Predição Iterativa de Vazão de Rede Utilizando Uma Abordagem de Janelas Deslizantes combinada com Imputação de Dados Dinâmica

Maria C. Mesquita[1], **Maria L. Linhares**[1], **Ariel L. Portela**[1], **Ivo A. Pimenta**[1], **Thelmo P. Araújo**[1], **Rafael L. Gomes**[1]

[1]Universidade Estadual do Ceará (UECE)

{clara.mesquita,malu.linhares,ariel.portela,

aguiar.pimenta}@aluno.uece.br,{thelmo.araujo, rafa.lopes}@uece.br

Abstract. *Throughput prediction is essential for proactive network management in Internet Service Providers (ISPs); however, real-world time series often exhibit gaps resulting from limitations in data collection mechanisms. This paper proposes an approach that operates in an autoregressive manner, utilizing actual values when available and predictions when observed data are missing, thereby integrating fault handling directly into the predictive process. Experiments conducted with real-world data from Ipê network infraestructure from National Education and Research Network (RNP) demonstrate that the approach maintains consistent performance even in scenarios with significant measurement failures, establishing it as a viable solution for supporting preventive network management.*

Resumo. *A predição de vazão é fundamental para a gestão proativa de redes em provedores de rede, porém séries temporais reais frequentemente apresentam lacunas decorrentes de limitações nos mecanismos de coleta. Este trabalho propõe uma abordagem que opera de forma autorregressiva, empregando valores reais quando disponíveis e predições quando dados observados estão ausentes, integrando o tratamento de falhas diretamente ao processo preditivo. Experimentos com dados reais da infraestrutura da rede Ipê da Rede Nacional de Pesquisa (RNP) demonstram que a abordagem mantém desempenho consistente mesmo em cenários com falhas significativas de medição, configurando-se como uma solução viável para suporte à gestão preventiva de redes.*

1. Introdução

A gestão eficiente de infraestruturas de rede de grande escala depende do monitoramento contínuo da vazão de tráfego, uma métrica central de Qualidade de Serviço e de Experiência (*Quality of Service/Quality of Experience*, QoS/QoE) que reflete diretamente a capacidade da rede, seu nível de utilização e a experiência dos usuários[Nobre et al. 2025, Gomes et al. 2016]. Nestes cenários, a medição precisa da vazão é essencial para caracterizar desempenho, utilização e gargalos em infraestruturas de rede [Mutter and Shannigrahi 2024, Gomes et al. 2013]. Contudo, o monitoramento da vazão é dificultado pela variação de latência e perda de dados inerentes aos modelos

de consulta tradicionais sob congestionamento, pela elevada sobrecarga gerada no plano de controle para garantir precisão em tempo real e pelas limitações de *hardware* que restringem a execução de cálculos matemáticos complexos diretamente nos dispositivos de rede [Pimenta et al. 2024, Gomes et al. 2014].

Dentre as etapas fundamentais de tratamento e preparação de dados para a aplicação de modelos de aprendizagem de máquina, o tratamento de dados faltantes destaca-se como uma das mais críticas, especialmente para Redes Neurais Recorrentes, já que lacunas não tratadas introduzem viés, perda de informação e instabilidade, comprometendo o aprendizado da dinâmica temporal [**?**, Brito et al. 2026]. No contexto de séries temporais univariadas de desempenho de rede, como tipicamente ocorre em dados de vazão [Na et al. 2023, Ferreira et al. 2025], essa questão é ainda mais sensível: embora métricas como vazão, latência e perda de pacotes apresentem fortes correlações espaço-temporais que viabilizam inferência de valores ausentes, a qualidade dessa recuperação depende fortemente do método adotado [Portela et al. 2024, Lopes Gomes and Roberto Mauro Madeira 2012].

A incompletude dos dados representa um desafio significativo para a análise de séries temporais e para a predição de tráfego de rede. Embora técnicas simples de imputação (como média ou interpolação linear) sejam computacionalmente eficientes, elas geralmente não preservam a dinâmica temporal dos dados, especialmente em cenários com taxas moderadas ou elevadas de ausência [Ferreira et al. 2025]. Estudos recentes demonstram que o tratamento inadequado de dados faltantes pode degradar substancialmente o desempenho de modelos preditivos, resultando em estimativas enviesadas e análises operacionais pouco confiáveis [Ferreira et al. 2024]. Nesse contexto, estratégias robustas de imputação tornam-se um requisito essencial para análises de rede fidedignas.

Avanços recentes na literatura propuseram métodos sofisticados de imputação voltados especificamente para dados de vazão de rede, explorando representações latentes e estruturas temporais de baixo posto para reconstruir valores ausentes de forma mais consistente. Esses trabalhos evidenciam que a preservação dos padrões temporais intrínsecos é determinante para a melhoria do desempenho de modelos de previsão subsequentes [Hu et al. 2024, Du et al. 2024, Ferreira et al. 2025, Ferreira et al. 2024]. No entanto, a maioria dessas abordagens ainda trata a imputação como uma etapa estática de pré-processamento, realizada uma única vez antes do treinamento do modelo [Du et al. 2024, Ferreira et al. 2025, Ferreira et al. 2024, Li 2024].

A predição de desempenho e tráfego de rede é um elemento central para o aprimoramento da QoS e da QoE em redes modernas, permitindo que Provedores de Serviços de Internet (ISPs) antecipem variações de carga, flutuações de vazão e potenciais gargalos, viabilizando estratégias proativas de alocação de recursos, balanceamento de carga e planejamento de capacidade [Kablaoui et al. 2024]. Nesse cenário, modelos baseados em aprendizado de máquina, especialmente arquitetura de Redes Neurais Recorrentes como *Long Short-Term Memory* (LSTM), destacam-se pela capacidade de capturar dependências temporais de curto e longo prazo, superando abordagens estatísticas tradicionais na predição de tráfego e vazão [Yalda et al. 2024]. Contudo, em ambientes reais de operação, a elevada variabilidade dos padrões de uso, a heterogeneidade do tráfego e as limitações inerentes aos dados de monitoração tornam a predição um problema complexo, exigindo modelos que sejam não apenas acurados, mas também robustos e adaptáveis às condições

dinâmicas das redes [Al-Thaedan et al. 2023].

