CONTROLE DE QUALIDADE DE DADOS PLUVIOMÉTRICOS MENSAIS COM SELEÇÃO HÍBRIDA DE ESTAÇÕES DE APOIO VIZINHAS

Walszon Terllizzie A. Lopes [1]; Elton Vinícius Krauss[1]; Gabriel Meldau Lemos[1]; Gabriel Reis de Vasconcelos[1]; Henrique Dantas Borges[1]; João Carlos Carvalho[1]; Taiara Sampaio Santos[1] & Taína Martins Magalhães[1]

RESUMO - Este trabalho apresenta uma metodologia automatizada de controle de qualidade (*Quality Control* - QC) aplicada a dados pluviométricos mensais, baseada na seleção híbrida de estações de apoio vizinhas. A abordagem combina critérios de proximidade espacial, correlação histórica e disponibilidade amostral para definir estações de referência e construir uma estimativa regional de precipitação para comparação com cada estação analisada. A rotina calcula indicadores de suporte espacial, resíduos robustos, envelopes sazonais, faixas de aceitabilidade e níveis hierárquicos de suspeição, diferenciando registros críticos, suspeitas importantes, evidências complementares e avisos operacionais. A aplicação resultou na análise de 20.565 registros mensais, distribuídos em 53 estações, no período de 1912 a 2025. Foram classificados como suspeitos 2.926 registros, correspondentes a 14,2% da base, dos quais 1.133 apresentaram *Score* igual ou superior a 3, indicando maior prioridade para revisão técnica. A estação 00340118 - CURRAL VELHO foi utilizada como exemplo para ilustrar os produtos gráficos gerados, incluindo mosaicos de auditoria, gráficos de controle de qualidade e mapas espaciais de precipitação, resíduos e anomalias. Os resultados demonstram que a metodologia é útil para triagem preliminar e priorização da análise de dados pluviométricos mensais.

ABSTRACT - This paper presents an automated quality control (QC) methodology applied to monthly rainfall data, based on a hybrid selection of neighboring support stations. The approach combines spatial proximity, historical correlation, and sample availability criteria to define reference stations and build a regional rainfall estimate for comparison with each target station. The routine computes spatial support indicators, robust residuals, seasonal envelopes, acceptability ranges, and hierarchical suspicion levels, distinguishing critical records, important suspicions, complementary evidence, and operational warnings. The application analyzed 20,565 monthly records from 53 rainfall stations, covering the period from 1912 to 2025. A total of 2,926 records were classified as suspicious, corresponding to 14.2% of the dataset, of which 1,133 presented a Score equal to or greater than 3, indicating higher priority for technical review. Station 00340118 - CURRAL VELHO was used as an example to illustrate the graphical products generated by the routine, including audit mosaics, quality control charts, and spatial maps of rainfall, residuals, and anomalies. The results show that the proposed methodology is useful as a preliminary screening tool to support the prioritization of monthly rainfall data analysis.

Palavras-Chave - Controle de qualidade; precipitação mensal; estações vizinhas.

[1] Agência Nacional de Águas e Saneamento Básico – ANA. Setor Policial, Área 5, Quadra 3, Bloco L. Fone: (61) 2109 5476. E-mail: walszon@ana.gov.br; elton.krauss@ana.gov.br; gabriel.lemos@ana.gov.br; gabriel.vasconcelos@ana.gov.br; henrique.borges@ana.gov.br; joao.carvalho@ana.gov.br; taiara.santos@ana.gov.br; taina.magalhaes@ana.gov.br.

1 INTRODUÇÃO

A disponibilidade de séries pluviométricas consistentes é fundamental para estudos hidrológicos, climatológicos e de gestão de recursos hídricos. Dados mensais de precipitação são utilizados na caracterização de regimes sazonais, avaliação de secas e cheias, estudos de disponibilidade hídrica, regionalização hidrológica, planejamento de redes de monitoramento e apoio à tomada de decisão. Entretanto, séries observadas em redes hidrometeorológicas extensas estão sujeitas a inconsistências associadas a falhas de leitura, erros de digitação, problemas de transcrição, lacunas temporais, valores acumulados indevidamente, falsos zeros e registros extremos incompatíveis com o comportamento regional.

