XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

CONTROLE DE QUALIDADE AUTOMÁTICO PARA PRECIPITAÇÃO SUBDIÁRIA: UMA ANÁLISE DOS DADOS DO RIO GRANDE DO NORTE

José Lindemberg Vidal-Barbosa [1] ; Gabriela da Luz Lins [2] ; Filipe Carvalho Lemos[3] ; Eduardo Gonçalves Patriota [4], Cinthia Maria de Abreu Claudino[5] ; Victor Hugo Rabelo Coelho[6], Guillaume Francis Bertrand[7]; Emerson da Silva Freitas[8]; Cristiano das Neves Almeida[9]

Abstract: Subdaily precipitation data are crucial for understanding extreme events and climate patterns in high-resolution. However, equipment failures, transmission issues, or storage inconsistencies can compromise data quality. This study proposes an automated quality control methodology for a subdaily precipitation dataset from EMPARN (Agricultural Research Corporation of Rio Grande do Norte). The methodology employs three approaches: subdaily and daily data analysis, and a hybrid model. The subdaily approach focuses on identifying repetitive patterns and anomalies specific to high-frequency data, while the daily analysis applies consistency checks. The hybrid model combines both methods to enhance data reliability. Validation was carried out using a reference dataset from CEMADEN, demonstrating the expected effectiveness. Results show that the hybrid model achieved the highest accuracy (86%) and precision (62%). Despite these results, challenges remain, such as minimizing false positives and improving low-quality data discharge.

Resumo: Os dados de precipitação subdiária são cruciais para compreender eventos extremos e padrões climáticos em alta resolução temporal. No entanto, falhas em equipamentos, problemas de transmissão ou inconsistências no armazenamento podem comprometer a qualidade dos dados. Este estudo propõe uma metodologia automatizada de controle de qualidade para o conjunto de dados de precipitação subdiária da EMPARN (Empresa de Pesquisa Agropecuária do Rio Grande do Norte). A metodologia utiliza três abordagens distintas: análise de dados subdiários e diários, e um modelo híbrido. A abordagem subdiária foca na identificação de padrões repetitivos e anomalias específicas de dados de alta frequência, enquanto a análise diária aplica verificações de consistência. O modelo híbrido combina essas metodologias para melhorar a confiabilidade. A validação foi realizada utilizando um conjunto de dados de referência do CEMADEN, demonstrando a eficácia da metodologia proposta. Os resultados mostraram que o modelo híbrido alcançou a maior acurácia (86%) e precisão (62%). Apesar desses resultados, desafios como a minimização de falsos positivos e a melhoria na deleção de dados de baixa qualidade.

Palavras-Chave – controle de qualidade; precipitação subdiária; monitoramento pluviométrico.

1) Doutorando em Engenharia Civil e Ambiental, UFPB, Campus João Pessoa, (85) 99972-9552, lindembergvidal@gmail.com
2) Doutoranda em Engenharia Civil e Ambiental, UFPB, Campus João Pessoa, (83) 99176-9150, gabyylins17@gmail.com
3) Doutorando em Engenharia Civil e Ambiental, UFPB, Campus João Pessoa, (83) 99964-3487, filipe_carvalho_l@hotmail.com
4) Doutorando em Engenharia Civil e Ambiental, UFPB, Campus João Pessoa, (83) 99655-3877, edugoncalvespatriota@gmail.com
5) Doutoranda em Engenharia Civil e Ambiental, UFPB, Campus João Pessoa, (83) 99618-9626, cinthiamariaac@gmail.com
6) Prof. Dr. do Departamento de Geociências, UFPB, Campus João Pessoa, (83) 98886-1663, victor.coelho@academico.ufpb.br
7) Prof. Dr. do Departamento de Engenharia Civil e Ambiental, UFPB, Campus João Pessoa; Chrono-environnement (UMR 6249), Université Marie et Louis Pasteur, +33 6 46 65 78 37, guillaume353@gmail.com
8) Prof. Dr. do Instituto Federal da Paraíba, Campus Picuí, (83) 99405-9283, emerson.freitas@ifpb.edu.br
9) Prof. Dr. do Departamento de Engenharia Civil e Ambiental, UFPB, Campus João Pessoa, (83) 99349-0880, almeida74br@yahoo.com.br

INTRODUÇÃO

Dados de precipitação são essenciais para estudos e obras necessárias para o desenvolvimento das localidades. Obras de drenagem urbana e tantas outras despertam a necessidade de utilização de dados precisos, pois estes evidenciam de fato o que acontece naquela localidade. Além do mais, permitem compreender, também, o padrão e influência das mudanças climáticas na região. Comumente, devido maior disponibilidade, são utilizados dados diários, mas uma alternativa para a coleta de dados trata-se de estações automáticas e com registros de dados subdiários, com passo de tempo inferiores a 24 horas, que podem inferir uma maior precisão em aplicações, tais como modelos hidrológicos, porém tem-se a necessidade de um controle de qualidade mais rigoroso para remoção de inconsistências.

