XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

CLASSIFICAÇÃO AUTOMATIZADA DA QUALIDADE DE DADOS PLUVIOGRÁFICOS: UMA APLICAÇÃO COM DADOS DO CEMADEN E TELEMETRIA NO ESTADO DO RIO GRANDE DO SUL

Abner Lins Silva [1] ; Filipe de Carvalho Lemos [2] Emerson da Silva Freitas[3] ; Victor Hugo Rabelo Coelho [4] ; Cristiano das Neves Almeida [5] ; José Lindemberg Vidal Barbosa[6]

Abstract: This study proposes the automation of a quality control method for sub-hourly precipitation data, with a specific focus on the state of Rio Grande do Sul, Brazil. Initially, a manual classification of data quality was performed using two sources: Cemaden and ANA's Telemetry network. The Cemaden data, previously analyzed and deemed reliable, were used as a reference to define representative intervals for 14 rainfall properties indicative of good quality. A Telemetry station was automatically classified as low quality when one or more of its property values fell outside these defined intervals. The rainfall properties were calculated using 30-minute Minimum Time Intervals (MIT) and compared with nearby Cemaden stations located within a 60 km radius. The accuracy of the method was assessed by comparing the automated classification with the manual reference classification, resulting in a confusion matrix with an accuracy of 88% and an F1-score of 70%.

Resumo: Este trabalho propõe a automatização de um método de controle de qualidade para dados de precipitação com resolução sub-horária, com foco específico no estado do Rio Grande do Sul. Inicialmente, foi realizada uma classificação manual da qualidade dos dados provenientes de duas fontes: Cemaden e Telemetria da ANA. Os dados da Cemaden, previamente analisados e confiáveis, foram utilizados como referência para definir intervalos admissíveis para os valores de 14 propriedades da chuva que seriam indicativas de boa qualidade. Quando uma estação da Telemetria apresentava uma ou mais propriedades fora desses intervalos, era automaticamente classificada como de baixa qualidade. As propriedades da chuva foram calculadas para o MIT (Mínimos Intervalos de Tempo) de 30 minutos e comparadas com estações Cemaden localizadas num raio de até 60 km. A acurácia do método foi avaliada por meio da comparação entre a classificação automática e a classificação manual de controle, resultando em uma matriz de confusão com acurácia de 88% e F1-score de 70%.

Palavras-Chave – Controle de qualidade, Pluviógrafos.

1) Afiliação: Laboratório de Recursos Hídricos e Engenharia Ambiental, UFPB, João Pessoa – PB, e-mail: abner.lins.silva@gmail.com
2) Afiliação: Laboratório de Recursos Hídricos e Engenharia Ambiental, UFPB, João Pessoa – PB, e-mail: filipe_carvalho_l@hotmail.com
3) Afiliação: Instituto Federal do Pernambuco, IFPE, Campus Pesqueira- PE, e-mail: e-mail: emerson.sfreitas@hotmail.com
4) Afiliação: Laboratório de Recursos Hídricos e Engenharia Ambiental, UFPB, João Pessoa – PB, e-mail: almeida74br@yahoo.com.br
5) Departamento de Geociências, UFPB, João Pessoa – PB, e-mail: victor-coelho@hotmail.com
6) Afiliação: Laboratório de Recursos Hídricos e Engenharia Ambiental, UFPB, João Pessoa – PB, e-mail: lindembergvidal@gmail.com

INTRODUÇÂO

A chuva é uma das principais fontes de água doce do planeta, responsável pelo abastecimento dos reservatórios e aquíferos, além de ser fundamental para a manutenção dos ecossistemas terrestres (Rozante et al., 2018; Souza; Azevedo; Araújo, 2012)
A qualidade dos dados de precipitação é um fator essencial para o desenvolvimento de estudos hidrológicos, modelagem ambiental e planejamento de políticas públicas relacionadas à gestão de recursos hídricos. No entanto, séries temporais de dados pluviométricos de alta resolução, como aquelas com intervalos sub-horários, estão sujeitas a diversas fontes de erro, incluindo falhas instrumentais, interferências atmosféricas e problemas de comunicação ou armazenamento dos dados (RUBENS; ALVES, 2022).

