Aplicando Redes Neurais e Análise Temporal para Predição Adaptativa de Desempenho de Rede

Silvio E. S. B. Ribeiro[1], **Rafael A. Menezes**[1], **Ariel L. C. Portela**[1], **Thelmo P. Araújo**[1], **Rafael L. Gomes**[1]

[1]Universidade Estadual do Ceará (UECE), Fortaleza, Ceará, Brasil.

{silvio.eduardo,rafael.menezes,ariel.portela}@aluno.uece.br

{thelmo.araujo,rafa.lopes}@uece.br

Abstract. *Network monitoring services are performed by several companies and Internet Service Providers (ISPs), which provide results of regular performance tests, such as throughput, loss, and delay, among others. These measurements support knowing the behavior of the network, as well as obtaining information for strategic planning. However, these tools still need to evolve in order to encompass more complex activities, such as performance prediction, especially within the current context of elastic demand. Within this context, this paper presents an adaptive network performance prediction model based on Neural Networks and Time Series Analysis, enabling the identification of future network performance in specific periods, according to past network measurements. The experiments carried out, using real data from the National Education and Research Network (RNP), show that the proposed model reaches high levels of accuracy in prediction and overcomes the existing prediction models.*

Resumo. *Serviços de monitoramento de rede são executados por diversas empresas e Provedores de Internet (ISPs), que fornecem resultados de testes regulares de desempenho, tais como vazão, perda, atraso, dentre outros. Estas medições auxiliam a conhecer o comportamento da rede, bem como obter informações para um planejamento estratégico. Contudo, estas ferramentas ainda precisam evoluir a fim de englobar atividades mais complexas, tal como predição do desempenho, principalmente dentro do contexto atual de demanda elástica. Dentro deste contexto, este artigo apresenta um modelo adaptativo de predição de desempenho de rede baseado em Redes Neurais e Análise de Séries Temporais, habilitando a identificação do desempenho futuro da rede em determinados períodos, de acordo com medições de rede passadas. Os experimentos realizados, usando dados reais da Rede Nacional de Ensino e Pesquisa (RNP), mostram que o modelo proposto consegue atingir altos níveis de acurácia na predição, bem como supera o uso dos modelos de predição existentes.*

1. Introdução

No contexto das Redes de Computadores e Telecomunicações, é importante haver uma análise mais detalhada dos diversos tipos de dados e métricas de desempenho que as infraestruturas de rede podem obter através de serviços de monitoramento, tais como vazão, perda, atraso, dentre outras [Mok et al. 2021]. Em particular, existe a análise por demanda de recursos de rede, que é uma forma de medir o perfil de consumo de um cliente. Um

cliente, sob o ponto de vista do provedor, pode ser uma conexão residencial, pública, empresarial, etc. Conhecer o seu perfil é muito útil para realizar uma melhor alocação de recursos, de acordo com a sua necessidade, levando em consideração a elasticidade da demanda em função do tempo.

Estas medições auxiliam as empresas e ISPs a conhecerem o comportamento da rede, bem como obterem informações para um planejamento estratégico, incluindo locais de expansão de capacidade de enlace, agendamento de manutenção e outras ações de gerenciamento. Contudo, estas ferramentas ainda precisam evoluir em diversos aspectos, e um dos pontos a serem evoluídos, em relação aos serviços de monitoramento, é a falta de integração com funcionalidades mais complexas, tal como a predição do desempenho da rede [Lara-Benítez et al. 2021]. A predição ocorre através de técnicas que permitem a compreensão do comportamento, por meio de observações anteriores, e a previsão de valores futuros [Hewamalage et al. 2021], permitindo a aplicação de tarefas proativas para evitar problemas como: lentidão, interrupção do serviço, desconexões constantes e planejar a infraestrutura de rede.

Contudo, as técnicas de predição existentes encontram dificuldade em lidar com o cenário de demanda elástica por recursos de rede. A demanda elástica ao longo do dia ocorre devido à mobilidade humana nas cidades, que modifica a forma de acesso à Internet dos usuários finais de acordo com o local de acesso [Gomes et al. 2020]. Esse comportamento elástico, quando não tratado adequadamente, pode causar problemas de lentidão, interrupção do serviço e desconexões constantes, resultando em usuários frustrados e quebra dos requisitos do Acordo de Nível de Serviço (SLA) [Tomic et al. 2022]. Consequentemente, há um impacto direto do contexto de demanda elástica nas métricas de Qualidade de Serviço (QoS).