Em redes pluviométricas de abrangência regional ou nacional, a revisão manual de todos os registros é operacionalmente limitada, sobretudo quando envolve longos períodos históricos e grande número de estações. Nesse contexto, procedimentos automatizados de controle de qualidade passam a ser instrumentos importantes para identificar registros potencialmente inconsistentes, priorizar a análise técnica e orientar etapas posteriores de consistência. A literatura especializada destaca que procedimentos de controle de qualidade devem combinar verificações estatísticas, temporais, espaciais e climatológicas, evitando tanto a aceitação acrítica de valores duvidosos quanto a exclusão automática de observações raras, porém plausíveis (Durre et al., 2010; WMO, 2018).

O controle de qualidade de dados de precipitação apresenta desafios específicos, pois a chuva possui elevada variabilidade espacial e temporal, especialmente em regiões sujeitas a eventos convectivos. Assim, um valor discrepante em relação à média regional pode representar tanto uma inconsistência quanto um evento real localizado. De modo semelhante, registros nulos ou muito baixos em uma estação, quando estações vizinhas indicam chuva generalizada, podem sugerir falhas de observação ou problemas operacionais. Por essa razão, a identificação de suspeitas deve ser compreendida como uma etapa de triagem e priorização da revisão técnica, e não como procedimento automático de correção, exclusão ou validação definitiva dos dados.

Uma forma recorrente de avaliar a coerência de registros pluviométricos é comparar a estação em análise com estações de apoio vizinhas. A premissa é que estações próximas e com comportamento histórico semelhante tendem a apresentar padrões mensais compatíveis, ainda que não idênticos. A seleção dessas estações de apoio, contudo, é etapa crítica do processo. Critérios baseados apenas na distância podem incluir postos próximos, mas pouco correlacionados; por outro lado, critérios baseados apenas na correlação podem selecionar estações distantes e menos

representativas de eventos locais. Dessa forma, a combinação de critérios espaciais, estatísticos e de disponibilidade amostral tende a tornar a comparação regional mais robusta.

Além da comparação espacial, a análise de séries mensais deve considerar a sazonalidade da precipitação. Em muitas regiões brasileiras, a distribuição anual das chuvas apresenta marcada alternância entre períodos chuvosos e secos. Consequentemente, um mesmo valor mensal pode ser usual em determinado mês e atípico em outro. Procedimentos de controle de qualidade precisam, portanto, incorporar referências sazonais, de modo a reduzir interpretações equivocadas de valores baixos ou elevados quando avaliados fora do contexto climatológico mensal. A importância da avaliação de consistência em séries climáticas mensais também é destacada em estudos voltados à comparação de métodos aplicados a dados históricos (Venema et al., 2012).

Diante desse contexto, este trabalho apresenta uma metodologia automatizada de controle de qualidade aplicada a dados pluviométricos mensais, baseada na seleção híbrida de estações de apoio vizinhas. A abordagem proposta combina critérios de proximidade, correlação e disponibilidade de dados para estimar o comportamento regional esperado e avaliar desvios entre a precipitação observada na estação analisada e a referência construída a partir das estações de apoio. O procedimento inclui indicadores de suspeição, classificação por níveis de severidade e produtos gráficos de auditoria.

O objetivo principal do trabalho é avaliar a capacidade de uma rotina automatizada de Controle de Qualidade (*Quality Control* - QC) em identificar registros mensais de precipitação que merecem revisão técnica. Como objetivos específicos, busca-se: selecionar estações de apoio por critérios combinados; calcular referências regionais para comparação com a estação analisada; detectar padrões compatíveis com falso zero, ausência prolongada, valores extremos, anomalias espaciais e discrepâncias sazonais; e produzir saídas tabulares e gráficas que auxiliem a priorização da análise de consistência.

A contribuição do trabalho está na estruturação de uma etapa preliminar, automatizada e reprodutível de controle de qualidade para séries pluviométricas mensais. Os resultados gerados não substituem a análise técnica especializada, mas fornecem uma base objetiva para direcionar esforços de revisão, reduzir subjetividades e organizar o processo de auditoria dos dados. Essa etapa pode ser posteriormente integrada a procedimentos complementares de validação, incluindo dados diários (Hamada et al., 2011), telemetria, estimativas por satélite e métodos adicionais de avaliação temporal.