Para superar essas limitações, este trabalho propõe uma solução de predição de tráfego baseada em janelas deslizantes com autorregressão. Inicialmente, realiza-se a imputação dos dados faltantes para o treinamento do modelo, com ajustes dinâmicos do tamanho da janela de *look-back* e do horizonte de predição. O principal diferencial da metodologia reside no uso das próprias predições do modelo como imputação de valores futuros ausentes, estabelecendo um ciclo contínuo de realimentação que ajusta o processo de treinamento à medida que novas predições são geradas. Ao integrar imputação e predição em um *framework* adaptativo unificado, a abordagem proposta mitiga inconsistências temporais e aumenta a precisão das predições, bem como a resiliência do processo de planejamento de redes.

A metodologia proposta é avaliada com dados reais de vazão da infraestrutura da rede Ipê da Rede Nacional de Pesquisa (RNP), coletados via MonIpê[1], evidenciando sua eficácia no tratamento de séries temporais incompletas e seu potencial para apoiar a gestão de redes de forma escalável. A avaliação utiliza métricas como Raiz do Erro Quadrático Médio (*Root Mean Square Error* - RMSE), Raiz do Erro Quadrático Médio Normalizado (*Normalized Root Mean Square Error* - NRMSE) e Erro Percentual Absoluto Médio Simétrico (*Symmetric Mean Absolute Percentage Error* - SMAPE), analisando a propagação do erro ao longo do horizonte de predição para garantir um desempenho preditivo consistente. De maneira geral, as contribuições deste artigo são o desenvolvimento de um *framework* que traz a abordagem de inferência robusta a dados faltantes utilizando máscaras e atualização híbrida do estado com janelas deslizantes.

O restante deste artigo está organizado da seguinte forma: A Seção 2 apresenta o estado da arte e os trabalhos relacionados ao contexto de imputação, predição e monitoramento de desempenho de rede, a Seção 3 descreve a proposta, enquanto a Seção 4 detalha os experimentos realizados, e a Seção 5 discute os resultados. Finalmente, a Seção 6 conclui o artigo e apresenta diversos trabalhos futuros.

2. Trabalhos Relacionados

Esta seção apresenta uma análise de trabalhos científicos que estão relacionados com o que pode ser o estado-da-arte no que se diz respeito a pesquisas sobre predição de dados considerando falhas importantes, pensando na melhoria da qualidade destes para o então uso em aplicações como na predição de dados de vazão de rede para possibilitar uma gestão adaptativa de recursos. Nesse contexto, trabalhos de predição com redes neurais e aprendizagem de máquina foram considerados, dando enfoque em casos onde a imputação de dados foi implementada e analisada, para avaliar a superioridade e relevância da presente proposta.

O trabalho de Na et al., [Na et al. 2023] propõe um sistema de predição de vazão de rede voltado a redes LTE, com foco em aplicações sensíveis à latência, como *streaming* de vídeo. A ideia central é empregar uma LSTM com mecanismo de atenção para melhorar a precisão da predição de vazão futura, explorando dependências temporais mais relevantes dentro da janela de observação. Ele utiliza para validação um conjunto de dados sintético derivado de *logs* tcp em redes LTE, sem uma sonda de vazão real (como

[1]https://memoria.rnp.br/pd/monipe.html

iPerf, perfSONAR, SNMP, etc.) completa e ideal. Nesse caso, ele não trata explicitamente dados faltantes nem dados reais, ignorando completamente os cenários de falhas de coleta comuns em ambientes reais. Em cenários reais, a atenção pode atribuir alto peso a um ponto artificial, amplificando erro, uma vez que o mecanismo de atenção necessariamente não sabe distinguir valor real observado e valor sintético (imputado ou predito), que pode acabar recebendo peso alto na janela.

Em seu trabalho, Neog et al. [Neog et al. 2026] propõe uma abordagem agnóstica e livre de imputação que utiliza o *Time-Feature Independent* (TFI) *Embedding*, onde é criado um *token* por par (tempo e variável) e só *embeddings* observados são processados, e o *Missing Feature-Aware Attention* (MFAA), variáveis faltantes são explicitamente ignoradas e apenas a informação observada é utilizada para formar a representação latente de cada tempo, mas ignorando variáveis faltantes. Contudo, o elevado custo computacional e a natureza estrutural (não temporal) do método limitam sua eficácia em séries de vazão complexas e não estacionárias, que exigem a captura de dinâmicas não-lineares. Além disso, o uso de mascaramento no treinamento pode comprometer a robustez do modelo em cenários reais [Qian et al. 2025]. Para sanar as limitações de dados faltantes no conjunto de predição, [Li 2024] propõe uma abordagem clássica de dois estágios, como de praxe [Ahn et al. 2021, Ferreira et al. 2025]: imputação de dados com aprendizagem de máquina, utilizando uma Conditional Generative Adversarial Network (C-GAN), que é treinada para imputar valores ausentes aprendendo a distribuição de dados completos; e, na fase de predição, aplica inferência bayesiana para quantificar a incerteza das previsões. Embora eficaz, a abordagem proposta em [Li 2024] apresenta limitações associadas ao uso de *Generative Adversarial Networks* (GANs), incluindo suscetibilidade a *overfitting* e elevado custo computacional. Além disso, a necessidade de reconstrução completa da série temporal antes da predição pode reduzir a eficiência do processo

Diferente das abordagens de dois estágios que tratam a imputação como um pré-processamento estático ou externo , a presente proposta utiliza uma arquitetura baseada em janelas deslizantes com realimentação recursiva. Enquanto métodos baseados em atenção ou decomposição latente enfrentam alto custo computacional e dificuldades em capturar dinâmicas temporais não lineares , este trabalho simplifica o pipeline ao utilizar as próprias predições do modelo para sustentar o estado da janela na ausência de medições. Essa integração direta entre o aprendizado da dinâmica temporal e o tratamento de falhas permite uma gestão de rede proativa e resiliente, mesmo em séries temporais univariadas com taxas de omissão superiores a 50%.