No Brasil, duas importantes fontes de dados pluviométricos em tempo quase real são o Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden) e a Rede Hidrometeorológica Nacional, operada por meio da plataforma de Telemetria da Agência Nacional de Águas e Saneamento Básico (ANA). Embora ambos os sistemas disponibilizem dados valiosos, é comum a ocorrência de divergências de qualidade entre suas respectivas séries, o que reforça a necessidade de métodos sistemáticos de controle de qualidade.

Tradicionalmente, o controle de qualidade é realizado de forma manual, envolvendo inspeção visual e análise estatística pontual. No entanto, esse processo pode ser altamente subjetivo, demorado e inviável para grandes volumes de dados com alta resolução temporal. Nesse contexto, torna-se relevante o desenvolvimento de abordagens automatizadas que utilizem critérios objetivos para classificar os dados com base em propriedades estatísticas e físicas das séries (OLIVEIRA et al., 2025).

Este trabalho tem como objetivo desenvolver e aplicar um método automatizado de controle de qualidade para dados sub-horários de precipitação, utilizando como base comparações inter-estacionais. A abordagem consiste em definir intervalos representativos de 14 propriedades estatísticas derivadas dos dados do Cemaden, considerados confiáveis, e utilizá-los como referência para classificar automaticamente os dados das estações da Telemetria da ANA. O método é aplicado ao estado do Rio Grande do Sul, e sua acurácia é avaliada por meio da validação cruzada com uma classificação manual de controle.