A coleta, a transmissão e o armazenamento dos dados de precipitação estão sujeitos a erros, despertando a demanda para um controle da qualidade dos dados obtidos, tais como métodos e algoritmos para automatizar o procedimento, uma vez que grandes bases de dados geram um volume impossível de ser analisado manualmente. A abordagem do procedimento de automatização de um controle pode se basear em diferentes aspectos de seu registro, consistindo em: (i) comparação com dados auxiliares de fontes externas, como imagens de radar, de satélite etc.; (ii) definição de limites para excluir valores improváveis; ou (iii) teste de consistência interna comparando estações vizinhas ou com informação do passado já validada de uma mesma estação.

Globalmente, iniciativas como o *Global Sub-Daily Rainfall* (GSDR) (Lewis et al., 2021) e o algoritmo GSDR-QC (25 verificações) avançaram na validação de dados subdiários. No Brasil, destacam-se o trabalho de Meira et al. (2022) com dados do CEMADEN (sub-horários) e as grades diárias de Xavier et al. (2016) e Xavier et al. (2022). Na Catalunha (nordeste da Espanha), Llabrés-Brustenga et al. (2019) desenvolveu um procedimento de controle de qualidade para precipitação diária contando com mais de 1.700 estações, cuja metodologia se dividia em três etapas: (i) um controle básico de qualidade para a detecção de valores fisicamente impossíveis; (ii) um controle de qualidade absoluto, que testa características internas de cada estação, tais como disponibilidade de dados, lacunas, valores atípicos e sazonalidade semanal; e (iii) um controle de qualidade relativo, que avalia a qualidade de cada valor diário de chuva baseado na similaridade com os valores das estações vizinhas. Estévez et al. (2022) reuniram dados de 1.947 estações de uma região semiárida na Espanha com um período entre 1870 até os dias atuais, aplicando uma metodologia muito similar a empregada por Llabrés-Brustenga et al. (2019). Todavia, foi feita uma adaptação ao clima da região, que é mais heterogêneo, predominantemente semiárido e possui grandes períodos de seca, semelhante ao semiárido brasileiro (SAB).

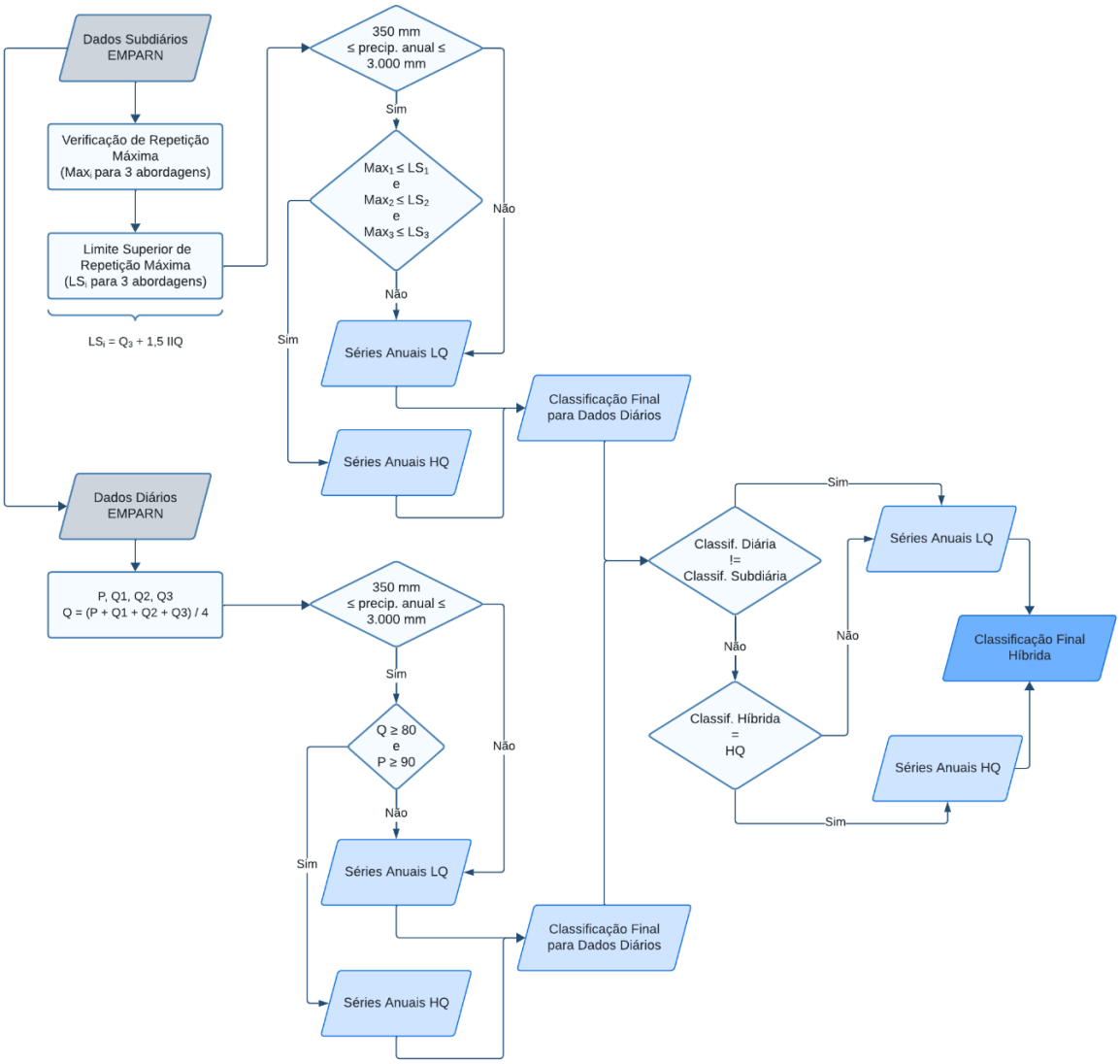
Contudo, lacunas persistem em metodologias para dados subdiários, especialmente em regiões semiáridas como o Nordeste Brasileiro (NEB), onde as características climáticas locais exigem abordagens específicas. O Rio Grande do Norte (RN), com 85% de seu território no semiárido, possui rede de monitoramento subdiário da EMPARN desde 2018, porém sem procedimentos automatizados de controle de qualidade.

