XXVI Simpósio Brasileiro de Recursos Hídricos
23 a 28 de novembro de 2025 - Vitória - ES

ABRHidro
Associação Brasileira de Recursos Hídricos

XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

IDENTIFICAÇÃO E CORREÇÃO DE INCONSISTÊNCIAS EM SÉRIES DE DADOS HIDROLÓGICOS

Emilly Meireles Braga Cornelio [1]; Hugo Pires Procópio [2]; Flávio Augusto Altieri dos Santos[3]; Tiago Alves da Fonseca[4]; Manuel Nascimento Dias Barcelos Júnior[5] & Jorge Andrés Cormane Angarita [6]

Abstract: The identification of outliers in hydrological time series is essential to ensure the quality of forecasting models used for the prevention of extreme hydrological events. In this context, an algorithm is proposed for the automatic detection of contextual outliers in river level data from the Amazon basin, based on the identification of nearby stations and the use of regression methods with sliding windows. For the detection of point and collective outliers, eleven different methods are employed, relying on the detection of abrupt variations and the definition of thresholds. The performance of these methods is evaluated by counting correct classifications (true positives and true negatives) and errors (false positives and false negatives). This evaluation shows that most of the analyzed methods achieve accuracy rates above 90%. In addition, regression-based methods for correcting these outliers are presented. This approach contributes to the overall improvement of hydrological data quality and enhances the reliability of analyses and decisions based on these records.

Resumo: A identificação de dados inconsistentes em séries temporais hidrológicas é imprescindível para garantir a qualidade dos modelos de previsão para prevenção de eventos hidrológicos extremos. Nesse contexto, é proposto um algoritmo para identificação automática de inconsistências contextuais em dados de nível de rios da bacia amazônica, considerando a identificação de estações próximas e métodos de regressão com janelas deslizantes. Para o caso das inconsistências pontuais e coletivas, utiliza-se 11 métodos distintos baseados na detecção de variações abruptas e na definição de limiares. O desempenho desses métodos é avaliado contabilizando os acertos (positivos verdadeiros e negativos verdadeiros) e erros (falsos positivos e falsos negativos). Essa avaliação mostra que a maioria dos métodos analisados apresenta eficiência acima de 90%. Além disso, são apresentados métodos de correção dessas inconsistências baseados em regressão. Essa abordagem contribui para a melhoria geral da qualidade dos dados hidrológicos e aumenta a confiabilidade das análises e decisões baseadas nesses registros.

Palavras-Chave – Detecção Automatizada; Adequação Supervisionada; Valores Atípicos; Análise Contextual; Gestão de Recursos Hidrológicos; Amazonia.

1) Faculdade de Ciências e Tecnologias em Engenharia, Universidade de Brasília (UnB), emilly.mbcornelio@gmail.com
2) Faculdade de Ciências e Tecnologias em Engenharia, Universidade de Brasília (UnB), hugopprocopio@gmail.com
3) CENSIPAM/CRBE, flavio.santos@sipam.gov.br
4) Faculdade de Ciências e Tecnologias em Engenharia, Universidade de Brasília (UnB), fonsecafga@unb.br
5) Faculdade de Ciências e Tecnologias em Engenharia, Universidade de Brasília (UnB), manuelbarcelos@unb.br
6) Faculdade de Ciências e Tecnologias em Engenharia, Universidade de Brasília (UnB), jcormane@unb.br

INTRODUÇÃO

A identificação de inconsistências em séries de dados hidrológicos é importante para garantir a qualidade dos mesmos. Isto, aprimora a gestão dos recursos hídricos e auxilia na prevenção e mitigação dos efeitos de desastres naturais (Pereira, Barbieiro e Quevedo, 2020). Os dados inconsistentes aparecem nas séries temporais de forma pontual e coletiva, e podem ser identificados por meio de métricas estatísticas convencionais, porém, em alguns casos mais complexos, se faz necessário comparar os dados da série temporal analisa com os dados advindos de estações de medição adjacentes. Nesse último caso, as inconsistências são chamadas de contextuais.