2 MATERIAIS E MÉTODOS

O trabalho foi desenvolvido com base em uma rotina automatizada de Controle de Qualidade (Quality Control - QC) aplicada a séries mensais de precipitação. A metodologia corresponde à primeira etapa de um pipeline de análise de dados de chuva mensal e contempla a seleção de estações de apoio vizinhas, a construção de uma referência regional e a aplicação de testes automáticos para identificação de registros potencialmente suspeitos. Neste estudo, são considerados apenas os resultados gerados por esse módulo inicial, sem incorporação das etapas posteriores de validação por satélite, consolidação, revisão conservadora de suspeitas e Curva de Dupla Massa (CDM).

2.1 Base de dados

Foram utilizados dados mensais de precipitação registrados em estações pluviométricas disponíveis no banco de dados Hidro da ANA. A rotina consulta os totais mensais de chuva, códigos das estações, datas, níveis de consistência e metadados básicos, incluindo nome, latitude e longitude. Para evitar duplicidade mensal, quando há mais de um registro para a mesma estação e mês, os dados são ordenados por estação, data e nível de consistência, mantendo-se apenas um registro por combinação estação-mês.

As datas são padronizadas para o primeiro dia do mês correspondente, de modo a uniformizar a unidade temporal da análise e permitir a correta aplicação dos testes baseados em sequência cronológica. A base analisada contempla 20.565 registros mensais, distribuídos em 53 estações pluviométricas da ANA localizadas nos estados do Ceará e Piauí, no período de janeiro de 1912 a dezembro de 2025. Cada estação foi avaliada individualmente, tendo como referência um conjunto de estações vizinhas selecionadas automaticamente.

2.2 Seleção das estações de apoio e construção da referência regional

Para cada estação em análise, denominada estação-alvo ou estação de análise, foi selecionado um conjunto de até cinco estações de apoio vizinhas. A seleção foi realizada por critério híbrido, combinando proximidade espacial, correlação histórica e disponibilidade de dados simultâneos. Inicialmente, foram consideradas candidatas todas as estações diferentes da estação-alvo. Para cada candidata, foram calculados a distância geográfica em relação à estação-alvo, a correlação de Pearson entre séries mensais coincidentes e o número de meses simultâneos.

Foram mantidas apenas as candidatas situadas a até 100 km da estação-alvo, com correlação igual ou superior a 0,70 e pelo menos 24 meses de dados em comum. As estações aprovadas nesses filtros receberam uma pontuação composta, formada pela combinação ponderada de distância, correlação e tamanho amostral, com pesos de 0,35, 0,45 e 0,20, respectivamente. As cinco estações com maior pontuação foram selecionadas como estações de apoio.

Após a seleção, foi construída uma série mensal de referência, denominada apoio combinado. Essa série foi calculada pela média ponderada dos totais mensais disponíveis nas estações de apoio, dada por $\hat{P}_t = \sum_{i=1}^{n_t} \omega_i P_{i,t} / \sum_{i=1}^{n_t} \omega_i$ com $\omega_i = 1/d_i^2$, em que $\hat{P}_t$ representa a precipitação estimada no mês t, $P_{i,t}$ é a precipitação da estação de apoio i, d_i é a distância entre a estação de apoio e a estação-alvo, ω_i é o peso atribuído à estação de apoio e n_t é o número de vizinhas com dado disponível no mês. Dessa forma, estações mais próximas exercem maior influência na estimativa regional. Para cada mês, o cálculo considera apenas as estações de apoio com dado disponível, sendo também registrados indicadores auxiliares, como número de vizinhas ativas, mediana das vizinhas, desvio-padrão entre vizinhas, percentual de vizinhas com chuva e suporte efetivo ponderado.

A rotina também considera a sazonalidade da precipitação, calculando estatísticas específicas para cada mês do ano. Entre os indicadores utilizados estão o Z-score sazonal da observação e envelopes empíricos de baixa e alta frequência, baseados nos quantis mensais da própria estação. Esse procedimento permite avaliar se determinado valor é atípico em relação ao comportamento esperado para o mês analisado.