Neste panorama, este artigo pretende propor uma metodologia de controle de dados automatizado para a base de dados subdiários de precipitação da EMPARN utilizando técnicas em diferentes resoluções temporais, que dispense uma base de dados suplementar, cujo produto é apresentar a primeira avaliação automática de qualidade para esse conjunto.

METODOLOGIA

Uma série de procedimentos foram usados para realizar o controle de qualidade automático (*Automatic Quality Control Procedure*, A-QCP) da base de dados subdiários da EMPARN, baseado em comportamentos e padrões dos dados pluviométricos, sem a necessidade de uma base verificada previamente. A figura 1 a seguir ilustra as etapas do processo:

Figura 1 – Fluxograma simplificado do controle de qualidade



Área de Estudo

O Estado do Rio Grande do Norte possui 52.809,599 km² de superfície, com população de 3.302.729 habitantes, de acordo com o Censo Demográfico de 2022 (IBGE, 2022), constituindo o 17º estado mais populoso do Brasil. A população se concentra majoritariamente em torno da Região Metropolitana de Natal (RMN), que apresenta mais de 1,6 milhão de pessoas. Outros núcleos populacionais importantes incluem as cidades de Mossoró e Caicó.

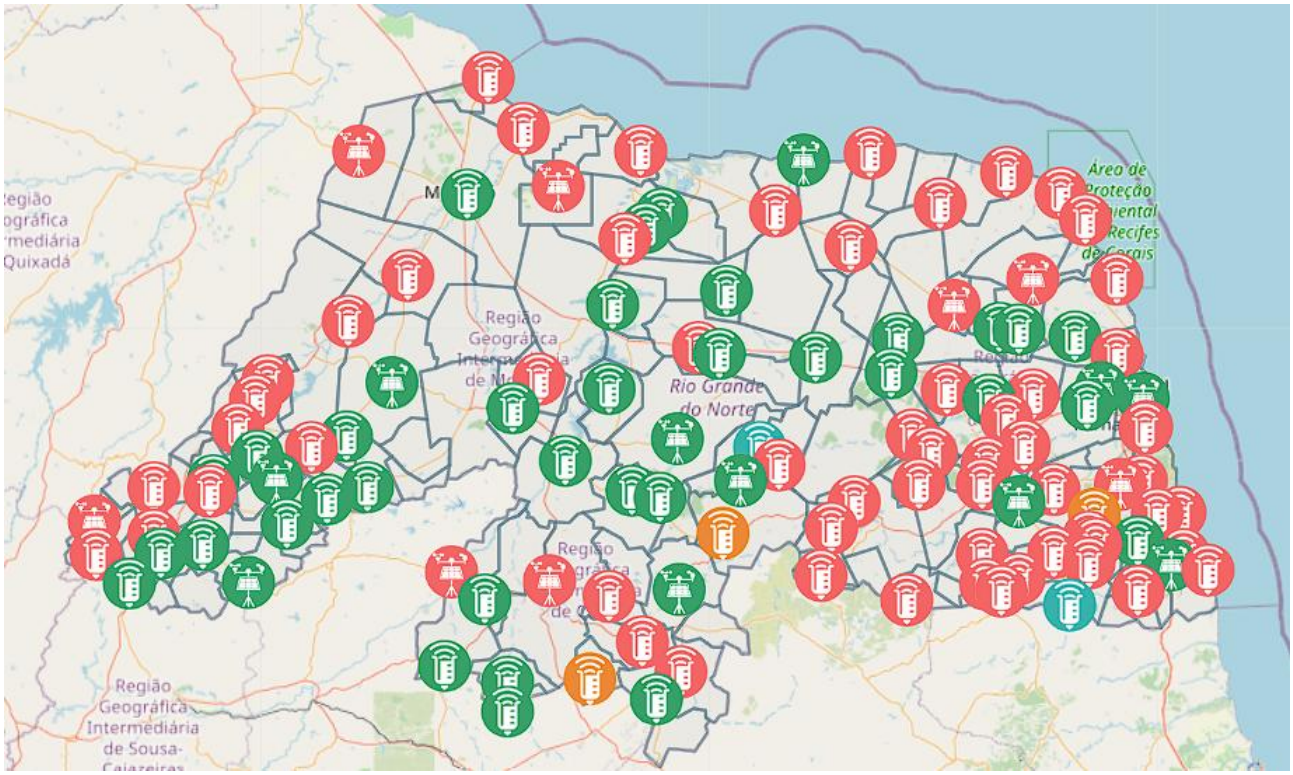
Cerca de 85% dos seus 167 municípios estão no semiárido, com a Caatinga como bioma predominante. O território potiguar é caracterizado por formações como a Chapada do Apodi, Serra de Santana e o Vale do Açu. Os climas característicos são o semiárido quente (BSh) e o tropical de savana (As), com algumas áreas recebendo uma precipitação anual de apenas 400 mm (Alvares et al.,

2014). Apesar da pequena variação térmica anual, a distribuição da precipitação é bastante irregular ao longo do ano. A altitude nas regiões de serra e a continentalidade no interior do estado provocam grandes variações na umidade. De modo geral, os índices pluviométricos anuais são relativamente baixos (com média menor do que 800 mm anuais), com exceção da região serrana e do litoral (Lucena; Cabral Júnior; Steinke, 2018).

Dados Subdiários de Precipitação (EMPARN)