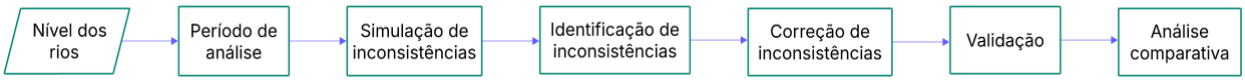
As inconsistências nos dados podem ser causadas por erro humano, mal funcionamento dos sensores ou até obstruções no rio, quando se trata de dados de vazão. Jamshidi et al. (2021) estudaram a implementação de métodos de forma individual e combinada para este caso, concluindo uma menor eficiência no caso da aplicação dos métodos combinados, já que o aumento das detecções pode levar à propagação de erros. Bae e Ji (2019), por sua vez, se basearam na média móvel exponencialmente ponderada para promover a remoção desses dados e a suavização dos mesmos como forma de melhorar a qualidade dos dados de nível de córregos. Na aplicação do método, os autores determinaram o tamanho ideal da janela deslizante por meio da diferença relativa entre a mediana móvel e a média móvel após a remoção das inconsistências. Gleeson et al. (2023) destacam a necessidade de escolher um algoritmo que seja apropriado à distribuição dos dados. Nesse sentido, no estudo de inconsistências em dados de águas subterrâneas, Kim et al. (2022) avaliaram varias adaptações de métodos bastante utilizados para a identificação de inconsistências de forma a adequá-los para distribuições de dados não-Gaussianas.

Considerado as premissas supracitadas, diferentes métodos e combinações são analisados para identificação e correção de inconsistências pontuais, coletivas e contextuais. Por fim, para validação dos mesmos, os métodos são validados por meio da substituição de 10 observações por valores aleatórios. A eficiência de cada método é calculada considerando os conceitos de precisão e sensibilidade partindo das quantidades de falsos positivos, falsos negativos e positivos verdadeiros encontrados.

METODOLOGIA

A metodologia usa neste trabalho é apresentada na Figura 1 na forma de fluxograma dos procedimentos implementados neste estudo.

Figura 1 - Fluxograma do estudo



As etapas decorrentes para a implementação prática do fluxograma que representa a metodologia são apresentadas a seguir:

i. Importação dos dados de nível fluviométrico;
ii. Escolha do período de análise, em que não há inconsistências identificadas;
iii. Substituição de pontos aleatórios por valores inconsistentes;
iv. Aplicação dos métodos de identificação de inconsistências pontuais, coletivas e contextuais;

v. Correção das inconsistências identificadas;
vi. Avaliação dos resultados.

A validação dos métodos de identificação de inconsistências pontuais e coletivas foi realizada a partir do cálculo de métricas como: Precisão, Sensibilidade, Eficiência. As métricas propostas consideram as quantidades de Falsos Positivos (FP), Falsos Negativos (FN) e Positivos Verdadeiros (TP), de acordo com as equações 1, 2 e 3. Já os métodos de identificação de inconsistências contextuais e de correção são validados visualmente nas estações hidrológicas de Manacapuru, Anamã Manaus em períodos distintos.

$$Precisão = \frac{TP}{TP+FP} \quad (1)$$

$$Sensibilidade = \frac{TP}{TP+FN} \quad (2)$$

$$Eficiência = 2\frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \quad (3)$$

Métodos Baseados no Cálculo de Limiares para Inconsistências Pontuais e Coletivas

Os métodos reportados na literatura utilizam parâmetros da estatística descritiva para o cálculo de limiares, tal e como mostrado na Tabela 1.

Tabela 1 - Métodos baseados em limiares

Método	Cálculo do Limiar	Equação
Z-Score Padrão (Jamshidi, et al., 2022)	$limiar = \mu \pm 3\sigma$	(4)
Z-Score Modificado com decomposição (Jamshidi, et al., 2022)	$limiar = \bar{x} \pm \frac{3MAD}{0{,}6745}$	(5)
Z-Score Modificado* (Berendrecht, van Vliet e Griffioen, 2023)	$limiar = \bar{r} \pm 3{,}7599MAD$	(6)
Z-Score com Média Móvel Exponencial (Jamshidi, et al., 2022)	$limiar = MME \pm \frac{3MAD}{0{,}6745}$	(7)
Intervalo Interquartílico (Tukey, 1977)	$limiar_{sup} = Q_3 + 1{,}5 \times IQR$ $limiar_{inf} = Q_1 - 1{,}5 \times IQR$	(8) (9)
Intervalo Interquartílico Modificado (Jeong, et al., 2017)	$limiar_{sup} = \mu + Q_3 + 1{,}5 \times IQR$ $limiar_{inf} = \mu - Q_1 - 1{,}5 \times IQR$	(10) (11)
Sigma Modificado (Kim, et al., 2022)	$\sigma_{up} = \min(f_{up} - \mu)$ $\sigma_{low} = \min(\mu - f_{low})$ $limiar_{sup} = \mu + 3\sigma_{up}$ $limiar_{sup} = \mu - 3\sigma_{low}$	(12) (13) (14) (15)

Em que, $\mu, \sigma, \bar{x}, Q_1, Q_3$ são a média, o desvio, a mediana, o primeiro e terceiro quartil das medições, respetivamente. $\bar{r}$ é a mediana dos resíduos entre as medições e as médias locais. MAD é desvio absoluto mediano. IQR é o intervalo interquartílico ($Q_3 - Q_1$. f_{up} e f_{low} são os vetores com valores correspondentes aos pontos superior e inferior no intervalo quantílico [0,5, 0,95]. σ_{up} e σ_{low} são os graus de dispersão direcional.