3. Proposta

A gestão eficaz de redes de computadores por meio de análises preditivas de vazão permite a identificação preventiva de degradações na transmissão de dados entre pontos distintos da infraestrutura, configurando uma funcionalidade essencial para ISPs. Apesar do avanço das plataformas de telemetria e dos serviços de monitoramento, ainda persistem desafios significativos na utilização de métricas como vazão para o treinamento de modelos de redes neurais, que demandam grandes volumes de dados com elevada qualidade. Nesse contexto, a proposta de um modelo baseado em janelas deslizantes busca lidar com a natureza aleatória e frequentemente incompleta dos fluxos de tráfego em redes de longa distância por meio de um *pipeline* de processamento que prioriza a integridade estrutural da série temporal durante a fase de predição.

Propõe-se um *framework* de predição autorregressiva de vazão baseado em janelas deslizantes que integra, de forma implícita, o tratamento de dados faltantes ao próprio processo preditivo. A Figura 1 apresenta uma visão geral do fluxo, com ênfase no mecanismo recursivo de predição por janela deslizante.

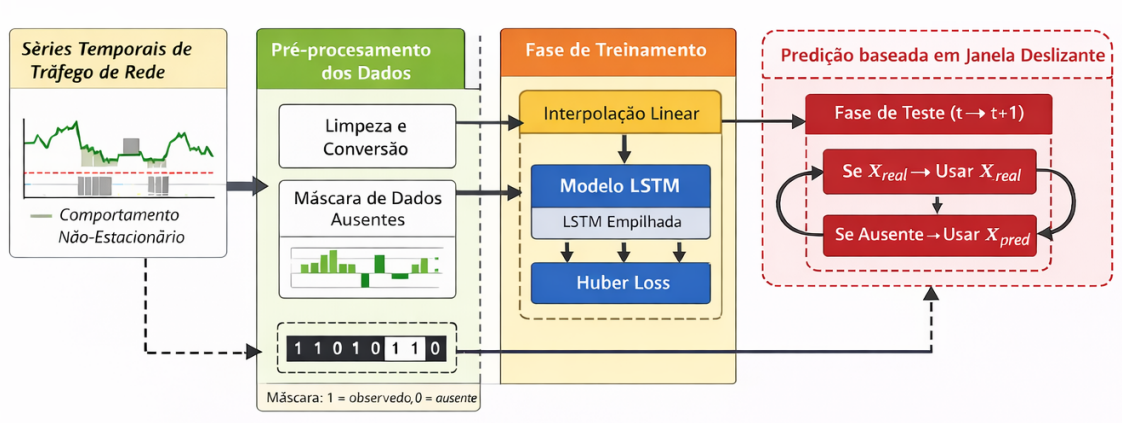

Figura 1. Visão geral da proposta do *framework* de predição

A abordagem utiliza uma LSTM operando de maneira autorregressiva, na qual cada previsão é realizada a partir de uma janela temporal (*look-back*) de observações passadas e, na construção das janelas subsequentes, incorpora-se o valor real quando disponível ou a própria predição quando ocorre falha de medição. Esse mecanismo permite uma imputação adaptativa guiada pela dinâmica temporal aprendida pelo modelo, preservando padrões estruturais da série e evitando etapas externas de imputação. O modelo é treinado com função de perda de Huber, adequada a cenários com alta variabilidade e presença de *outliers*, e seus hiperparâmetros são ajustados por meio de *grid search* com validação temporal, assegurando realismo experimental.

3.1. Coleta de Dados

Os experimentos deste trabalho utilizam dados reais de vazão coletados pelo MonIPÊ, serviço oficial de monitoramento de desempenho da Rede Nacional de Ensino e Pesquisa (RNP). A plataforma foi concebida para acompanhar a saúde da rede, permitindo identificar degradações de desempenho que nem sempre se manifestam como indisponibilidade total. Para esse fim, o MonIPÊ realiza medições contínuas de métricas essenciais à operação de redes acadêmicas, como vazão, latência, *jitter* e perda de pacotes, fornecendo uma base observacional adequada para estudos de predição sob condições realistas, incluindo ruído, não estacionariedade e falhas de medição.

Do ponto de vista arquitetural, o MonIPÊ segue o padrão internacional perfSONAR, amplamente adotado em redes acadêmicas, executando testes recorrentes em malha completa entre pontos de presença distribuídos nacionalmente. As medições são conduzidas com ferramentas especializadas, como o iPerf3 para estimativa de vazão e oOWAMP para medição precisa de latência unidirecional, além do *Network Diagnostic Tool* (NDT) para diagnósticos de conectividade sob a perspectiva do usuário final. Devido a limitações históricas da API, a coleta foi realizada de forma incremental ao longo de aproximadamente 18 meses, compreendendo os anos de 2024 e 2025, com posterior reamostragem das séries em intervalos regulares de 6 horas, a fim de padronizar a granularidade temporal

e reduzir irregularidades do processo de medição. Todos os experimentos consideraram exclusivamente medições de vazão associadas ao protocolo BBR, dada sua relevância em cenários de alta capacidade e sua adoção crescente em redes acadêmicas.

3.2. Predição Autorregressiva com Imputação Implícita Baseada em Janelas Deslizantes

Embora a vazão seja uma métrica central para a gestão preventiva de redes por ISPs, séries temporais reais obtidas por plataformas de monitoramento, como mostrado nas sessões anteriores, frequentemente contêm lacunas, leituras inválidas e ruído, o que pode inviabilizar o uso direto de modelos preditivos convencionais. Assim, propõe-se um *framework* de predição autorregressiva que integra o tratamento de dados faltantes ao próprio processo de inferência, preservando a coerência temporal da série e mantendo a operação mesmo sob falhas persistentes de coleta. A seguir, detalha-se as etapas da solução.

3.2.1. Pré-processamento de Dados

Para cada enlace monitorado, a série temporal de vazão é carregada e convertida para uma unidade consistente (e.g., Mbps), tratando representações explícitas de ausência (como -1) como valores faltantes. Seja $\{x_t\}_{t=1}^{T}$ a série de vazão (convertida para Mbps). Essa máscara é preservada durante a inferência e governa a lógica híbrida (observação vs. predição) do processo recursivo.