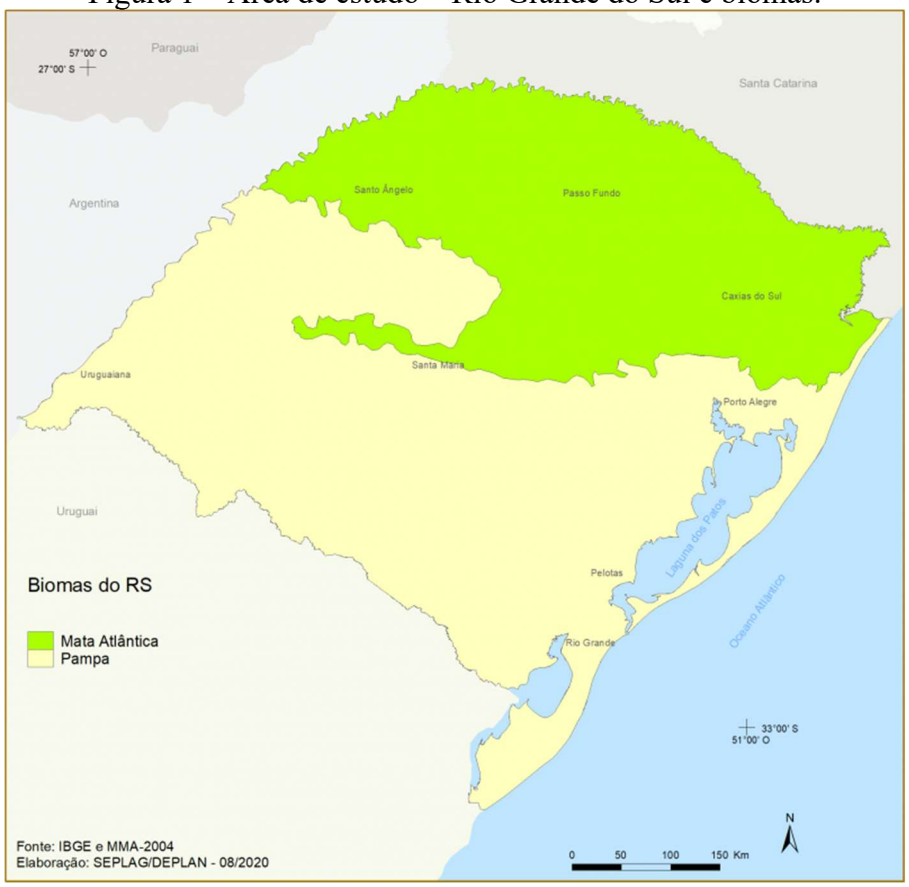
METODOLOGIA

Área de Estudo

O presente estudo abrange todo o território do estado do Rio Grande do Sul (Figura 1), situado na região Sul do Brasil. Embora a escala de análise seja regional, é relevante contextualizar que o Brasil, país de dimensões continentais, possui uma área de aproximadamente 8,5 milhões de km², estendendo-se entre as latitudes 5°16′N e 33°45′S e as longitudes 34°47′W e 73°59′W. A precipitação anual no território brasileiro apresenta elevada variabilidade espacial, com totais que variam entre aproximadamente 380 mm e mais de 4.000 mm, dependendo das características regionais e das condições climáticas predominantes (Gadelha et al., 2019). O estado do Rio Grande do Sul possui clima subtropical úmido, caracterizado por chuvas bem distribuídas ao longo do ano, mas com significativa variabilidade interanual e ocorrência de eventos extremos, como estiagens e tempestades intensas. A precipitação média anual no estado varia entre 1.200 mm e 2.200 mm, sendo que as regiões mais úmidas se concentram no norte e nordeste, especialmente na região serrana, com totais

anuais entre 1.800 mm e 2.200 mm. Por outro lado, as regiões mais secas localizam-se no sudoeste e oeste, próximas à fronteira com a Argentina e o Uruguai, onde os volumes anuais oscilam entre 1.200 mm e 1.500 mm.

Figura 1 – Área de estudo – Rio Grande do Sul e biomas.



Fonte: atlassocioeconomico.rs.gov

Coleta de dados

Os dados de precipitação observados em escala sub-horária utilizados neste estudo estão disponíveis no site do Centro Nacional de Monitoramento e Alerta de Desastres Naturais (CEMADEN). Atualmente, a rede de monitoramento do CEMADEN conta com mais de 3.800 pluviômetros automáticos, que registram a precipitação em tempo real com uma resolução temporal de 10 minutos durante eventos de chuva. Em períodos sem precipitação, essa resolução passa a ser de 60 minutos.

Adicionalmente, foram utilizados dados da rede de telemetria da Agência Nacional de Águas e Saneamento Básico (ANA), que está em operação desde 2000.Essa rede é composta por mais de [preencher a informação] pluviômetros automáticos, com resoluções temporais de 15 e 30 minutos. Assim como os dados do CEMADEN, os dados da ANA passaram por rigoroso controle de qualidade, sendo selecionadas apenas estações consideradas de alta qualidade, ou seja, aquelas que operaram corretamente durante todo o ano analisado.

O processo de pré-seleção das estações incluiu a aplicação dos seguintes critérios:

Períodos de falha prolongada: exclusão de estações com mais de 60 dias consecutivos sem registros;

- Marcadores consecutivos: identificação e exclusão de dados com repetição sistemática de valores;
- Validação de dados: verificação de valores inválidos ou inconsistentes;
- Comparação entre propriedades: análise da coerência dos dados em relação à precipitação observada em estações próximas;
- Análise visual: inspeção detalhada dos dados para identificação de possíveis inconsistências não detectadas pelos métodos automáticos.

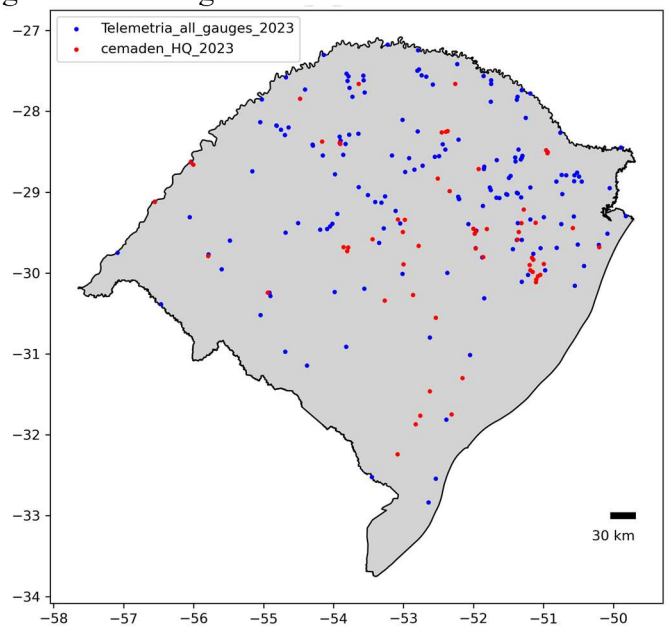
Controle de qualidade dos dados de precipitação