Métodos Baseados na Cálculo da Variação para Inconsistências Pontuais e Coletivas

Esses métodos utilizam a primeira e a segunda derivada para detectar variações abruptas nos dados, que indicariam que os dados são inconsistentes. A primeira derivada é usada para medição da variação dos dados, e a segunda derivada identifica as variações abruptas. É definido o limite de 20 cm/dia^2 para essa análise.

A técnica das janelas deslizantes é utilizada para avaliar padrões locais e aumentar a sensibilidade da detecção de inconsistências. A série temporal foi dividida em janelas de 15 dias com passo de 7 dias para garantir a sobreposição das janelas para uma análise mais precisa (Yu et al., 2014).

Além disso, utilizou-se o desvio absoluto mediano (MAD) como medida de dispersão dos dados na forma de uma terceira modificação do Z-Score de acordo com a equação (16). Para o dado ser considerado inconsistente é necessário escolher um critério de rejeição para M_i. Os valores sugeridos são: 3 (muito conservador), 2,5 (moderadamente conservador) e 2 (pouco conservador). Neste trabelho, valores acima de 2,5 são considerados como inconsistentes.

$$M_i = \frac{|x_i - \bar{x}|}{\text{MAD}} \tag{16}$$

Para garantir maior robustez, utiliza-se o estimador S_n, de acordo com a equação (17), em conjunto com a aplicação do Z-Score Modificado às segundas derivadas, conforme descrito na equação (18) (Rousseeuw e Croux, 1993).

$$S_n = 1{,}1926\left(|x_i - x_j|\right) \tag{17}$$

$$Z_i = \frac{|d_i - \bar{d}|}{S_n} \tag{18}$$

Esses métodos são empregados em diferentes combinações: derivada, janela deslizante com MAD nas medições, janela deslizante com MAD na derivada, janela deslizante com Sn na derivada.

Métodos para Identificação de Inconsistências Contextuais

Para a identificação de inconsistências contextuais se faz necessário identificar as estações de medição próximas à estação de interesse, o que é feito por meio das coordenadas geográficas de acordo com a equação de Haversine, descrita na equação (19).

$$d = 2r\,sin^{-1}\left(\sqrt{sin^2\left(\frac{lat_2 - lat_1}{2}\right) + \cos(lat_1)\cos(lat_2)\,sin^2\left(\frac{long_2 - long_1}{2}\right)}\right) \tag{19}$$

Onde, d é a distância entre os dois pontos geográficos, r é o raio da Terra em km, lat_1 e lat_2 são as latitudes dos pontos em radianos. $long_1$ e $long_2$ são as longitudes dos pontos em radianos.

Tal inconsistência é identificada quando um segmento da série temporal da estação em análise, devidamente ajustado no tempo, apresenta comportamento discrepante em relação às séries temporais de estações de medição adjacentes.

A abordagem proposta envolve, também, a aplicação de janelas deslizantes junto ao coeficiente de correlação de Pearson e do coeficiente angular da regressão linear. O coeficiente de correlação pode ser incorporado como um parâmetro no método da janela deslizante para investigar a correlação entre as curvas de estações adjacentes. Essa aplicação requer a definição de um limite inferior para a correlação entre as curvas, sendo que janelas que exibem correlação acima desse limite são identificadas como inconsistentes. As janelas que exibem coeficiente de correlação acima de 0,3 são consideradas inconsistentes.

Uma maneira convencional de saber como um dado se relaciona com outro, é calcular a regressão linear entre eles. É possível assumir que o coeficiente angular da regressão entre duas estações adjacentes sofre pouca alteração ao longo da série temporal. Portanto, podemos usar a janela deslizante ao longo da série temporal para avaliar a variação do coeficiente angular da regressão linear, e definir como inconsistente uma janela em que o coeficiente angular apresenta uma grande variação em relação a série inteira.