Dentro deste contexto, este artigo apresenta o Predição de Desempenho de Rede (PDR), um modelo adaptativo de predição de desempenho de rede, baseado em Redes Neurais e Análise de Séries Temporais, o qual habilita a identificação do desempenho futuro da rede em determinados períodos, de acordo com medições passadas. O PDR ajusta a série temporal das medições realizadas através de um conjunto de análises estatísticas, tais como decomposição e interpolação, criando uma série temporal padronizada que possibilita o treinamento de um modelo de predição mais eficaz. Essa série temporal tratada é usada como insumo para diversas redes neurais que irão realizar o processo de predição de uma nova série com o desempenho de rede esperado.

A fim de validar a solução e analisar seu desempenho no que se refere a capacidade de predição, foram realizados experimentos usando dados reais do Serviço de monitoramento da rede Ipê (Monipê)[1] da Rede Nacional de Ensino e Pesquisa (RNP). Os resultados obtidos mostram que o modelo proposto consegue atingir altos níveis de acurácia na predição, bem como supera o uso dos modelos de predição existentes em cerca de 25% no que se refere a acurácia e outras métricas de avaliação.

O restante deste artigo está organizado da seguinte forma: A Seção 2 apresenta os trabalhos relacionados existentes. A Seção 3 apresenta o modelo de predição de desempenho proposto, enquanto a Seção 4 descreve os resultados dos experimentos realizados. Por fim, a Seção 5 conclui o artigo e apresenta trabalhos futuros.

[1] monipe-central.rnp.br

2. Trabalhos Relacionados

Esta seção descreve os principais trabalhos relacionados, e recentemente publicados pela comunidade científica, sobre predição no contexto de redes de computadores, incluindo aspectos de desempenho e qualidade de serviço.

Syu et al. [Syu et al. 2019] descrevem a aplicação de programação genética (GP) para atributos dinâmicos de QoS com reconhecimento de tempo para serviços da Web em ambientes de nuvem. Os autores realizaram vários experimentos (estudos e análises empíricas), atestando que as abordagens de aprendizado de máquina são geralmente superiores aos métodos estatísticos de séries temporais amplamente utilizados em termos de precisão de modelagem e precisão de predição. No entanto, os autores estudam especificamente o contexto dos atributos dinâmicos de QoS dos serviços em nuvem, que é diferente da demanda elástica de recursos.

Baig et al. [Baig et al. 2019] propõem um método para identificar, de forma adaptativa e automática, o modelo mais adequado para estimar, com precisão, a utilização dos recursos do *data center*. Os autores treinam um classificador com base no histórico do uso de recursos para decidir o modelo de predição a ser usado para determinadas observações. Estas observações são coletadas durante um intervalo de tempo específico. Apesar da abordagem de predição aplicada, esta proposta é baseada em cenários de *data center*, que possuem um comportamento de rede distinto quando comparado ao desempenho de rede, principalmente quando a demanda elástica por recursos é considerada.

Katris et al. [Katris and Daskalaki 2019] desenvolveram modelos de séries temporais no tráfego de vídeo como parte de três esquemas de alocação dinâmica de largura de banda. Os autores tentam melhorar a precisão das previsões de tráfego de vídeo usando a Média Móvel AutoRegressiva Integrada Fracionada (FARIMA) e a Heterocedasticidade Condicional Autoregressiva Generalizada (GARCH) com redes neurais. As previsões de tráfego aplicam um método de seleção de não linearidade a ser combinado com o modelo. O uso do modelo de predição implementado é específico para o tráfego de vídeo, uma vez que considera informações de *buffer* e taxa de utilização. Desta forma, a sua aplicação no contexto de medição de desempenho é limitada.

Yang et al. [Yang et al. 2021] apresentam um método de predição de tráfego de rede baseado no ARIMA (*Autoregressive Integrated Moving Average model*)-BPNN (*Back Propagation Neural Network*) a fim de aumentar a precisão para predição de tráfego futuro. O ARIMA-BPNN realiza uma avaliação do potencial de mineração de comportamentos lineares e não lineares dos dados históricos de tráfego de rede, melhorando assim a precisão da predição. No entanto, o método proposto não considera a predição do desempenho da rede, somente do tráfego em si.