2.3 Testes de controle de qualidade

Os testes de controle de qualidade foram organizados em camadas hierárquicas, segundo a natureza e a severidade da inconsistência identificada. A Camada 1 (C1) reúne ocorrências críticas; a Camada 2 (C2) reúne indícios importantes de inconsistência; e a Camada 3 (C3) reúne evidências complementares, úteis à interpretação, mas que isoladamente não caracterizam o registro como suspeito. Além dessas camadas, são gerados avisos operacionais relacionados ao suporte espacial, ao nível de consistência e à estabilidade da relação entre a estação-alvo e as estações de apoio.

Na Camada 1, são avaliados registros com valor negativo, falso zero severo, ausência prolongada em contexto regional úmido e sequência quase constante. Na Camada 2, são avaliados outliers de resíduo robusto, extremos raros não corroborados pelas vizinhas, pico isolado em campo seco, seco isolado em campo úmido e falso zero moderado. Na Camada 3, são avaliadas anomalias espaciais em relação à faixa de aceitabilidade, valores fora do envelope sazonal e desvios em relação ao campo local quando as estações vizinhas apresentam baixa dispersão.

O resíduo foi definido como a diferença entre a precipitação observada na estação-alvo e o apoio combinado, isto é, $R_t = P_t - \hat{P}_t$, em que R_t é o resíduo no mês t, P_t é a precipitação observada na estação-alvo e $\hat{P}_t$ é a precipitação estimada pelo apoio combinado. Para reduzir a influência de extremos, utiliza-se uma padronização robusta baseada na mediana e no desvio absoluto mediano. A faixa de aceitabilidade é obtida por modelo linear entre a estação-alvo e o apoio combinado, com

intervalo de predição de 99%. Esses critérios permitem avaliar tanto discrepâncias absolutas quanto desvios em relação ao comportamento histórico e regional esperado.

2.4 Pontuação, classificação e produtos gerados

Cada registro mensal recebe uma pontuação de suspeição, denominada *Score*, definida conforme os testes acionados. Registros sem indícios relevantes recebem Score igual a zero. Registros que acionam testes das camadas principais são classificados como suspeitos e recebem a descrição dos motivos correspondentes. Para facilitar a interpretação, a rotina registra separadamente os motivos associados às camadas C1, C2 e C3, além dos avisos operacionais (AV). Também é atribuída uma camada dominante, definida de forma hierárquica: C1, C2, C3 e aviso (AV).

A classificação de um registro como suspeito não implica, necessariamente, que o valor esteja errado. Considerando a elevada variabilidade espacial da precipitação, especialmente em eventos localizados, os resultados devem ser entendidos como triagem preliminar para orientar a auditoria e a revisão técnica, e não como procedimento automático de exclusão ou correção dos dados.

Como produtos, a rotina gera uma base consolidada com os registros analisados, valores observados, apoio combinado, indicadores de suporte, resíduos, Score, classificação de suspeição, camada dominante e motivos da sinalização. Também são produzidos relatórios gráficos por estação, incluindo mapa de localização, correlações com estações de apoio, disponibilidade temporal, regime sazonal, dispersão entre estação-alvo e apoio combinado, resíduos, totais anuais, curva de permanência e gráfico quantil-quantil. Para eventos de maior prioridade, são gerados mapas espaciais de precipitação, resíduos e anomalia percentual, utilizados como apoio visual à interpretação dos registros sinalizados.

3 RESULTADOS E DISCUSSÃO

A aplicação da rotina automatizada de controle de qualidade aos dados pluviométricos mensais resultou em uma base consolidada com 20.565 registros, distribuídos em 53 estações, abrangendo o período de janeiro de 1912 a dezembro de 2025. Para cada registro mensal, foram calculados indicadores de comparação entre a precipitação observada na estação-alvo e a estimativa regional obtida a partir das estações de apoio vizinhas, além de métricas auxiliares relacionadas ao suporte espacial, sazonalidade, resíduo robusto, faixa de aceitabilidade e classificação da suspeição.