O método do coeficiente angular consiste no cálculo dos coeficientes da regressão linear entre duas estações adjacentes (b_1) e dentro de cada janela (a_1). Calcula-se o ângulo entre as duas regressões de acordo com a equação (20). Se o ângulo calculado for acima de 15, a janela é considerada inconsistente.

$$\alpha = \tan^{-1}(b_1) - \tan^{-1}(a_1) \tag{20}$$

Métodos de correção de inconsistências

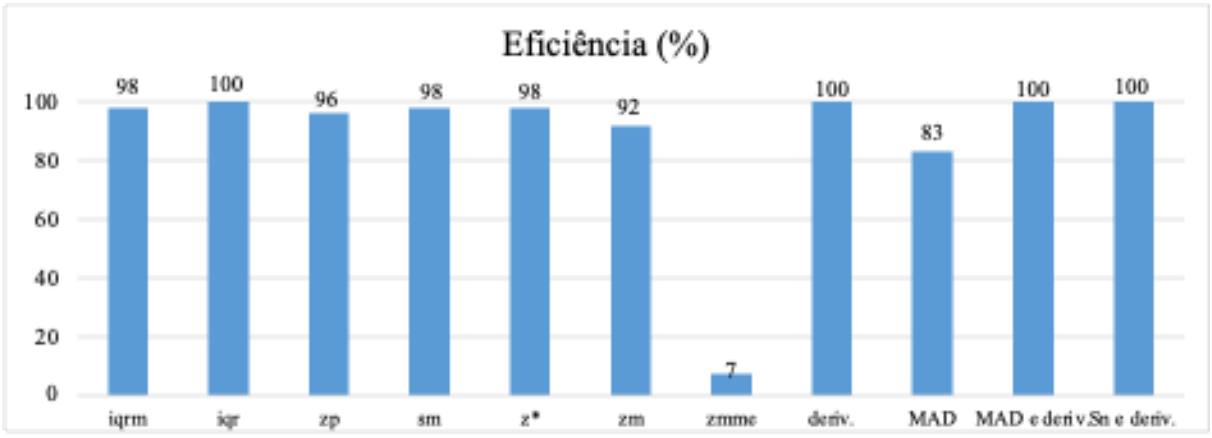
Para o caso de inconsistências pontuais e coletivas consiste em eliminar as janelas consideradas inconsistentes e substituir esses dados pela técnica de *spline* cúbica e de regressão polinomial. Esses métodos são comparados em relação ao coeficiente de determinação r^2.

Para as inconsistências contextuais se faz necessário usar dados da estação adjacente para corrigir os dados por meio de interpolação e regressão com defasagem temporal. Considerando a correlação entre as estações de medição, é possível estimar um valor por meio da medição tomada em uma estação próxima. Entretanto, um evento que ocorre em uma estação só faz efeito na estação a jusante depois de um determinado tempo. Para estimar a melhor defasagem entre os dados das estações, utiliza-se o coeficiente de correlação linear. Tendo o intervalo em que há o maior coeficiente de correlação, emprega-se a regressão aos dados defasados para estimar os dados da estação de referência.

RESULTADOS

A eficiência dos métodos de identificação de inconsistências pontuais e coletivas é apresentado no gráfico da Figura 2.

Figura 2 - Eficiência dos métodos utilizados para identificação de inconsistências pontuais e coletivas