A presente base de dados reúne medições de precipitação horária obtidas a partir de 120 pluviômetros automáticos (Figura 2) operados pela Empresa de Pesquisa Agropecuária do Rio Grande do Norte (EMPARN) no período de 2018 a 2023, fornecendo cerca de 481 séries anuais distintas. Os dados de precipitação da sua rede de monitoramento são gerados por estações do tipo PCDs (Plataformas de Coleta de Dados) e TELEPLUs (Telemetria de Pluviômetros) de hora em hora.

Figura 2 – Distribuição espacial da rede de monitoramento da EMPARN (EMPARN, 2024).



Dados Subdiários de Precipitação de Referência (CEMADEN)

A base de referência para inspeção visual (REF-VI) foi desenvolvida a partir de dados subdiários do CEMADEN, contendo registros em intervalos de 10 minutos durante eventos de precipitação e medições horárias em períodos secos. Inicialmente, aplicou-se um filtro para excluir pluviômetros com mais de 60 dias de falhas anuais, classificados automaticamente como de baixa qualidade (LQ). Os dados remanescentes foram submetidos a uma análise visual detalhada, onde as séries temporais de cada estação foram comparadas com dados mensais e sub-horários de suas cinco estações mais próximas, utilizando um protocolo de dupla avaliação independente com arbitragem por terceiro especialista em caso de discordância. As estações foram então classificadas como de alta qualidade (HQ) ou de baixa qualidade (LQ), sendo esta última identificada por três padrões principais:

(i) períodos extensos de valores zero durante eventos de chuva registrados nas estações vizinhas; (ii) ocorrência de picos pluviométricos superiores a 40 mm em intervalos de 10 minutos, inconsistentes com o histórico da estação ou da região; (iii) e registros persistentes de valores mínimos (0,2 mm) por longos períodos, sugerindo possíveis obstruções no equipamento, enquanto as estações adjacentes apresentavam variações típicas de precipitação. Esta abordagem metodológica rigorosa permitiu estabelecer uma base de referência confiável para a validação dos algoritmos automáticos de controle de qualidade (QCP), garantindo a identificação precisa de estações com problemas de qualidade sem comprometer a integridade dos dados classificados como confiáveis.