Do total analisado, 2.926 registros foram classificados como suspeitos, correspondendo a 14,2% da base. Desses, 633 registros foram enquadrados na Camada 1, associada a ocorrências críticas, e 2.293 registros na Camada 2, relacionada a indícios importantes de inconsistência. Outros registros apresentaram apenas evidências complementares ou avisos operacionais, os quais não caracterizam, isoladamente, erro no dado, mas auxiliam na interpretação da robustez da comparação

realizada. Também foram identificados 1.133 registros com Score igual ou superior a 3, correspondendo a aproximadamente 5,5% da base, os quais representam o subconjunto de maior prioridade para auditoria técnica.

Entre os principais padrões de sinalização observados, destacaram-se outliers de resíduo robusto, valores fora do envelope sazonal, anomalias espaciais em relação à faixa de aceitabilidade, seco isolado em campo úmido, pico isolado em campo seco e ausência prolongada de chuva em contexto regional úmido. Esses resultados são compatíveis com inconsistências esperadas em séries pluviométricas mensais, tais como falhas de observação, problemas de transcrição, valores acumulados indevidamente, registros nulos em meses chuvosos ou eventos locais não corroborados pelas estações vizinhas. Contudo, considerando a elevada variabilidade espacial da precipitação, a classificação como suspeita deve ser interpretada como indicação de necessidade de revisão, e não como confirmação automática de erro.

Para exemplificar os produtos gerados pela rotina, foi selecionada a estação 00340118[2], denominada CURRAL VELHO. Essa estação apresentou 153 registros mensais no período de abril de 2013 a dezembro de 2025. Foram identificados 9 registros suspeitos, correspondendo a 5,9% da série analisada, todos classificados na Camada 2. Também foram observados 21 registros com evidências complementares de Camada 3, sem classificação final como suspeitos. A estação apresentou aviso de baixa consistência em todos os meses analisados, em razão do nível de consistência NC = 1 (dados brutos), aspecto que reforça a necessidade de revisão técnica posterior.

A Figura 1 apresenta o mosaico de auditoria gerado para a estação 00340118. O produto reúne, em uma única visualização, mapa de localização da estação-alvo e das estações de apoio, correlações, disponibilidade temporal, regime sazonal, curvas de dupla massa, dispersões entre estação e apoio, análise de resíduos, totais anuais, curva de permanência e gráfico quantil-quantil. A integração desses painéis permite avaliar se as suspeitas decorrem de comportamento pontual, diferenças sistemáticas entre séries, limitação do suporte espacial ou alteração temporal da relação entre a estação-alvo e as estações de apoio.

Além do mosaico geral, foram produzidos gráficos específicos de controle de qualidade, apresentados na Figura 2. Esses gráficos detalham a relação entre a precipitação observada e o apoio combinado, a faixa de aceitabilidade e o comportamento sazonal da observação. No caso da estação 00340118, parte das sinalizações esteve associada a outliers de resíduo robusto e a valores fora do

[2] O código oficial da estação pluviométrica é composto por oito dígitos, conforme a codificação adotada no cadastro de estações hidrometeorológicas gerenciado pela ANA. Na ferramenta automatizada utilizada neste trabalho, os zeros à esquerda foram omitidos nos produtos gráficos; assim, o código 340118 apresentado nas figuras corresponde à estação 00340118 – CURRAL VELHO.

envelope sazonal, especialmente em meses nos quais a precipitação observada diferiu do comportamento esperado com base nas estações de apoio.