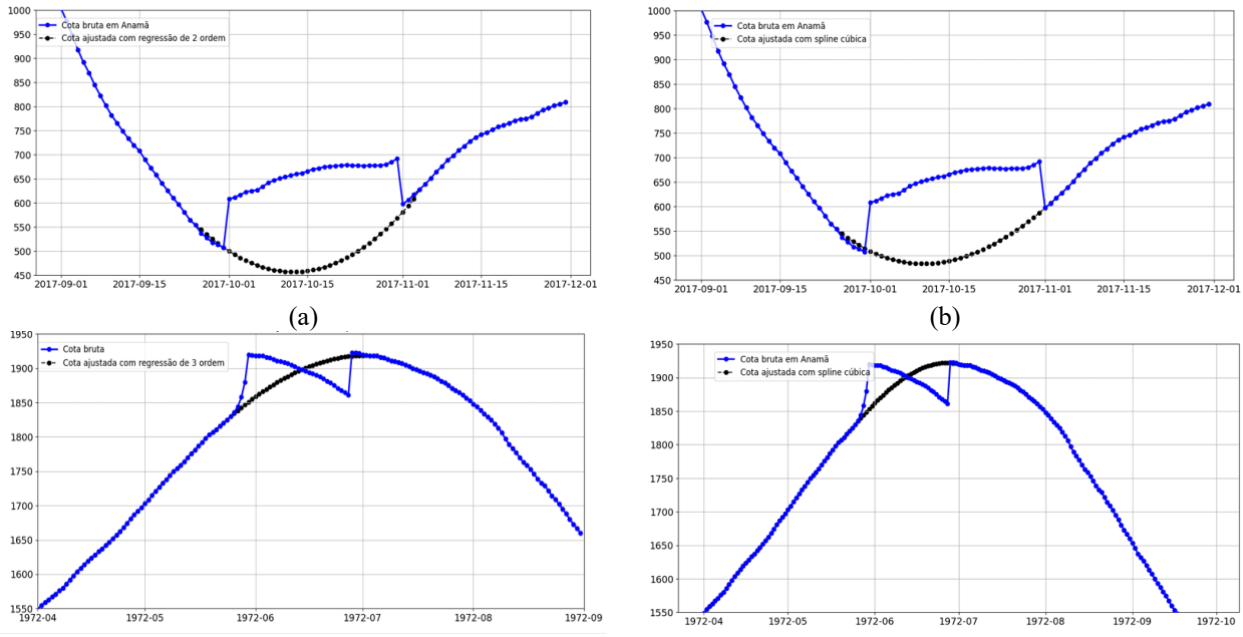


Os métodos apresentados são: intervalo interquartílico modificado (iqrm), intervalo interquartílico (iqr), z-score padrão (zp), sigma modificado (sm), z-score modificado* (z*), z-score modificado (zm), z-score com média móvel exponencial (zmme) e derivada (deriv).

Os resultados dos testes indicam que a maioria dos métodos apresentam respostas confiáveis, com eficiência acima de 90%, com boa precisão e sensibilidade. A baixa precisão do z-score com média móvel exponencial, é possivelmente causada pela grande incidência de falsos positivos, o que sugere a necessidade de uma adaptação para que o método se adeque aos dados estudados.

A Figura 3 mostra os resultados obtidos para correção de inconsistências pontuais e coletivas por *spline* cúbica e regressão polinomial da estação de Anamã. A figura 3(a) apresenta a correção dos dados com regressão polinomial. A Figura 3(b) mostra correção dos dados com *spline* cúbica. A Figura 3(c) apresenta a correção dos dados com regressão polinomial. A Figura 3(d) mostra a correção dos dados com *spline* cúbica.

Figura 3. Correção de inconsistências pontuais e coletivas por Spline Cúbica e Regressão Polinomial da estação de Anamã



(a) (b)

(c) (d)

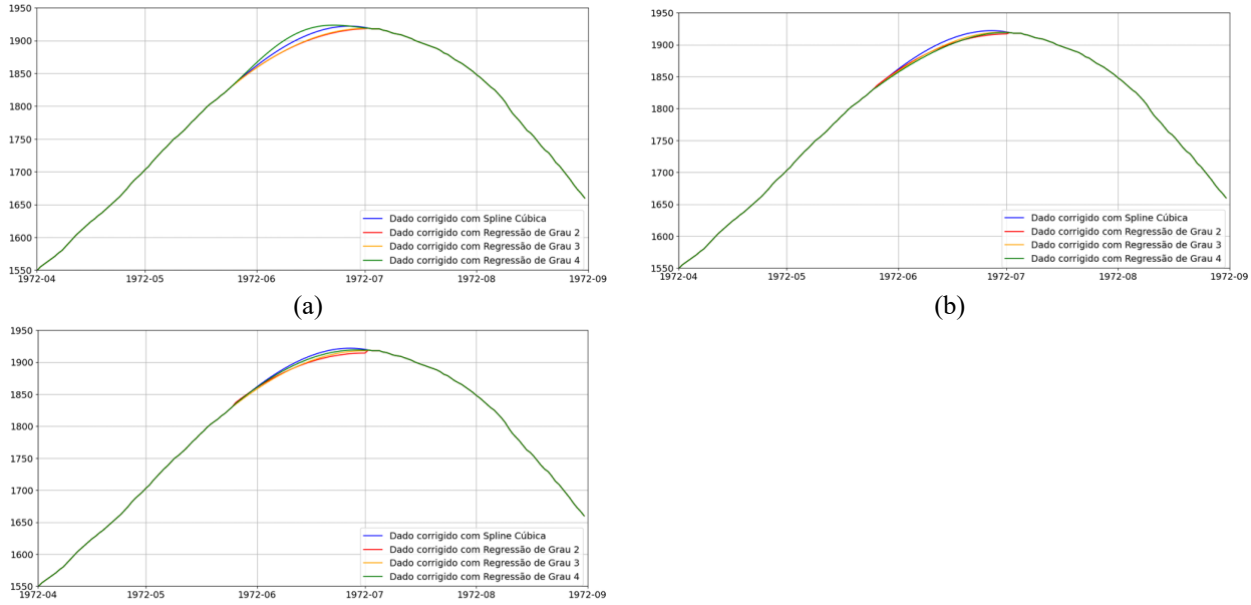
Os resultados dos cálculos para o coeficiente de determinação decorrentes da correção de inconsistências pontuais e coletivas por *Spline* Cúbica e Regressão Polinomial da estação de Anamã são apresentados na Tabela 2.