Controle de Qualidade

O controle de qualidade dos dados de precipitação foi estruturado em três abordagens complementares: análise subdiária, avaliação diária e um modelo híbrido integrado. A metodologia fundamentou-se em algoritmos capazes de identificar inconsistências nas séries temporais, como valores atípicos, repetições sistemáticas, lacunas de dados e registros inconsistentes, seguindo referências consolidadas na literatura. Uma etapa preliminar de filtragem eliminou automaticamente estações com séries anuais com precipitação acumulada inferior a 350 mm ou superior a 3.000 mm, classificando-as como Baixa Qualidade (LQ) devido à impropriedade física desses valores extremos para a região.

Para os dados subdiários, desenvolveu-se um método específico para detectar padrões de mau funcionamento em pluviômetros automáticos, particularmente aqueles relacionados a entupimentos. O processo considerou a precisão instrumental típica de 0,2 mm/h (com exceção de quatro estações em 2022, que operavam com 0,1 mm/h), aplicando três critérios sequenciais de avaliação. O primeiro analisou ocorrências consecutivas do valor mínimo exato, o segundo examinou registros não nulos iguais ou inferiores ao limite mínimo, e o terceiro incorporou uma dimensão temporal, avaliando intervalos entre 60 e 360 minutos. Os limiares de aceitação foram definidos pelo método do Intervalo Interquartil, classificando como HQ apenas as séries que atenderam aos critérios estabelecidos.

A avaliação dos dados diários adaptou metodologias validada por Estévez et al. (2022), calculando um Índice de Qualidade (Q) composto por quatro parâmetros interdependentes. A disponibilidade de dados (P) mensurou a completude das séries, enquanto o parâmetro de lacunas (Q_1) penalizou falhas prolongadas. A uniformidade na distribuição semanal de chuvas (Q_2) detectou possíveis vícios operacionais, e a análise de valores atípicos (Q_3) identificou precipitações mensais inconsistentes. Séries com $Q \geq 80$ e $P \geq 90$ receberam classificação HQ, garantindo um padrão mínimo.

O modelo híbrido surgiu como etapa final de integração, cruzando os resultados das análises subdiária e diária. Adotou-se um critério conservador, onde apenas as estações classificadas como HQ nos dois métodos mantiveram essa qualificação. Casos de discordância foram automaticamente reclassificados como LQ, priorizando a detecção de possíveis inconsistências. O processo de validação empregou matrizes de confusão e métricas estatísticas (precisão, acurácia e recall) para quantificar o desempenho do sistema, assegurando que os critérios adotados equilibravam sensibilidade e especificidade na identificação de dados problemáticos. Essa estrutura metodológica multicamadas permitiu uma avaliação abrangente da qualidade dos dados, incorporando diferentes escalas temporais e perspectivas analíticas para garantir a confiabilidade do conjunto final.

O processo de validação dos três métodos de controle de qualidade foi conduzido utilizando uma matriz de confusão, que permitiu avaliar a precisão das classificações das estações como HQ ou LQ. Os resultados foram categorizados como verdadeiros positivos (VP), representando as estações HQ corretamente classificadas; verdadeiros negativos (VN), referentes às estações LQ corretamente identificadas; falsos positivos (FP), ou erros do tipo I, em que estações LQ foram incorretamente