Em seguida, a série é particionada de forma causal em treino e teste, sem embaralhamento:

$$\mathcal{D}_{\text{train}} = \{x_t\}_{t=1}^{\lfloor 0.8T \rfloor}, \qquad \mathcal{D}_{\text{test}} = \{x_t\}_{t=\lfloor 0.8T \rfloor+1}^{T}. \tag{1}$$

Para garantir continuidade mínima durante o ajuste dos pesos, aplica-se interpolação linear apenas na primeira janela do conjunto de treino (evitando inserção de sinal artificial no futuro). Em seguida, ajusta-se um escalonamento Min–Max somente com o treino e aplica-se a mesma transformação ao teste:

$$z_t = \text{scale}(x_t) = \frac{x_t - x_{\min}}{x_{\max} - x_{\min} + \varepsilon}, \quad x_{\min} = \min(\mathcal{D}_{\text{train}}),\ x_{\max} = \max(\mathcal{D}_{\text{train}}), \tag{2}$$

onde ε é um termo pequeno para estabilidade numérica. Denota-se por $\{z_t\}$ a série normalizada.

A série normalizada é convertida em um problema supervisionado por meio de janelas deslizantes de tamanho k (parâmetro *look-back*). Para cada instante $t > k$, define-se a entrada (estado) e o alvo:

$$\mathbf{s}_t = [z_{t-k}, z_{t-k+1}, \ldots, z_{t-1}]^\top \in \mathbf{R}^k, \qquad y_t = z_t. \tag{3}$$

O conjunto supervisionado é então $\{(\mathbf{s}_t, y_t)\}_{t=k+1}^{|\mathcal{D}_{\text{train}}|}$.

3.2.2. Fase de treinamento

Treina-se um modelo LSTM empilhado com duas camadas recorrentes e uma camada densa de saída:

$$\hat{y}_t = f_\theta(\mathbf{s}_t), \tag{4}$$

onde f_θ é parametrizado por pesos θ. O treinamento minimiza a função de perda Huber, mais robusta a rajadas e *outliers* típicos de vazão:

$$\mathcal{L}_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{se } |y - \hat{y}| \leq \delta, \\ \delta\left(|y - \hat{y}| - \frac{1}{2}\delta\right), & \text{caso contrário.} \end{cases} \tag{5}$$

No sistema implementado, utiliza-se $\delta = 0.25$, e a otimização é feita com Adam, com taxa de aprendizado avaliada no *grid search*. Para reduzir sobreajuste, aplica-se *early stopping* monitorando a perda de validação, com restauração dos melhores pesos.

3.2.3. Predição baseada em Janelas Deslizantes

A inferência ocorre de forma sequencial no conjunto de teste, com atualização recursiva do estado. Seja $\mathbf{s}_t$ o estado corrente em escala normalizada e $\hat{z}_t$ a predição do modelo para o próximo ponto. Denote por m_t a máscara associada ao instante t no teste. O valor efetivamente utilizado para avançar a janela é:

$$\tilde{z}_t = \begin{cases} z_t, & \text{se } m_t = 1 \text{ (observado)}, \\ \hat{z}_t, & \text{se } m_t = 0 \text{ (ausente)}. \end{cases} \tag{6}$$

Então, o deslocamento da janela (janela deslizante) é:

$$\mathbf{s}_{t+1} = [z_{t-k+1}, z_{t-k+2}, \ldots, z_{t-1}, \tilde{z}_t]^\top. \tag{7}$$

Esse mecanismo realiza a imputação de modo *implícito* durante a predição: quando há observação real, o estado é ancorado no dado observado; quando há lacuna, a própria predição sustenta o estado, mantendo o processo operacional sem uma etapa externa de imputação no período de teste.

4. Experimentos

A seleção dos conjuntos de dados adotados neste trabalho foi guiada por critérios estatísticos mínimos e representativos, capazes de refletir desafios reais enfrentados na predição de vazão em redes de longa distância. Os enlaces escolhidos apresentam taxas expressivas de dados faltantes em janelas de 6 horas, variando aproximadamente entre 29% e 53%, o que caracteriza um cenário adverso para abordagens tradicionais de predição. Essa propriedade é central para a proposta baseada em janelas deslizantes, uma vez que o método busca manter a coerência temporal local mesmo na presença de lacunas frequentes. Adicionalmente, o número de registros disponíveis em cada enlace assegura séries temporais suficientemente longas para avaliação consistente do modelo ao longo do tempo.

Do ponto de vista estatístico, os enlaces selecionados exibem médias elevadas de vazão, compatíveis com enlaces de *backbone*, além de diferentes níveis de variabilidade relativa, capturados pelo coeficiente de variação. A utilização do intervalo interquartil como medida de dispersão robusta permite evidenciar a heterogeneidade do tráfego sem influência excessiva de *outliers*, comuns em ambientes acadêmicos sujeitos a transferências científicas massivas. Por fim, a diversidade geográfica dos enlaces, conectando estados de diferentes regiões do país, reforça a representatividade da amostra ao incorporar distintos padrões de demanda, distância física e comportamento de uso da rede, fortalecendo a validade externa da avaliação da abordagem proposta. A tabela 1 traz um resumo de algumas dessas informações que ajudaram a escolher a amostra para testes.

Tabela 1. Resumo estatístico dos links analisados (agregação em 6h)

Link	Quant.	Faltantes (%)	Média	CV	IQR
ac-to	2213	39.09	7.91×10^8	0.118	1.31×10^8
es-ce	1897	29.47	9.15×10^8	0.032	1.17×10^7
es-pr	2213	52.96	9.12×10^8	0.029	1.06×10^7
go-sp	783	49.17	9.65×10^8	0.038	1.18×10^7
ms-to	2207	41.10	9.04×10^8	0.040	1.69×10^7
sc-to	2212	46.34	9.41×10^8	0.040	7.85×10^6

4.1. Definição de Janela Deslizantes

Para simular operação real e limitar custo computacional, as previsões no teste são executadas em blocos de tamanho W (`WINDOW_SIZE`), preservando o estado entre blocos. No experimento, $W = 28$, que no caso equivale a uma semana. Em cada passo do bloco: (i) prediz-se $\hat{z}_t$ a partir do estado corrente; (ii) obtém-se a predição em unidade original por inversão da normalização; (iii) atualiza-se a janela via Eq. (6) e (7). Esse procedimento corresponde a uma avaliação temporal baseada em janelas deslizantes e permite analisar propagação de erro e robustez sob faltas persistentes.