Para realização do controle de qualidade dos dados de precipitação do CEMADEN e da ANA, as estações de monitoramento foram submetidas a um rigoroso controle de qualidade, passando pelos seguintes procedimentos segundo (Freitas et al., 2020; Meira et al., 2022):

- Cada estação foi comparada com as suas 5 vizinhas mais próximas;
- Foram analisadas as séries de precipitação mensal e sub-horárias (resolução temporal de 10,15 e 30 minutos), a fim de identificar inconsistência nos dados medidos;
- As estações que apresentavam diferenças com relação às suas vizinhas eram excluídas;
- Foi realizado o teste duplo cego, ou seja, dois pesquisadores avaliavam a mesma estação individualmente e ao fim as classificações eram confrontadas e analisadas por uma terceira pessoa, para identificar e eliminar as possíveis divergências.

Assim, foi possível construir um banco de dados confiável com séries de precipitação de 2014 a 2024. Os dados da CEMADEN foram usados como referência na calibração do método, enquanto os dados validados da Telemetria serviram posteriormente como controle para avaliação do método.

Figura 2 – Pluviógrafos das redes Telemetria e Cemaden (HQ)



Fonte: autoral (2025)

Criação do método automático de controle de qualidade

A base de dados de estações pluviográficas da rede CEMADEN, considerada de alta confiabilidade, vem sendo atualizada semestralmente desde 2014 utilizando o método descrito acima. No entanto, até o momento, os dados da rede Telemetria ainda não haviam sido integrados com um procedimento sistemático de controle de qualidade. Este estudo propõe um método automatizado como alternativa à verificação manual, visando classificar a qualidade dos dados da rede Telemetria.

Foram utilizadas 14 propriedades de chuva. Dentre elas, quatro propriedades foram calculadas a partir da identificação de eventos com tempo mínimo entre eventos de 30 minutos (MIT = 30 min):

- Tempo seco entre eventos (Dry Time)
- Lâmina de precipitação por evento (Rainfall Depth)
- Duração do evento (Rainfall Duration)
- Intensidade média do evento (Rainfall Intensity)

Para cada uma dessas quatro variáveis, foram determinados os valores médios, máximos e desvios padrão, totalizando 12 propriedades derivadas. Essas métricas foram calculadas para todos os eventos de chuva registrados no ano de 2023 em cada estação da rede Telemetria.

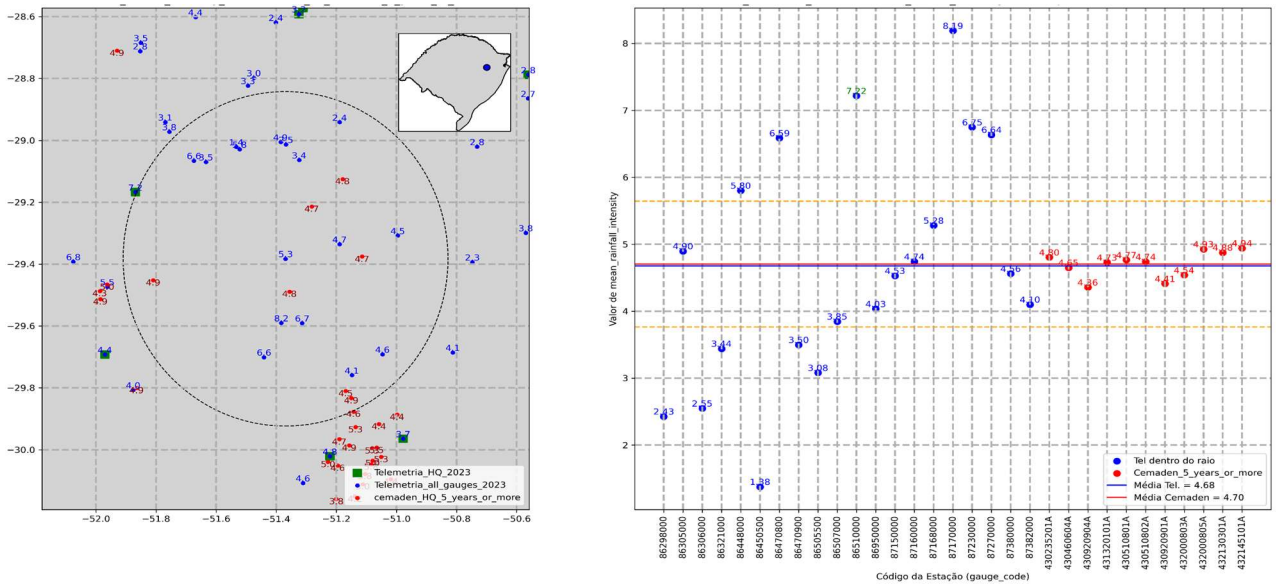
Além dessas 12 propriedades, foram incluídos dois indicadores adicionais:

- Precipitação total anual
- Número total de eventos chuvosos (identificados com MIT de 30 min)

Com isso, cada estação passou a contar com 14 variáveis descritivas para posterior avaliação de qualidade.

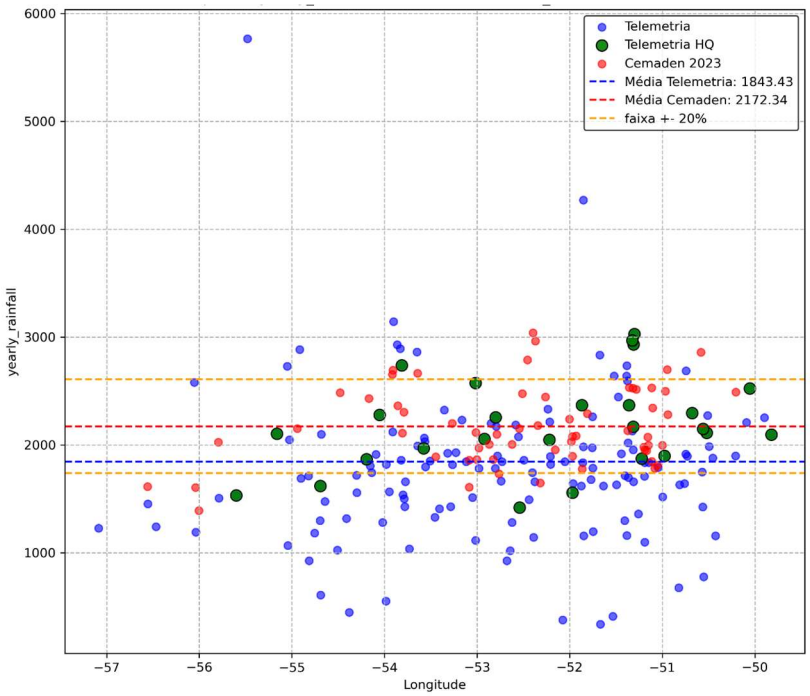
O procedimento de comparação consistiu em identificar, para cada estação Telemetria, as estações CEMADEN localizadas num raio de até 60 km, consideradas como referência de alta qualidade e foi calculada a média de cada uma das 14 propriedades. Em seguida, estabeleceu-se uma faixa aceitável de valores em torno dessa média dentro da qual os dados da estação Telemetria eram classificados como de boa qualidade. Valores fora dessa faixa foram considerados indicativos de má qualidade dos dados.

Figura 3 – Faixa de valores para a propriedade para uma estação típica a ser analizada (mean rainfall intensity)



Fonte: autoral (2025)

Figura 4 – Faixa de valores hipotética para a chuva anual para toda a área de estudo



Fonte: autoral (2025)

As faixas de aceitação foram determinadas com base no método da exaustão, que consiste em repetir sistematicamente o cálculo das classificações para um elevado número de combinações possíveis dos parâmetros envolvidos. Neste caso, considerou-se que a faixa aceitável para cada propriedade seria definida como:

$$[média\ x\ P2\ ;média\ x\ P1]$$

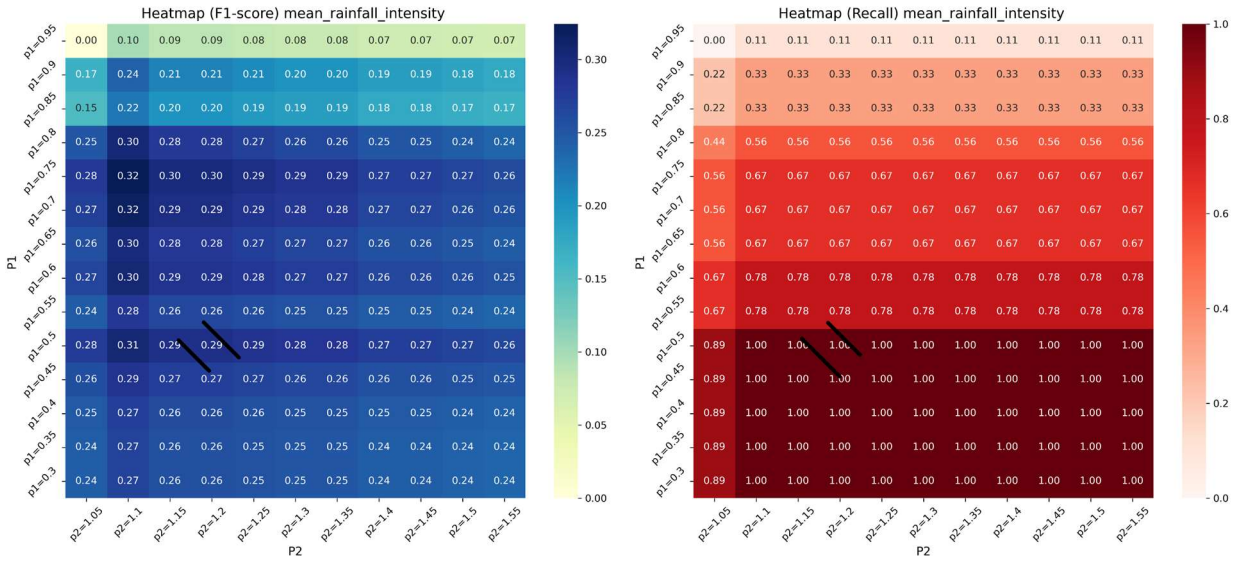
onde:

- $p_1 > 1$ expande o limite superior da faixa
- $p_2 < 1$ reduz o limite inferior da faixa

Foram testadas diversas combinações de p_1 e p_2, e para cada combinação foram geradas as respectivas matrizes de confusão comparando os rótulos de qualidade estimados com uma base de referência previamente validada. A qualidade da classificação foi avaliada por meio das métricas F1-score e Recall, buscando-se valores próximos de 1, indicativos de alta precisão e sensibilidade.

As métricas resultantes foram então visualizadas em heatmaps, permitindo a identificação dos ótimos locais.

Figura 5 – Heatmap do F1-Score e Recall para Combinações de p_1 e p_2 para intensidade média (mean rainfall intensity)



Fonte: autoral (2025)

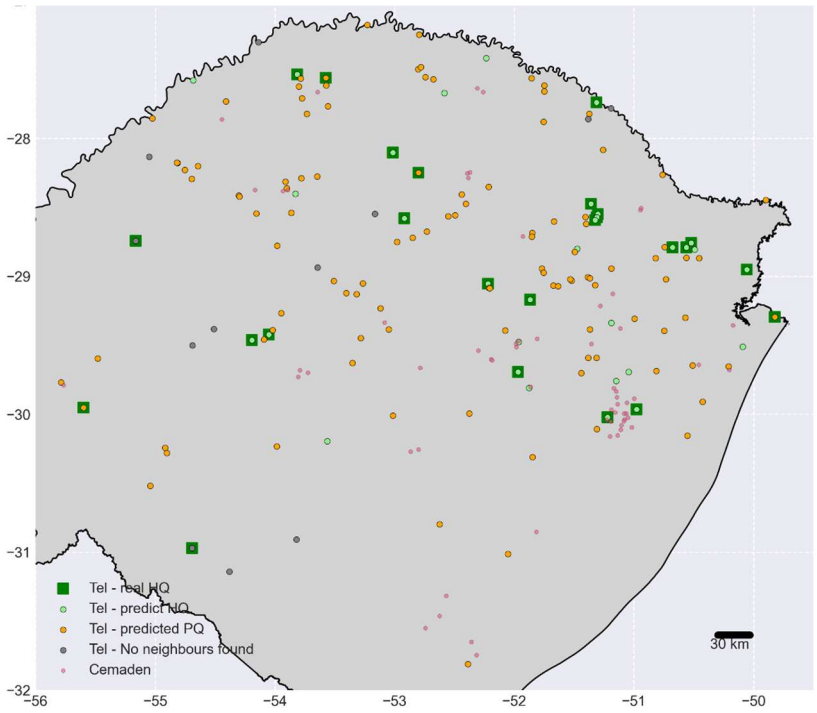
O procedimento foi repetido para cada uma das 14 propriedades, possibilitando a calibração individual das faixas de aceitação conforme a variabilidade de cada variável. Uma estação foi classificada como de alta qualidade pelo modelo apenas se todos os seus 14 valores estivessem dentro dos intervalos definidos a partir das estações de referência.