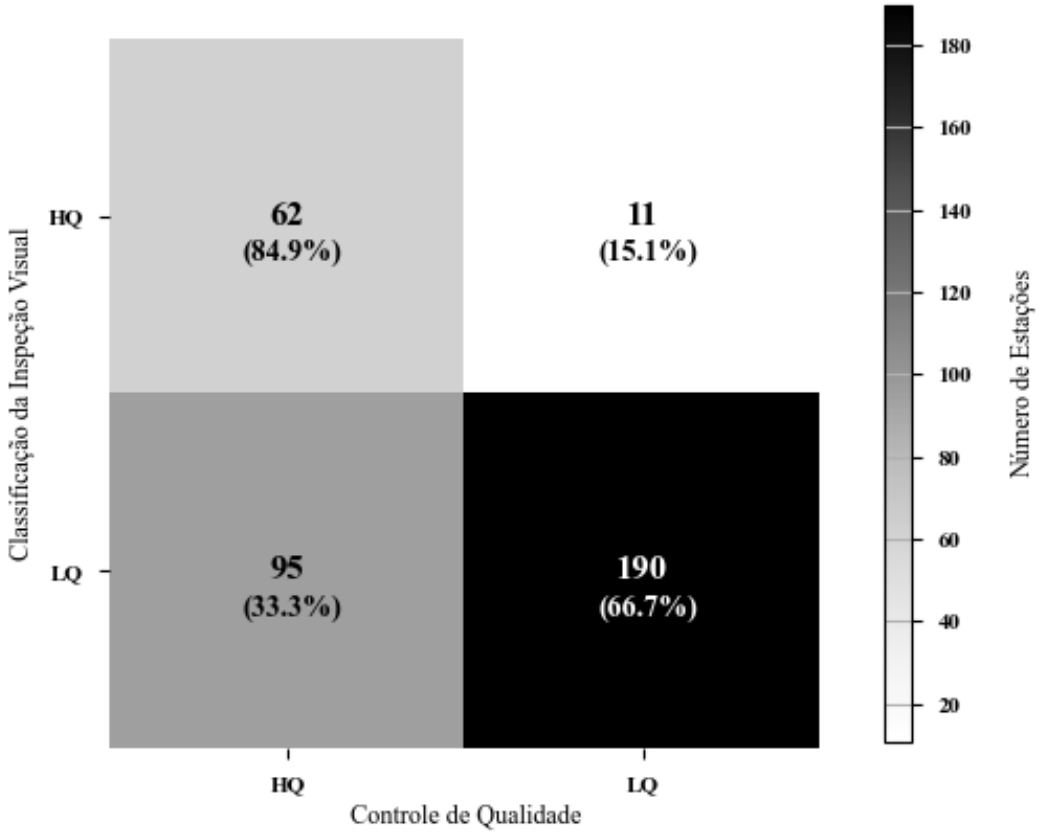
classificadas como HQ; e falsos negativos (FN), ou erros do tipo II, em que estações HQ foram equivocadamente classificadas como LQ. Além da matriz de confusão, foram calculadas métricas adicionais para avaliar o desempenho do método: a precisão, que mede a proporção de falsos positivos, indicando a confiabilidade na classificação das estações HQ; a acurácia, que avalia a eficácia geral do modelo na classificação correta de HQ e LQ; e o recall, ou revocação ou sensibilidade, que verifica a proporção de falsos negativos, refletindo o rigor do método em identificar as estações de baixa qualidade.

RESULTADOS E DISCUSSÕES

O presente estudo avaliou três abordagens distintas de controle de qualidade para séries temporais de precipitação no Rio Grande do Norte, utilizando dados do período de 2018 a 2023. A análise compreendeu 358 séries anuais provenientes de 70 estações do CEMADEN, que serviram como referência para validação, sendo 20,4% classificadas como alta qualidade (HQ) e 79,6% como baixa qualidade (LQ) através de inspeção visual criteriosa.

Para os dados subdiários (Figura 3), observou-se uma classificação de 49,7% das 481 séries da EMPARN como HQ, proporção significativamente superior à encontrada na base de referência. A validação contra os dados do CEMADEN revelou uma acurácia geral de 0,70, com revocação de 0,85 para a classe HQ, indicando boa capacidade de identificação de séries de qualidade. Contudo, a precisão de apenas 0,40 e a taxa de 33,3% de falsos positivos sugerem limitações na discriminação de séries LQ, apontando para a necessidade de refinamento nos critérios de classificação.

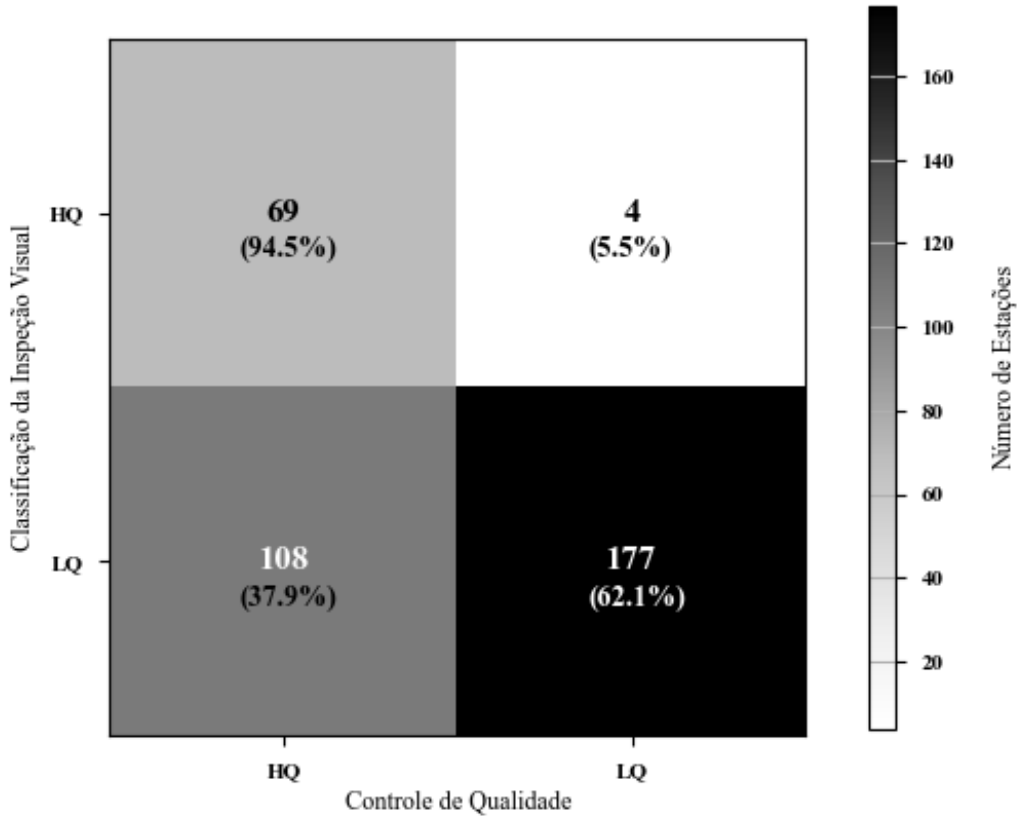
Figura 3 – Matriz de Confusão para Controle de Qualidade Subdiário (CEMADEN)