Tabela 2 – Cálculo do r^2 para a *Spline* Cúbica e Regressão Polinomial

Pontos	**Método**	**Regressão grau 2 (%)**	**Regressão grau 3 (%)**	**Regressão grau 4 (%)**
10	Em relação com a *Spline*	99,5434	99,5859	98,1023
10	regressão	99,9943	99,9958	99,9980
20	Em relação com a *Spline*	99,7254	99,6680	99,2392
20	regressão	99,9195	99,9916	99,9922
40	Em relação com a *Spline*	99,7820	99,6240	99,9224
40	regressão	99,8293	99,9794	99,9871

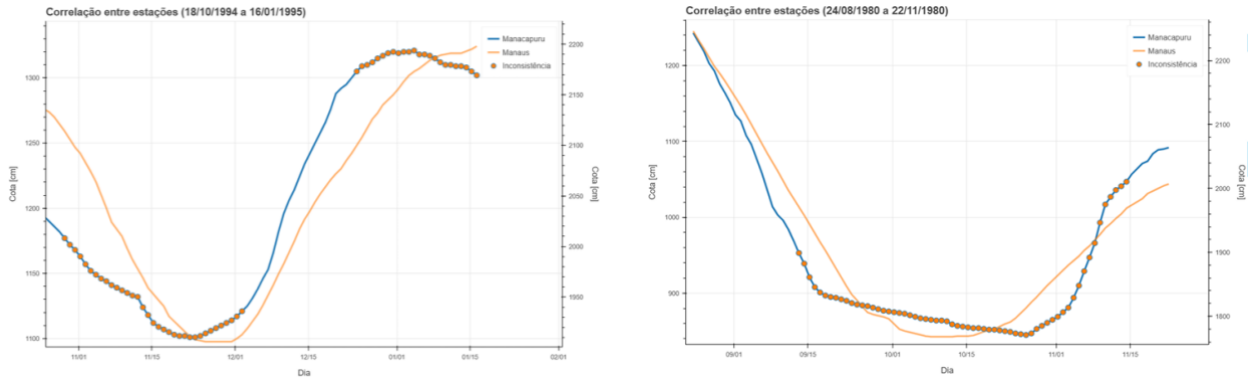
Para a regressão com 10 pontos totais, é possível ver que a regressão de 4° grau se aproxima melhor dos dados originais, mas se afasta do resultado da *Spline* cúbica. Enquanto as regressões de grau 2 e 3 se aproximam tanto dos dados originais quanto da *Spline* cúbica. Já com 20 pontos, em relação ao caso anterior, o r^2 em relação ao dado original para todas as regressões diminuiu. Porém, todos as correções se aproximaram da *Spline* cúbica, com um aumento significativo no r2 da regressão de grau 4. Com 40 pontos utilizados para a regressão, as curvas se afastaram mais do dado original, gerando uma descontinuidade no gráfico da regressão de 2° grau, mesmo com uma proximidade maior ao resultado da *Spline* cúbica. Os resultados anteriores são ilustrados na Figura 4. A Figura 4(a) regressão com 10 pontos totais; 4(b) regressão com 20 pontos e 4(c) regressão com 40 pontos.

Figura 4 - Correção de inconsistências pontuais e coletivas com diferentes quantidades de pontos

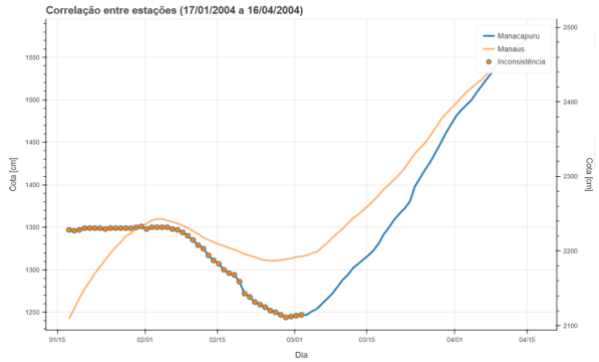


Na Figura 5, são apresentados os resultados da detecção de inconsistências contextuais. Três segmentos das series temporais da estação de Manacapuru foram avaliados com os dados da serie temporal da estação adjacente de Manaus. É possível notar visualmente nas Figuras 4(a), 4(b) e 4(c), as diferenças entre as séries temporais no trecho indicado como inconsistente, comprovando a coerência dos resultados obtidos.