RESULTADOS

A aplicação do método automatizado de controle de qualidade aos dados sub-horários de precipitação resultou em uma classificação eficiente das estações da rede Telemetria da ANA, com base nos intervalos de referência obtidos a partir dos dados da rede Cemaden.

A Figura 6 apresenta a distribuição espacial das estações analisadas no estado do Rio Grande do Sul, destacando a classificação final obtida (boa ou má qualidade). Observa-se uma maior conformidade da classificação do modelo e a classificação de controle (manual).

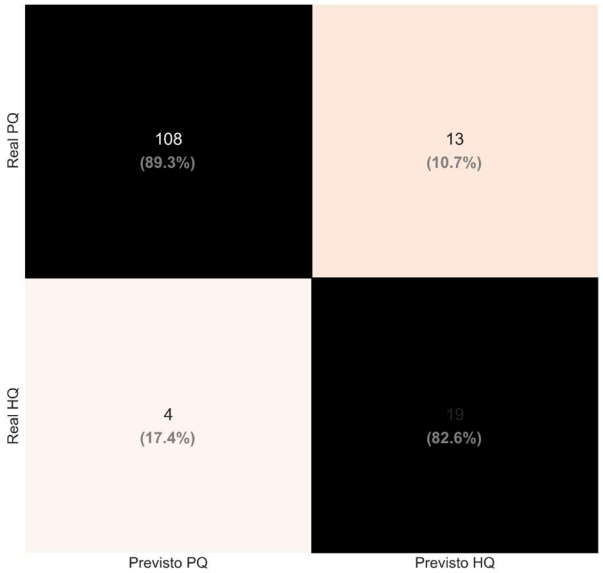
Figura 6 – Mapa da classificação de qualidade das estações Telemetria



Fonte: autoral (2025)

A performance do método foi validada por meio da comparação com a classificação manual de controle. A matriz de confusão (Figura 7) resultante apresentou uma **acurácia de 88%**, indicando que o modelo automatizado consegue reproduzir, com alta fidelidade, os julgamentos realizados manualmente. O **F1-score de 70%** evidencia um bom equilíbrio entre precisão e sensibilidade, sobretudo considerando a natureza binária da classificação (boa/má qualidade) e o desequilíbrio entre as classes.

Figura 7– Matriz de confusão da calibração final.

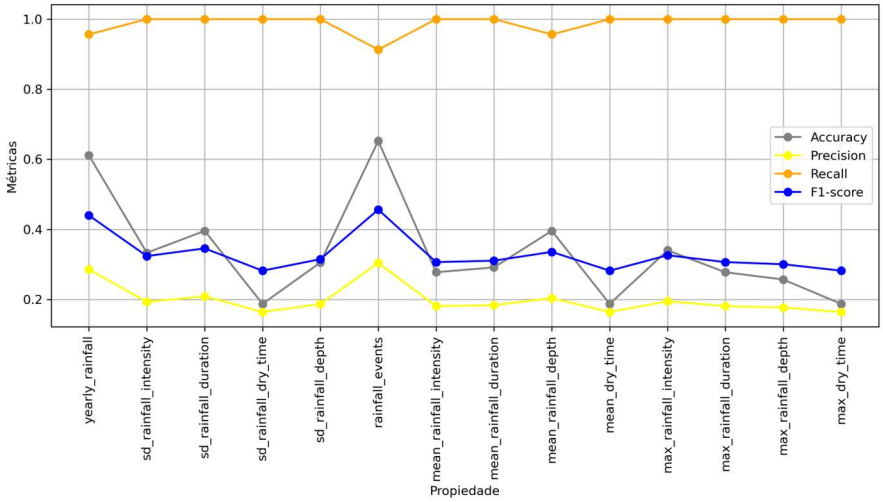


Fonte: autoral (2025)

A matriz de confusão obtida demonstra um bom desempenho do modelo, com destaque para a alta concentração de acertos, principalmente nos verdadeiros positivos e falsos negativos (representados pelos quadrados pretos). Observa-se uma baixa incidência de erros de classificação, como falsos positivos e verdadeiros negativos. Os indicadores de desempenho reforçam essa avaliação, com uma **precisão de 60%** e um **recall de 83%**, evidenciando que o modelo é mais eficaz em identificar corretamente os dados de boa qualidade, mesmo que com alguma limitação na exclusão dos dados inconsistentes.