4.2. Definição de Hiperparâmetros

Os hiperparâmetros são selecionados por *grid search* acoplado à validação cruzada temporal com K divisões (TimeSeriesSplit), preservando causalidade em cada dobra. Para cada combinação $\lambda \in \Lambda$ (taxa de aprendizado), $E \in \mathcal{E}$ (épocas), $B \in \mathcal{B}$ (*batch size*) e $P \in \mathcal{P}$ (*patience*), treina-se o modelo em cada dobra e calcula-se a perda média de validação $\overline{\mathcal{L}}_{\text{val}}$ ao longo das épocas (com *early stopping*).

Na implementação, utilizam-se $K = 5$ dobras; $\Lambda = \{10^{-3}\}$, $\mathcal{E} = \{100, 300, 500\}$, $\mathcal{B} = \{32, 64, 128\}$ e $\mathcal{P} = \{10\}$, com restauração dos melhores pesos por dobra. Em seguida, re-treina-se o modelo final com os hiperparâmetros selecionados no treino (com validação interna), e aplica-se a inferência recursiva no teste. Em termos de arquitetura, empregam-se 64 unidades por camada LSTM, duas camadas recorrentes empilhadas e uma saída densa (valor escalar).

4.2.1. Métricas de Desempenho

Para avaliação da proposta, foram escolhidas as MAE, RMSE, SMAPE e Kappa de Cohen. Essas métricas se complementam para avaliar tanto a magnitude do erro quanto a fidelidade ao padrão temporal da vazão. A MAE mede o erro médio em unidades reais (quanto, em média, a previsão se distancia do valor observado), sendo fácil de interpretar e menos sensível a picos extremos. A RMSE também mede erro em unidades reais, mas penaliza mais fortemente erros grandes, o que é útil quando erros em rajadas/picos são especialmente críticos em cenários de rede.

A sMAPE avalia o erro percentual de forma mais "equilibrada" (erro relativo), ajudando a comparar desempenho quando a série tem variações de escala ao longo do tempo ou quando links têm amplitudes diferentes. Já o Kappa de Cohen foca na aderência do comportamento: indica o quanto a predição acompanha o formato da série (tendên-

cia, oscilações e sazonalidade), mesmo quando a magnitude não é perfeita, o que é algo importante em vazão, onde preservar padrões temporais pode ser tão relevante quanto minimizar o erro absoluto.

Além das métricas pontuais, avalia-se a degradação do erro em inferência recursiva, caracterizando a propagação de erro quando previsões sucessivas são geradas ao longo do teste. O processo segue um *rollout* causal: a cada passo t, o estado é atualizado com o valor real quando disponível ou com a própria predição quando o ponto é ausente:

$$\tilde{y}_t = \begin{cases} y_t, & \text{se } m_t = 1, \\ \hat{y}_t, & \text{se } m_t = 0. \end{cases} \tag{8}$$

No método proposto, o estado corresponde a uma janela de tamanho k:

$$\mathbf{s}_t = [\tilde{y}_{t-k}, \ldots, \tilde{y}_{t-1}], \qquad \hat{y}_t = f(\mathbf{s}_t), \qquad \mathbf{s}_{t+1} = [\tilde{y}_{t-k+1}, \ldots, \tilde{y}_t]. \tag{9}$$

A degradação é computada por horizonte $h \in \{1, \ldots, H\}$, agregando o erro apenas nos instantes observados. Define-se:

$$\mathrm{RMSE}(h) = \sqrt{\frac{1}{|\mathcal{I}_h|} \sum_{t \in \mathcal{I}_h} (\hat{y}_t - y_t)^2}, \tag{10}$$

onde $\mathcal{I}_h$ representa os índices válidos (com $m_t = 1$) associados ao horizonte h. Por fim, sumariza-se o efeito do horizonte por:

$$\Delta\mathrm{RMSE} = \mathrm{RMSE}(H) - \mathrm{RMSE}(1), \qquad \mathrm{Deg}_{rel} = \frac{\Delta\mathrm{RMSE}}{\mathrm{RMSE}(1)}. \tag{11}$$

5. Resultados

Esta seção apresenta os resultados obtidos da análise de experimentos de avaliação conduzidos com um conjunto de dados real, discutindo os principais pontos referentes à validação e eficácia do modelo de predição de dados de vazão autorregressivo baseado em janelas deslizantes.

5.1. Desempenho Global

Analisando os resultados das métricas de avaliação do modelo, na figura 2, a vantagem é mais evidente em ac-to, o enlace mais difícil: enquanto os baselines ficam próximos de 8,9% (SARIMA) e 9,6–9,8% (GRU/LSTM), a proposta deste artigo reduz para 7,8%. Em enlaces com erros baixos, como go-sp, e também mantém a liderança (0,6%), e em enlaces intermediários (ms-to, sc-to, es-pr) os ganhos são menores, mas consistentes, mantendo-se como o menor SMAPE.

O mesmo padrão aparece em RMSE e MAE, nas figuras 3 e 4. Em ac-to, a proposta reduz o RMSE de 89–94 para 78–79 e o MAE de 60–67 para 52–53, indicando ganho relevante no cenário mais crítico. Em es-ce e sc-to, e também supera os baselines com folga (especialmente contra GRU/LSTM em RMSE), enquanto em ms-to as diferenças são mais discretas, porém ainda favoráveis.