Sone et al. [Sone et al. 2020] mostram uma análise temporal e espacial do tráfego de rede usando dados reais de uma rede corporativa (compreendendo quase quinhentos pontos de acesso), concluindo que o grupo de pontos de acesso com alta utilização de tráfego possui fortes padrões de sazonalidade (autocorrelações e correlações cruzadas). Adicionalmente, os autores estudam a relação entre o número de usuários conectados e o tráfego gerado, e mostram que nem sempre usuários mais conectados significam mais dados de tráfego e vice-versa. Esses estudos foram baseados em *Holt-Winters*, *Seasonal Auto-Regressive Integrated Moving Average* (SARIMA), *Long Short-Term Memory*

(LSTM), *Gated Recurrent Unit* (GRU), and *Convolutional Neural Network* (CNN) para prever o tráfego de rede.

A partir do levantamento bibliográfico realizado, nota-se que nenhum artigo da literatura se concentrou no desenvolvimento de um modelo de predição de desempenho de rede que seja adaptável ao contexto de medição sendo realizada, que é o foco deste artigo. Nossa proposta realiza ações estatísticas para aprimorar o processo de predição, criando um modelo padronizado de amostras que facilita o processo de predição das técnicas.

3. Modelo de Predição de Desempenho de Rede (PDR)

Esta seção apresenta a descrição detalhada do modelo adaptativo de predição de desempenho de rede proposto neste artigo, chamado PDR. O PDR visa ajustar os dados de medição, através de técnicas de análise em séries temporais, para situações de demanda elástica, possibilitando assim uma maior eficiência das técnicas de predição existentes. O modelo de predição é, de forma geral, executado em três etapas: (i) Coleta de dados necessários para predição, (ii) Tratamento da série temporal (aplicando técnicas de decomposição e interpolação) e (iii) Aplicação das técnicas de predição. Uma visão geral da estruturação da ferramenta proposta neste projeto é apresentada na Figura 1.

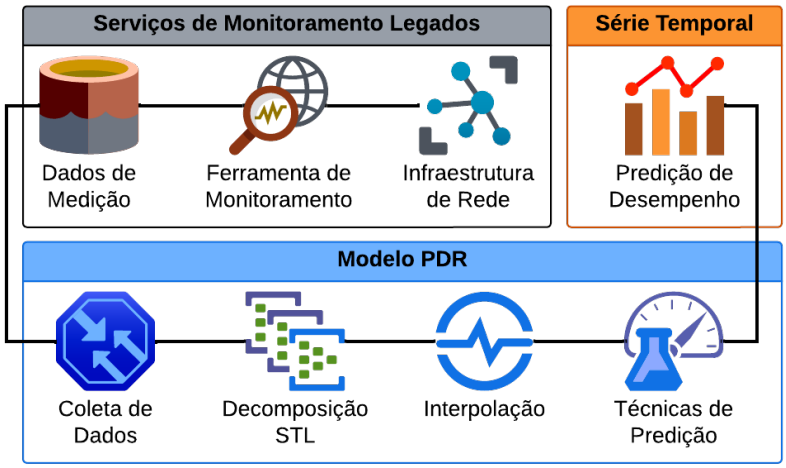


Figura 1. Visão Geral da Solução.

Primeiramente, a solução resgata os dados de medição oriundos do serviço de monitoramento existente (elementos em cinza da Figura 1), onde qualquer ferramenta pode ser considerada, tais como Perfsonar, SolarWinds, Auvik, etc. Posteriormente, em posse dos dados, aplica-se um modelo proposto sobre às medições de rede na etapa de tratamento da série (elementos em azul da Figura 1), com o objetivo de garantir um ajuste nos dados, ou seja, adequar a análise temporal ao contexto de demanda elástica e particularidades de cada contexto de comunicação. Por fim, o resultado do tratamento e análise da série temporal é usado para treinar um modelo de predição, ou seja, aplica-se alguma técnica de predição para gerar uma nova série temporal de predição em relação ao desempenho analisado.

3.1. Coleta dos Dados

No processo de predição, o passo inicial é a coleta de informações, que refere-se a um conjunto de dados que representam as medições entre dois pontos de comunicação. Ide-

almente, os serviços de monitoramento possuem séries temporais completas, formadas pelas medições periódicas realizadas com sucesso. Contudo, podem ocorrer falhas de medições que resultam em lacunas e, consequentemente, uma quebra da série temporal. Este problema ocorre principalmente em casos de medições de alta granularidade, que ocorrem diversas vezes por dia. Esses dados faltantes precisam ser preenchidos a fim de recompor a série temporal.