Figura 5 - Detecção de inconsistências contextuais



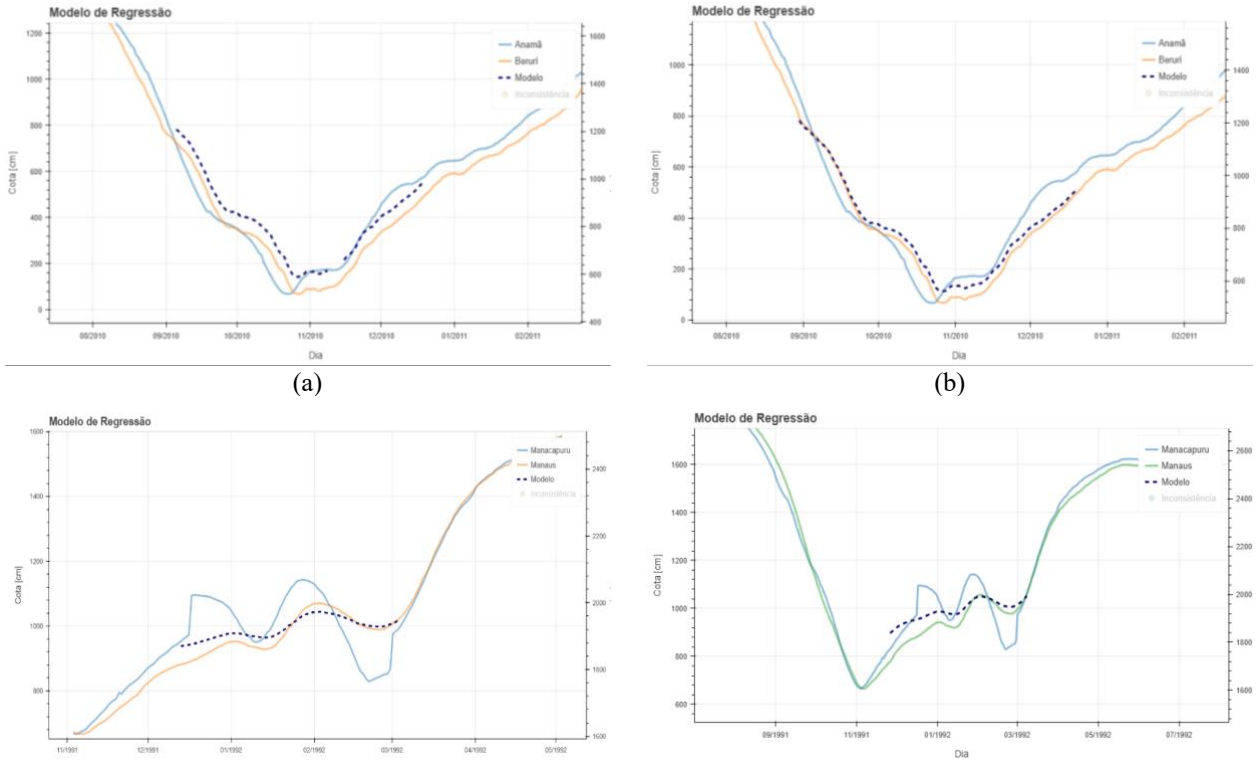
(a) (b)



(c)

Os resultados obtidos para correção de inconsistências contextuais, por sua vez, são apresentados na Figura 6, que compara a correção com e sem defasagem temporal. Figura 6(a) correção por regressão com defasagem temporal em Anamã (2010). Figura 6(b) correção sem defasagem em Anamã (2010). Figura 6(c) correção com defasagem para Manacapuru (1992). Figura 6(d) correção sem defasagem para Manacapuru (1992). Nota-se que os dois casos apresentam bons resultados, mas com melhor ajuste quando se considera defasagem temporal.