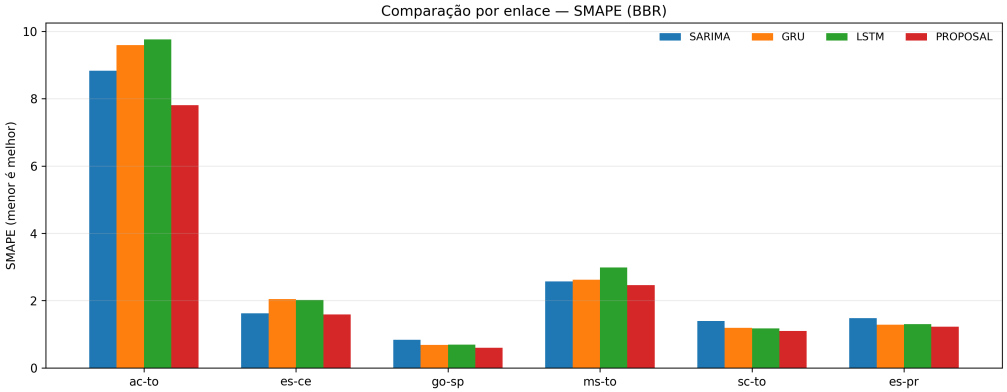

Figura 2. Comparação de SMAPE (%)

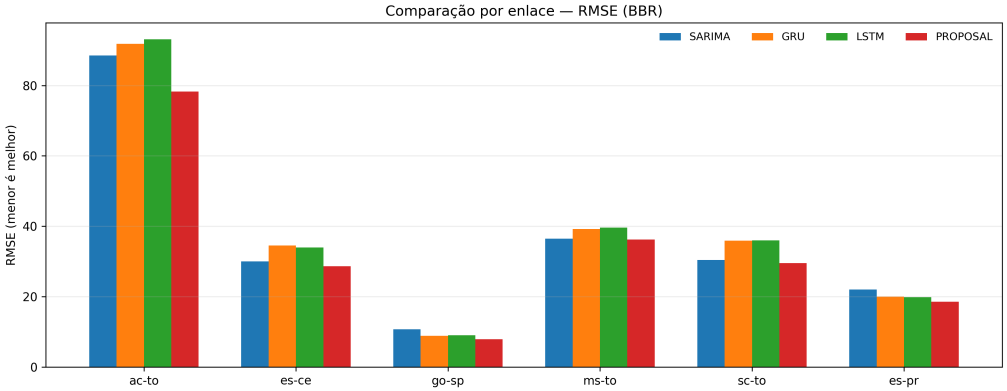

Figura 3. Comparação de RMSE

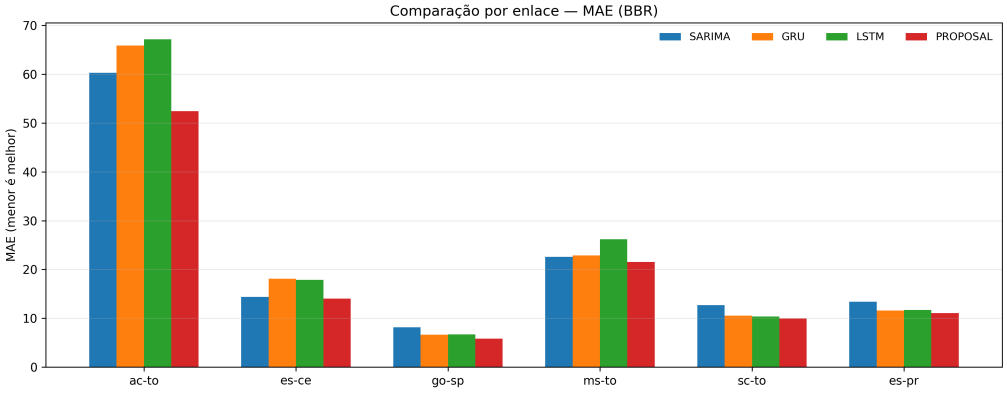

Figura 4. Comparação de MAE

Por fim, na figura 5, a métrica reforça a preservação de padrão temporal: a proposta é a melhor em todos os enlaces, com destaque para ac-to (0,35) e sc-to (0,26). Além disso, em enlaces onde alguns *baselines* apresentam correlação negativa (go-sp, ms-to, es-ce), o *framework* de predição proposto eleva a correlação para valores positivos; já em es-pr todos permanecem negativos.Entretanto, observa-se uma redução na negatividade, o que sugere que este enlace é estruturalmente mais difícil de capturar em termos de forma, mesmo com a melhoria nos erros absolutos.

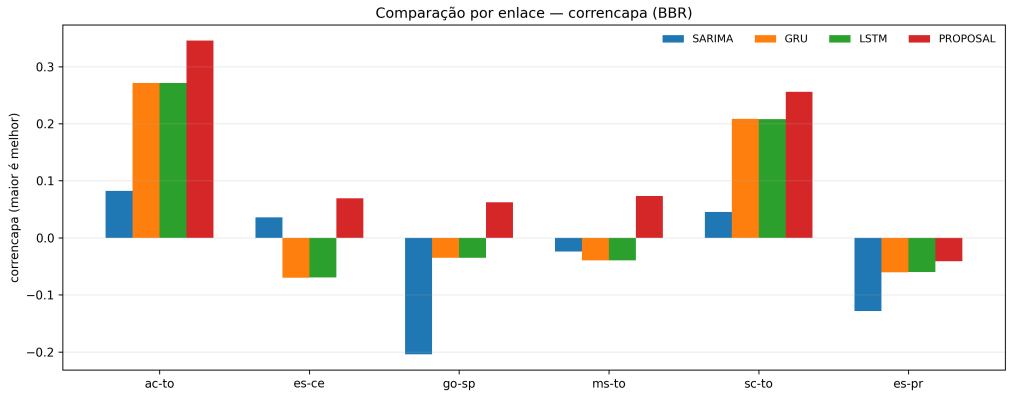

Figura 5. Kappa de Cohen

5.2. Análise de Degradação de Horizonte

Um dos maiores desafios em predição recursiva é a propagação de erros: ao usar uma predição para gerar a próxima, o modelo corre o risco de divergir rapidamente da realidade. A Figura 6 apresenta a análise de degradação do RMSE médio sobre um horizonte de 30 horas, comparando a proposta com as arquiteturas GRU e LSTM.