Para resolver este problema, foi desenvolvido um algoritmo que preenche os dados faltantes com a média de valores de cada ciclo da série (por exemplo, a cada dia), dessa forma, evitando lacunas na coleta e mantendo a integridade da série temporal. Portanto, durante a etapa de coleta de dados, o PDR analisa a série temporal original, identifica essas lacunas de medição e recompõe a série, habilitando esta a ser analisada e ajustada para posterior predição.

3.2. Análise e Tratamento da Série Temporal

Em geral, modelos tradicionais de predição precisam trabalhar em cima de processos estacionários para gerar bons resultados. Esse processo auxilia na estruturação de uma série temporal com condições favoráveis de análise, ao mesmo tempo que proporciona melhores predições, como será mostrado ao longo deste trabalho. Entretanto, alguns modelos mais robustos podem ter maior flexibilidade quanto a isso, ou seja, não havendo necessidade de um cenário muito restrito para gerar predições satisfatórias. Tais modelos, que utilizam Redes Neurais, por serem mais poderosos computacionalmente, têm uma tolerância maior a erros[Lara-Benítez et al. 2021].

Alguns padrões são encontrados em séries temporais [Hewamalage et al. 2021, Lara-Benítez et al. 2021]: (1) Tendência (T), essa componente é vista quando a série temporal segue alguma direção, crescente, decrescente e não exatamente linear; (2) Sazonalidade (S), padrão que se repete com uma certa frequência dentro de um período; e, (3) Ciclo (C), diferente da sazonalidade, este padrão se repete, mas não obedece uma periodicidade fixa. Z_t representa uma informação em série observada no período t, S_t é a componente sazonal do período t e E_t representa a parte não capturada do modelo no período t, chamada de erro (ou resíduo). Também é válido ressaltar que uma série temporal pode ser representada de forma aditiva ou multiplicativa, sendo Z_t composto pela combinação de componentes de tendência, sazonalidade e ciclo. Geralmente, o critério usado para definir uma série multiplicativa é se a magnitude dos dados cresce junto com a tendência, sendo $Z_t = T_t \times S_t \times E_t$. Caso a magnitude dos dados tenha uma tendência relativamente constante, a série deve ser modelada no formato aditivo $Z_t = T_t + S_t + E_t$.

No PDR, para a análise de séries temporais sazonais é utilizada a *Seasonal and Trend Decomposition using Loess* (STL), que é um método para destacar as componentes de sazonalidade e tendência, que utiliza o método *Loess* para estimar relações não-lineares entre os dados, o que permite uma melhor compreensão da série.

Primeiramente, são suavizados os valores da série temporal usando um filtro *low-pass*, que permite a passagem de ciclos de baixa frequência e atenua ciclos distantes da medida de corte, para remover o ruído de alta frequência e obter um componente de tendência suave. Em seguida, é subtraído o componente de tendência da série temporal original para obter uma série livre da tendência, para ser aplicado um filtro sazonal à esta mesma série sem tendência, obtendo-se uma estimativa de sazonalidade. Esta estimativa

é subtraída posteriormente para se obter uma série residual. A série temporal original pode então ser reconstruída adicionando os componentes de tendência, sazonalidade e residuos novamente, o que torna a decomposição STL particularmente útil para analisar e modelar dados com fortes padrões sazonais [Cleveland et al. 1990]. É um método flexível que pode lidar com uma ampla gama de períodos sazonais e pode ser aplicado a séries temporais aditivas e multiplicativas.

Posteriormente, para o tratamento de valores perdidos e *outliers*, é utilizada interpolação linear na série resultante [Costarelli et al. 2020]. O processo de interpolação linear pode ser usado para estimar o valor de uma série temporal em um ponto específico no tempo entre dois valores conhecidos. Por exemplo, obtém-se uma série temporal que representa a vazão entre dois pontos de comunicação durante um período, e deseja-se estimar o valor em um dia específico, que esteja entre dois dias consecutivos que possuam valores conhecidos. Para executar a interpolação linear neste caso, é preciso dois valores conhecidos pelos quais a série temporal passa em dois momentos específicos, (t_1, y_1) e (t_2, y_2). Também é necessário o tempo t no qual deseja estimar o valor da série temporal, que é chamado de ponto de interpolação [Blu et al. 2004]. Utilizando a fórmula de interpolação linear descrita na Equação 1, temos y sendo o valor estimado da série temporal (que se deseja obter) no ponto de interpolação t, e t_1 e t_2 sendo os tempos conhecidos.