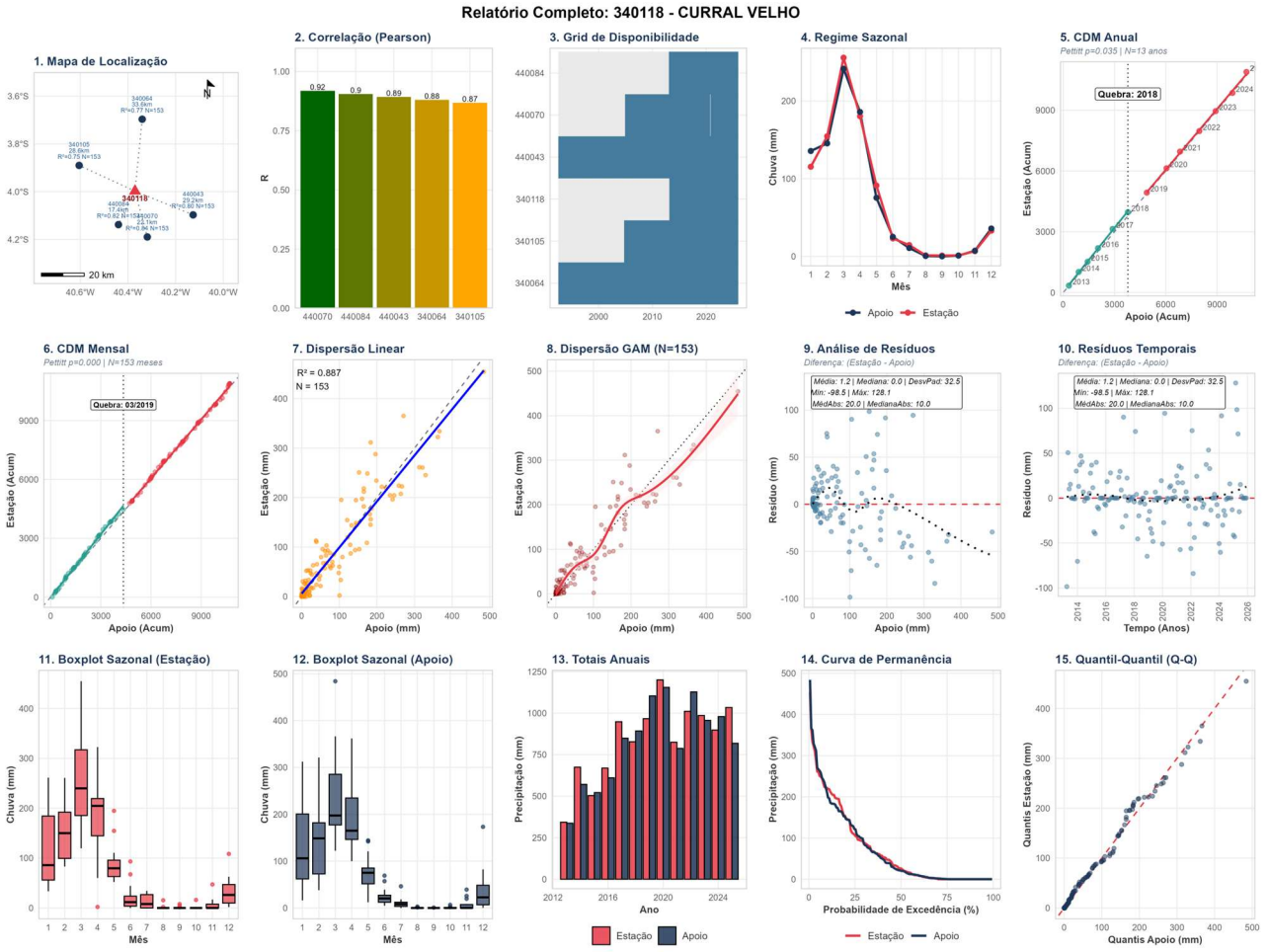


Figura 1 - Mosaico de auditoria gerado para a estação 00340118 - CURRAL VELHO.

Um dos eventos destacados ocorreu em outubro de 2024, quando a estação 00340118 registrou 16,1 mm, enquanto o apoio combinado estimou aproximadamente 0,3 mm, com as cinco estações de apoio disponíveis e sem chuva relevante nas vizinhas. A Figura 3 apresenta os mapas espaciais de precipitação, anomalia percentual e resíduos para esse evento. Os mapas evidenciam que a precipitação registrada na estação-alvo se destacou em relação ao entorno, caracterizando um possível evento localizado ou uma inconsistência a ser verificada. A distinção entre essas possibilidades exige análise complementar, podendo envolver consulta a dados diários, boletins pluviométricos, registros operacionais, informações de campo, telemetria ou estimativas independentes de precipitação.

Os resultados demonstram que a metodologia proposta é adequada para organizar a triagem inicial de inconsistências ou de dados suspeitos em séries mensais de precipitação. A combinação de seleção híbrida de estações de apoio, comparação espacial, tratamento sazonal e classificação hierárquica permite diferenciar suspeitas críticas, indícios importantes, evidências complementares e

avisos operacionais. Essa estrutura reduz interpretações binárias simplificadas e fornece uma base objetiva para priorização da análise técnica.