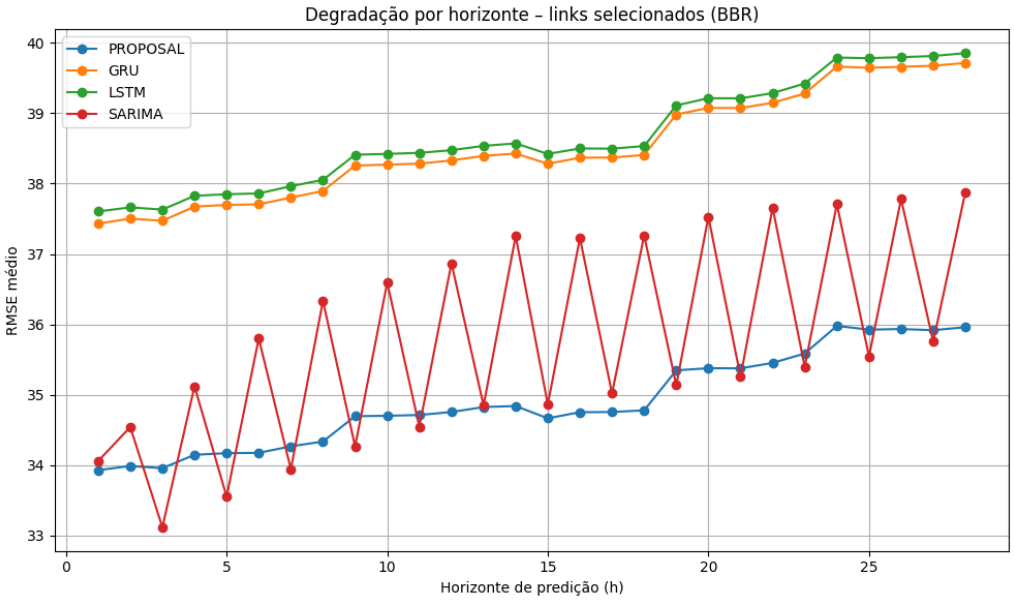

Figura 6. Degradação do RMSE médio em função do horizonte de predição para os enlaces selecionados.

Os resultados mostram que a proposta mantém uma estabilidade significativamente superior ao longo do tempo. Enquanto o erro do GRU e do LSTM cresce de forma acentuada após as primeiras 15 horas, a proposta exibe uma curva de degradação mais suave. Isso é quantificado pelos valores de inclinação (*slope*) do RMSE na 2

A menor inclinação da proposta (0.0791) indica que o mecanismo de imputação implícita, combinado com a Perda Huber, atua como um regularizador eficaz. Ao suavizar o impacto de erros pontuais no *loop* autorregressivo, o modelo consegue manter a

utilidade da predição para fins de planejamento de capacidade por períodos muito mais longos do que abordagens convencionais.

O RMSE inicial da proposta (33.9) já é consideravelmente inferior ao das demais arquiteturas, e o fato de o erro final após 28 horas (36.0) ainda ser menor que o erro inicial do GRU e do LSTM demonstra a robustez da técnica para cenários operacionais de longa duração.

Tabela 2. Degradação do RMSE médio do horizonte curto (h=1) ao horizonte final (h=H).

Modelo	$\mathbf{RMSE_{h=1}}$	$\mathbf{RMSE_{h=H}}$	$\mathbf{\Delta RMSE}$	**Degradação Relativa**
GRU	37.431279	39.714563	2.283284	0.060999
LSTM	37.604484	39.854778	2.250294	0.059841
SARIMA	34.056727	37.879669	3.822942	0.112252
Proposta	33.926224	35.959137	2.032913	0.059922

5.3. Discussão Final

Os resultados demonstram que a solução proposta supera os modelos tradicionais e de aprendizado de máquina (baselines) em cenários de alta taxa de dados faltantes. Ao avaliar o desempenho nos casos mais críticos, a proposta reduziu o erro percentual (SMAPE) para 7,8%, enquanto modelos como SARIMA, GRU e LSTM apresentaram erros de quase 10%. Além da precisão absoluta, o uso do Kappa de Cohen confirmou que a solução preserva melhor o padrão temporal e a forma das oscilações de tráfego, mesmo quando a magnitude exata não é atingida, superando as correlações negativas observadas nos concorrentes em determinados enlaces.

Desta forma, a proposta habilita a integração direta do tratamento de falhas ao processo de inferência por meio de um mecanismo de realimentação recursiva. Diferente de abordagens estáticas que exigem uma etapa prévia e onerosa de imputação de dados, esta abordagem utiliza as próprias predições do modelo para sustentar o estado da janela na ausência de medições reais. Essa imputação implícita atua como um regularizador, permitindo que o sistema mantenha a operabilidade mesmo em séries temporais univariadas com taxas de omissão superiores a 50%, sem introduzir a complexidade computacional de redes generativas como GANs.

Em termos de aplicabilidade industrial, a solução é altamente viável para provedores de Internet que operam infraestruturas de grande escala sujeitas a congestionamentos e falhas de telemetria. A solução permite uma gestão proativa de recursos e planejamento de capacidade, facilitando a identificação preventiva de gargalos mesmo quando os mecanismos de coleta de dados são intermitentes. Por ser um pipeline simplificado e adaptável a dados reais, a tecnologia possui escalabilidade para ser implementada em sistemas de monitoramento em tempo real que exigem resiliência operacional e baixo custo computacional.

6. Conclusão

Este trabalho apresentou um método de predição autorregressiva de vazão que lida com dados faltantes durante a inferência, usando janelas deslizantes e uma regra simples de

realimentação: quando há amostra observada, ela é usada, e quando há lacuna, o modelo utiliza a própria predição para manter a continuidade da janela. Nos enlaces avaliados com tráfego real (BBR), a proposta obteve os menores valores de SMAPE, RMSE e MAE na comparação com SARIMA, GRU e LSTM, bem como apresentou os melhores resultados em relação ao Kappa de Cohen, indicando melhor preservação do padrão temporal.