Figura 6 - Correção de inconsistências contextuais com e sem defasagem temporal



CONSIDERAÇÕES FINAIS

Por meio da validação proposta, foi possível provar a eficácia dos métodos para a identificação de inconsistências em séries temporais de dados hidrológicos, destacando-se como ferramentas indispensáveis para garantir a confiabilidade dos dados.

Na análise comparativa, notaram-se as elevadas sensibilidade e precisão da maioria dos métodos, resultando em uma alta eficiência. A exceção é o z-score combinado com média móvel exponencial, provando que o uso de métodos combinados pode levar à propagação de erros, a notar pela grande quantidade de falsos positivos. Para este caso indica-se a necessidade de ajustes para que a técnica seja aplicável aos dados estudados.

Quanto à correção de inconsistências pontuais, a análise comparativa entre as técnicas de correção utilizando regressão polinomial e *Spline* Cúbica. A análise revela que cada abordagem possui vantagens e limitações que devem ser consideradas. A *Spline* Cúbica demonstra maior consistência na suavização e correção das inconsistências, já que essa técnica garante continuidade nos dados. As regressões polinomiais também mostram capacidade de ajuste dos dados, se aproximando da correção com *Spline* Cúbica. Porém, ao utilizar um maior número de pontos no modelo, podem gerar descontinuidades que podem comprometer a qualidade da correção.

A correção de inconsistências contextuais apresentou bons resultados com os dois métodos propostos, mas com maior suavidade do ajuste na análise com defasagem temporal.

AGRADECIMENTOS - O presente estudo foi realizado no âmbito do Termo de Execução Descentralizada Nº 06/2023, celebrado entre a UnB e o CENSIPAM, cujo objeto é "Desenvolvimento de ferramenta computacional baseada em métodos matemáticos para a detecção automática e a adequação supervisionada de inconsistências em séries temporais de dados coletados de postos de medição fluviométrica (nível, vazão, medição de descargas liquida, calibração de curva chave) localizadas na Bacia Amazônica. Processo Nº 60092.000169/2023-12.

REFERÊNCIAS

BAE, I.; JI, U. "*Outlier detection and smoothing process for water level data measured by ultrasonic sensor in stream flows*". Water, v. 11, n. 5, p. 951, 2019, https://www.mdpi.com/2073-4441/11/5/951.

BERENDRECHT, W., VAN VLIET, M. & GRIFFIOEN, J. (2023). "*Combining statistical methods for detecting potential outliers in groundwater quality time series*". Environmental Monitoring and Assessment, v. 195, n. 1, p. 85, https://doi.org/10.1007/s10661-022-10661-0.

GLEESON, K.; HUSBAND, S.; GAFFNEY, J.; BOXALL, J. (2023). "*A data quality assessment framework for drinking water distribution system water quality time series datasets*". AQUA-Water Infrastructure, Ecosystems and Society, v. 72, n. 3, p. 329-347, https://doi.org/10.2166/aqua.2023.228.

JAMSHIDI, E.; YUSUP, Y.; KAYODE, J.; KAMARUDDIN, M. (2022). "*Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on*

surface water temperature". Ecological Informatics, v. 69, p. 101672, https://doi.org/10.1016/j.ecoinf.2022.101672.

JEONG, J; PARK, E,; HAN, W. S.; KIM, K; CHOUNG, S.; CHUNG, I. (2017) "*Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends*", Journal of Hydrology, v. 548, p. 135-144, https://doi.org/10.1016/j.jhydrol.2017.02.058.

KIM, K; JEONG, J; PARK, H; KWON, M; CHO, C; JEONG, J. (2022). "*Development of a data-driven ensemble regressor and its applicability for identifying contextual and collective outliers in groundwater level time-series data*", Journal of Hydrology, v. 612, Part A, https://doi.org/10.1016/j.jhydrol.2022.128127.

PEREIRA, M.; BARBIEIRO, B.; QUEVEDO, D. (2020). "*Importance of river basin monitoring and hydrological data availability for the integrated management of water resources*". Sociedade & Natureza, v. 32, p. 292-303. Disponível em: https://doi.org/10.14393/SN-v32-2020-43458.

ROUSSEEUW, P. J.; CROUX, C. (1993). "*Alternatives to the median absolute deviation. Journal of the American Statistical association*", v. 88, n. 424, p. 1273-1283, https://doi.org/10.1080/01621459.1993.10476408.

TUKEY, J. (1977). "*Exploratory data analysis*". Addison-Wesley.

Yu, Y.; Zhu, Y.; Li, S.; Wan, D. (2014). "*Time series outlier detection based on sliding window prediction*". Mathematical problems in Engineering, v. 2014, p. 879736.