No controle de qualidade subdiário da EMPARN foram classificadas 239 séries anuais como HQ (49,7%) e 242 estações como LQ (50,3%) das 481 séries anuais disponíveis, ou seja, com a proporção de HQ muito superior a encontrada na base de validação.

A abordagem diária (Figura 4) demonstrou desempenho distinto, com revocação excepcionalmente alta (0,94) para séries HQ, porém acompanhada de significativa taxa de falsos positivos (37,9%). A acurácia de 0,68 foi similar à obtida no método subdiário, mas a baixa precisão (0,39) indica que, embora eficaz na detecção de séries boas, o modelo apresenta dificuldades em rejeitar adequadamente séries problemáticas. Este padrão sugere que os parâmetros adotados podem ser excessivamente permissivos para a realidade local.

Figura 4 – Matriz de Confusão para Controle de Qualidade Diário (CEMADEN)



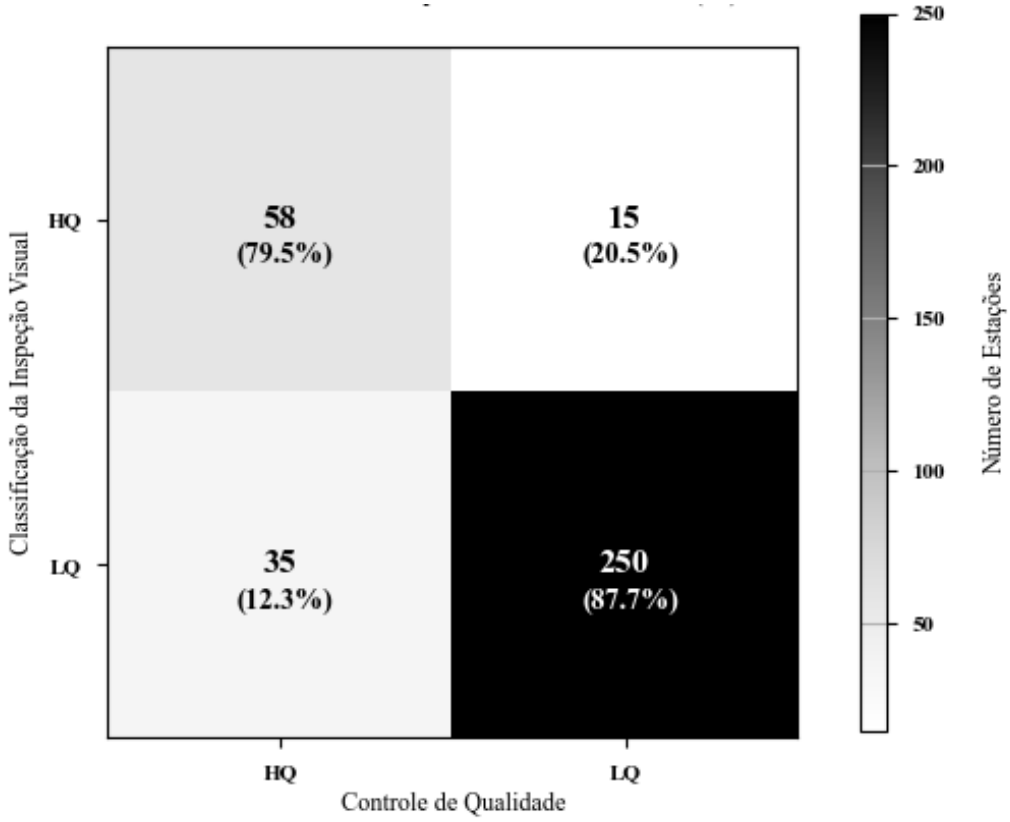
No controle de qualidade diário da EMPARN foram classificadas 189 séries anuais como HQ (39,3%) e 292 estações como LQ (60,7%) das 481 séries anuais disponíveis, ou seja, com a proporção de HQ também muito superior a encontrada na base de validação.