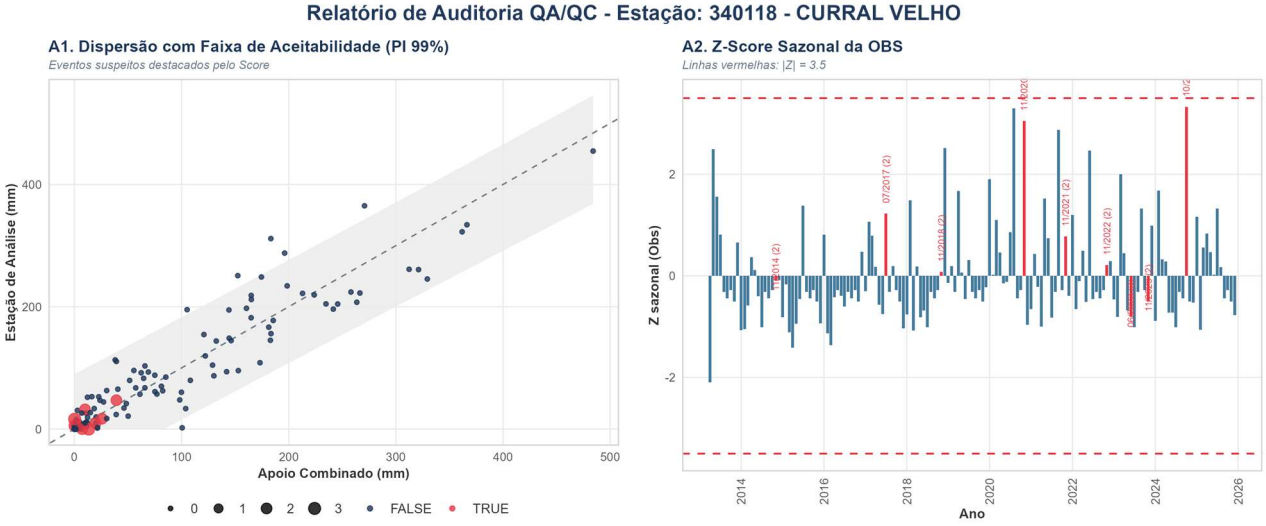


Figura 2 - Gráficos de controle de qualidade gerados para a estação 00340118 - CURRAL VELHO.

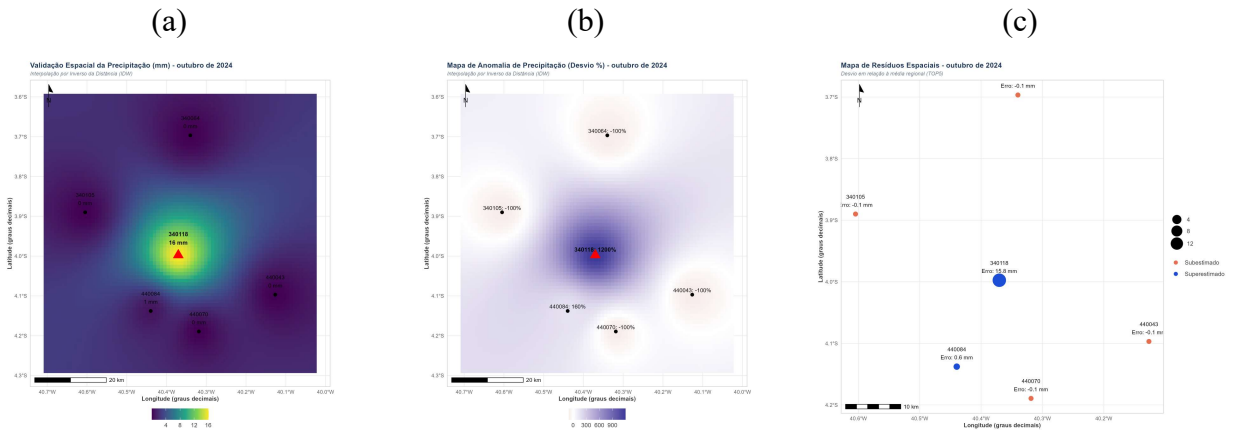


Figura 3 - Mapas espaciais de precipitação (a), anomalia percentual (b) e resíduos (c) para o evento de outubro de 2024 na estação 00340118 - CURRAL VELHO.