Entre as 14 propriedades analisadas o número de eventos, foi a mais eficaz em identificar anomalias. aquelas relacionadas à consistência temporal, como o número de intervalos com zero precipitação e o tempo entre picosA Figura 8 ilustra a distribuição de uma dessas propriedades, destacando os valores-limite usados para a classificação.

Figura 8 – Métricas de desempenho para cada propriedade



Fonte: autoral (2025)

Isoladamente, nenhuma propriedade foi suficiente para identificar todos os dados inconsistentes, mas a combinação das 14 variáveis permitiu um ajuste ótimo do modelo, resultando em alta precisão na detecção de dados de baixa qualidade.

CONCLUSÃO

Este estudo propôs um método automatizado de controle de qualidade para dados de precipitação sub-horária, com base em propriedades estatísticas derivadas de dados confiáveis da rede CEMADEN. Aplicado ao estado do Rio Grande do Sul, o método obteve resultados promissores, com 88% de acurácia e F1-score de 70%, validado por classificações manuais. A abordagem demonstrou ser eficaz na detecção de dados inconsistentes, oferecendo uma alternativa objetiva e eficiente ao controle manual, especialmente em grandes volumes de dados.

O uso de comparações espaciais em um raio de 60 km entre estações contribuiu para a identificação de padrões regionais de qualidade. A metodologia é flexível, podendo ser adaptada a outras regiões e resoluções temporais.

Futuras melhorias incluem ajustes sazonais dinâmicos e o uso de inteligência artificial. Também está em andamento a aplicação do método em escala nacional, com validação baseada em dados de controle manual.

REFERÊNCIAS

Agência Nacional de Águas e Saneamento Básico (ANA). (2019). **HidroWeb** v3.2.7. Brasil. https://www.snirh.gov.br/hidroweb/apresentacao

BLENKINSOP, S., LEWIS, E., CHAN, S. C., & FOWLER, H. J. (2017). Quality-control of hourly rainfall dataset and climatology of extremes for the UK**. International Journal of Climatology,** 37(2), 722–740. https://doi.org/10.1002/joc.4735

FREITAS, E. et al. The performance of the IMERG satellite-based product in identifying sub-daily rainfall events and their properties. **Journal of Hydrology**, v. 589, 5 jun. 2020.

GUPTA, Hoshin V. et al. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. **Journal of hydrology**, v. 377, n. 1-2, p. 80-91, 2009.

HAJANI, E. Climate change and its influence on design rainfall at-site in New South Wales State, Australia**. Journal of Water and Climate Change**, v. 11, n. S1, p. 251–269, 8 jul. 2020.

MEIRA, M. A. Quality control procedures for sub-hourly rainfall data: An investigation in different spatio-temporal scales in Brazil. **Journal of Hydrology**, p. 13, 2022.

OLIVEIRA, A. C. DE et al. Avaliação de Desempenho de Classificadores a Partir de Dados Relacionados à Precipitação Pluviométrica Coletados por Estação Meteorológica Automática. **Revista Eletrônica de Iniciação Científica em Computação**, v. 23, p. 24–29, 14 abr. 2025.

RUBENS; ALVES, R. Simulação de valores ausentes em séries temporais de precipitação para avaliação de métodos de imputação. **Revista Brasileira de Climatologia**, v. 30, p. 691–714, 10 jun. 2022

AGRADECIMENTOS

Os autores deste artigo agradecem aos seguintes órgãos: CNPq (bolsa PQ); CAPES e FAPESQPB (bolsas de mestrado e doutorado); CEMADEN (Centro Nacional de Monitoramento e Alerta a Desastres Naturais) pela disponibilização dos dados de precipitação.