O modelo híbrido (Figura 5), integrando as análises subdiária e diária, apresentou o melhor desempenho global, com acurácia de 0,86 e precisão de 0,62 - valores substancialmente superiores aos das abordagens individuais. A taxa de falsos positivos reduziu-se para 12,3%, enquanto a revocação manteve-se em nível satisfatório (0,79). Esses resultados indicam que a combinação de múltiplas perspectivas temporais permite um equilíbrio mais adequado entre sensibilidade e especificidade na classificação de qualidade.

No controle de qualidade híbrido da EMPARN foram classificadas 175 séries anuais como HQ (36,4%) e 306 estações como LQ (63,6%) das 481 séries anuais disponíveis, ou seja, com a proporção

de HQ também superior a encontrada na base de validação, porém menor do que nos modelos anteriores.

Figura 5 – Matriz de Confusão para Modelo Híbrido (CEMADEN)



A comparação entre as abordagens revela um *trade-off* característico entre revocação e precisão. Enquanto o método diário mostrou excelente sensibilidade para HQ (94,5%), o híbrido destacou-se pela capacidade de minimizar falsos positivos sem comprometer excessivamente a identificação de séries boas. Esta melhoria no equilíbrio entre métricas sugere que a integração de diferentes escalas temporais pode superar limitações inerentes a abordagens unimodais.

CONCLUSÃO

O estudo avaliou três abordagens de controle de qualidade para séries meteorológicas da EMPARN: subdiária, diária e híbrida. O modelo híbrido destacou-se com maior acurácia (0,86) e precisão (0,62), demonstrando eficácia na classificação de séries de HQ e LQ. Embora sua revocação (0,79) fosse inferior às demais, manteve um bom desempenho na identificação de HQ.

O método subdiário apresentou limitações, com baixa precisão (0,40) e alta taxa de falsos positivos, indicando necessidade de refinamento na discriminação entre classes. Já a análise diária obteve excelente revocação (0,95) para HQ, mas revelou uma taxa significativa de falsos positivos (37,9%), apontando para desafios na classificação de LQ.

Em síntese, embora o modelo híbrido tenha se mostrado superior, todas as abordagens exigem melhorias, como redução de falsos positivos e aprimoramento na detecção de LQ. Futuros estudos podem explorar técnicas de aprendizado de máquina e bases de dados mais diversificadas para aumentar a confiabilidade das classificações. Este trabalho não apenas avança nas metodologias de controle de qualidade, mas também orienta a integração de modelos híbridos em sistemas de monitoramento, visando maior precisão na análise de séries meteorológicas em larga escala.

REFERÊNCIAS

EMPARN – EMPRESA DE PESQUISA AGROPECUÁRIA DO RIO GRANDE DO NORTE. "*RELATÓRIO DE VARIÁVEIS METEOROLÓGICAS*". Disponível em: <https://meteorologia.emparn.rn.gov.br/relatorios/relatorio-variaveis>. Acesso em: 3 set. 2024.

ESTÉVEZ, J. *et al.* "*A quality control procedure for long-term series of daily precipitation data in a semiarid environment*". Theoretical and Applied Climatology, v. 149, n. 3-4, p. 1029–1041, 25 maio 2022.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Censo Demográfico 2022*. Rio de Janeiro: IBGE, 2022.

LEWIS, E. *et al.* "*Quality control of a global hourly rainfall dataset*". Environmental Modelling & Software, v. 144, p. 105169, out. 2021.

LLABRÉS-BRUSTENGA, A. *et al.* "*Quality control process of the daily rainfall series available in Catalonia from 1855 to the present*". Theoretical and Applied Climatology, v. 137, n. 3-4, p. 2715–2729, 16 jan. 2019.

LUCENA, R. L.; CABRAL JÚNIOR, J. B.; STEINKE, E. T. "*Comportamento Hidroclimatológico do Estado do Rio Grande do Norte e do Município de Caicó*". Revista Brasileira de Meteorologia, v. 33, n. 3, p. 485–496, set. 2018.

MEIRA, M. A. *et al.* "*Quality control procedures for sub-hourly rainfall data: An investigation in different spatio-temporal scales in Brazil*". Journal of Hydrology, v. 613, p. 128358, out. 2022.

XAVIER, A. C. *et al. "New improved Brazilian daily weather gridded data (1961–2020)"*. International Journal of Climatology, v. 42, n. 16, p. 8390–8404, 1 jun. 2022.

XAVIER, A. C.; KING, C. W.; SCANLON, B. R. *"Daily gridded meteorological variables in Brazil (1980-2013)"*. International Journal of Climatology, v. 36, n. 6, p. 2644–2659, 2016.