Na análise de degradação de horizonte, a proposta manteve crescimento de erro mais controlado ao longo de 30 horas, com $RMSE_{h=1} = 33.93$ e $RMSE_{h=H} = 35.96$, resultando em $\Delta RMSE = 2.03$ e degradação relativa de 0.0599, competitiva com GRU/LSTM e superior ao SARIMA. Como limitações, os experimentos ainda cobrem um conjunto restrito de enlaces e a avaliação não isola completamente o efeito de diferentes padrões de missingness. Como trabalhos futuros, pretende-se ampliar a quantidade de enlaces e períodos analisados e estudar uma regra explícita de decaimento/confiança para valores sintetizados em lacunas longas.

Agradecimentos

Esta pesquisa é parte do INCT de Redes de Comunicação e Internet das Coisas Inteligentes (ICoNIoT), financiado por CNPq (proc. 405940/2022-0), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88887.954253/2024-00. Adicionalmente, os autores gostariam de agradecer ao CNPq (N^o 305946/2025-0) e CAPES (N^o 88887.972043/2024-00) pelo apoio financeiro.

Referências

Ahn, H., Sun, K., and Kim, K. (2021). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials and Continua*, 70:767–779.

Al-Thaedan, A., Shakir, Z., Mjhool, A. Y., Alsabah, R., Al-Sabbagh, A., Salah, M., and Zec, J. (2023). Downlink throughput prediction using machine learning models on 4g-lte networks. *International Journal of Information Technology*, 15(6):2987–2993.

Brito, M. L. L., Ferreira, M. C. M., Portela, A. L. C., and Gomes, R. L. (2026). Ai-based estimation of bandwidth availability for data offloading in edge-cloud computing. *IEEE Networking Letters*, 8:69–73.

Du, W., Wang, J., Qian, L., Yang, Y., Ibrahim, Z., Liu, F., Wang, Z., Liu, H., Zhao, Z., Zhou, Y., Wang, W., Ding, K., Liang, Y., Prakash, B. A., and Wen, Q. (2024). Tsibench: Benchmarking time series imputation.

Ferreira, M., Linhares, M., Araújo, T., and Gomes, R. (2025). Aplicando decomposição de valores singulares na predição de vazão de rede. In *Anais do XLIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 630–643, Porto Alegre, RS, Brasil. SBC.

Ferreira, M. C., Ribeiro, S. E., Nobre, F. V., Linhares, M. L., Araújo, T. P., and Gomes, R. L. (2024). Mitigating measurement failures in throughput performance forecasting. In *2024 20th International Conference on Network and Service Management (CNSM)*, pages 1–7.

Gomes, R. L., Bittencourt, L. F., and Madeira, E. R. (2014). A bandwidth-feasibility algorithm for reliable virtual network allocation. In *2014 IEEE 28th International Conference on Advanced Information Networking and Applications*, pages 504–511.

Gomes, R. L., Bittencourt, L. F., Madeira, E. R., Cerqueira, E., and Gerla, M. (2016). Bandwidth-aware allocation of resilient virtual software defined networks. *Computer Networks*, 100:179–194.

Gomes, R. L., Bittencourt, L. F., and Madeira, E. R. M. (2013). A framework for sla establishment of virtual networks based on qos classes. In *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, pages 1175–1178.

Hu, H., Qian, S., Yang, D., Cao, J., and Xue, G. (2024). Iterative time series imputation by maintaining dependency consistency. *ACM Trans. Knowl. Discov. Data*, 19(1).

Kablaoui, R., Ahmad, I., Abed, S., and Awad, M. (2024). Network traffic prediction by learning time series as images. *Engineering Science and Technology, an International Journal*, 55:101754.

Li, X. (2024). Time series forecasting with missing data using generative adversarial networks and bayesian inference. *Information*, 15(4).

Lopes Gomes, R. and Roberto Mauro Madeira, E. (2012). A traffic classification agent for virtual networks based on qos classes. *IEEE Latin America Transactions*, 10(3):1734–1741.

Mutter, E. and Shannigrahi, S. (2024). Science dmz networks: How different are they really? In *2024 IEEE 49th Conference on Local Computer Networks (LCN)*, page 1–9. IEEE.

Na, H., Shin, Y., Lee, D., and Lee, J. (2023). Lstm-based throughput prediction for lte networks. *ICT Express*, 9(2):247–252.

Neog, A., Daw, A., Khorasgani, S. F., Sawhney, M., Pradhan, A., Lofton, M. E., McAfee, B. J., Breef-Pilz, A., Wander, H. L., Howard, D. W., Carey, C. C., Hanson, P., and Karpatne, A. (2026). Investigating a model-agnostic and imputation-free approach for irregularly-sampled multivariate time-series modeling.

Nobre, F. V. J., Silva, D. d. S., Ferreira, M. C. M. M., Brito, M. L. M. L., de Araújo, T. P., and Gomes, R. L. (2025). Time-weighted correlation approach to identify high delay links in internet service providers. *Journal of Internet Services and Applications*, 16(1):419–430.

Pimenta, I., Silva, D., Moura, E., Silveira, M., and Gomes, R. L. (2024). Impact of data anonymization in machine learning models. In *Proceedings of the 13th Latin-American Symposium on Dependable and Secure Computing*, pages 188–191.

Portela, A., Linhares, M. M., Nobre, F. V. J., Menezes, R., Mesquita, M., and Gomes, R. L. (2024). The role of tcp congestion control in the throughput forecasting. In *Proceedings of the 13th Latin-American Symposium on Dependable and Secure Computing*, pages 196–199.

Qian, L., Yang, Y., Du, W., Wang, J., Dobsoni, R., and Ibrahim, Z. (2025). Beyond random missingness: Clinically rethinking for healthcare time series imputation.

Yalda, K., Jamal Hamad, D., Tapus, N., and Okumus, I. T. (2024). Network traffic prediction performance using lstm. *Romanian Journal of Information Science and Technology*, 27:336–347.