH, Zain. LI, Zheng. PyOD: A Python Toolbox for Scalable Outlier Detection. Journal of Machine Learning Research, Pensilvânia, Estados Unidos, v.20, n.96, p.1-7, maio, 2019.

Disponível em: https://www.jmlr.org/papers/volume20/19-011/19-011.pdf. Acesso em: 26 jan. 2025.