A aplicação à estação CURRAL VELHO evidencia, ainda, a importância da integração entre produtos tabulares e gráficos. Enquanto os resultados tabulares indicam os meses prioritários para revisão, os mosaicos, gráficos e mapas ajudam a compreender a natureza da sinalização. Dessa forma, o módulo avaliado constitui uma etapa preliminar útil para auditoria e priorização da consistência hidrológica, podendo ser posteriormente integrado a fontes independentes e métodos complementares de validação.

4 CONCLUSÕES

Este trabalho apresentou uma metodologia automatizada de Controle de Qualidade (Quality Control - QC) aplicada a séries mensais de precipitação, baseada na seleção híbrida de estações de

apoio vizinhas. A abordagem combinou critérios espaciais, estatísticos e amostrais para construir uma referência regional de comparação e identificar registros potencialmente suspeitos.

A aplicação da rotina demonstrou sua utilidade para organizar a triagem inicial de inconsistências em bases pluviométricas mensais extensas. A classificação hierárquica adotada permitiu diferenciar ocorrências críticas, indícios importantes, evidências complementares e avisos operacionais, evitando a interpretação automática de todo valor atípico como erro. Assim, os registros sinalizados devem ser entendidos como prioridades para revisão técnica, e não como indicações definitivas de exclusão ou correção.

Os resultados indicaram 2.926 registros suspeitos, correspondentes a 14,2% da base analisada, dos quais 1.133 apresentaram Score igual ou superior a 3. Esse subconjunto representa os casos de maior prioridade para auditoria, demonstrando o potencial da rotina para direcionar esforços de consistência e reduzir o volume de registros a serem avaliados manualmente.

O estudo de caso da estação 00340118 - CURRAL VELHO evidenciou a relevância dos produtos gráficos gerados pelo módulo, incluindo mosaicos de auditoria, gráficos de controle de qualidade e mapas espaciais. Esses produtos ampliam a transparência da análise e auxiliam a interpretação dos motivos de sinalização, especialmente em situações nas quais é necessário distinguir eventos localizados de possíveis inconsistências.

Conclui-se que o módulo avaliado constitui uma etapa preliminar robusta, reprodutível e operacionalmente útil para o controle de qualidade de dados pluviométricos mensais. Como continuidade, recomenda-se integrar os registros sinalizados a fontes independentes e métodos complementares de validação, incluindo dados diários, telemetria, estimativas por satélite, análise de consistência temporal e Curva de Dupla Massa.

5 REFERÊNCIAS

DURRE, I.; MENNE, M. J.; GLEASON, B. E.; HOUSTON, T. G.; VOSE, R. S. (2010). “Comprehensive automated quality assurance of daily surface observations”. Journal of Applied Meteorology and Climatology 49(8), pp. 1615-1633.

HAMADA, A.; ARIMOTO, Y.; YATOMI, H.; YAMADA, T.; TAKARA, K. (2011). “An automated quality control method for daily rain-gauge data”. Hydrological Research Letters 5, pp. 2-12.

VENEMA, V. K. C.; MESTRE, O.; AGUILAR, E.; AUER, I.; GUIJARRO, J. A.; DOMONKOS, P.; VERTACNIK, G.; SLEZAK, P.; STEPANEK, P.; ZAHRADEK, P.; VIARRE, J.; MÜLLER-WESTERMEIER, G.; LAKATOS, M.; WILLIAMS, C. N.; MENNE, M. J.; LINDBERG, C.; RASO, J.; RUSTEMEIER, E.; KOLLER, S.; MALINEN, M.; TARDIVO, G.; BEGERT, M.; BRUNET, M.; PETERSON, T. C. (2012). “Benchmarking homogenization algorithms for monthly data”. Climate of the Past 8, pp. 89-115.

WMO - WORLD METEOROLOGICAL ORGANIZATION (2018). Guide to Climatological Practices. WMO-No. 100. Geneva.