$$y = y_1 + (y_2 - y_1) \times (t - t_1) \div (t_2 - t_1) \tag{1}$$

Por exemplo, diante da seguinte série temporal (Tempo t, Valor y) com três medições, sendo $(10, 100)$, $(12, 110)$ e $(14, 120)$, fosse preciso estimar o valor da série temporal no tempo 11, usa-se da interpolação linear com os valores de 10 e 12. O ponto de interpolação é 11 (ou seja, t) e os pontos conhecidos são (10/10, 100) e (12/10, 110). Usando a fórmula de interpolação linear, é possível estimar o valor no tempo 11 como $y = 100 + (110 - 100) \times (11 - 10) \div (12 - 10)$. Assim, o valor estimado da série temporal no tempo 11 seria 105, mostrando que é possível estimar valores com base nos valores conhecidos e com um grau de precisão coerente com todo o contexto da série temporal.

A Figura 2 ilustra um exemplo de serie temporal oriunda de uma medição de rede antes (em preto) e depois da análise temporal realizada pelo PDR (em azul). Percebe-se que o PDR consegue realizar uma amortização de ciclos irregulares, criando um comportamento mais padronizado em relação a série temporal como um todo. Portanto, após a aplicação do PDR, a série temporal a ser aplicada como entrada para o treinamento com as técnicas de predição está mais adequada e compreensível para as redes neurais.

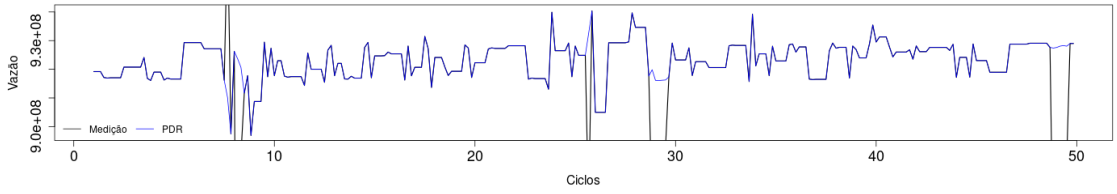


Figura 2. Ilustração dos Dados de Medição Antes e Depois do PDR.

3.3. Técnicas de Predição

No caso de predição de séries temporais, pode-se utilizar modelos mais tradicionais, tais como ARIMA, *Holt-Winters*, além da Rede Neural Auto-Regressiva (*Neural Network Autoregression* - NNAR), os quais foram aplicados durante as propostas anteriores [Oliveira et al. 2021, Menezes et al. 2021] com dados de uso de volume de tráfego. Além disso, é possível usar técnicas mais modernas que fazem uso de GPU, como por exemplo GRU e LSTM. Inicialmente foram avaliadas as técnicas tradicionais e uma técnica que utiliza redes neurais, mas apenas o modelo NNAR apresentou um desempenho satisfatório, dado o novo contexto deste trabalho. Portanto, durante os experimentos foram consideradas as técnicas NNAR, LSTM e GRU, as quais serão descritas a seguir.

3.3.1. Neural Network Autoregression (NNAR)

O modelo NNAR, também conhecido como modelo autorregressivo não linear, é um tipo de modelo de previsão de séries temporais baseado em redes neurais artificiais (ANNs). Uma das principais vantagens dos modelos NNAR é a sua capacidade de capturar relações não lineares entre os dados, o que os torna mais flexíveis e poderosos do que modelos lineares como o ARIMA [Maleki et al. 2018]. Eles também são capazes de lidar com várias entradas e dados ausentes, o que pode ser importante para algumas soluções. No entanto, os modelos NNAR podem ser mais difíceis de projetar e ajustar do que os modelos lineares, e podem ser computacionalmente custosos para treinar e executar. A rede é treinada para predição de uma etapa, sendo as predições de várias etapas calculadas recursivamente e para dados não sazonais, desta forma as previsões são obtidas pela combinação linear ou não linear das entradas.

3.3.2. Long Short-Term Memory (LSTM)

O modelo LSTM é um poderoso sistema neural recorrente especialmente projetado para superar os problemas de gradiente de explosão/desaparecimento que normalmente surgem ao aprender dependências de longo prazo, mesmo quando os intervalos de tempo mínimos são muito longos [Rafi et al. 2021]. No geral, isso pode ser evitado usando um *Constant Error Carousels* (CEC), que mantém o sinal de erro dentro da célula de cada unidade. Na verdade, tais células são elas próprias redes recorrentes com uma arquitetura melhorada para estender o CEC com recursos adicionais, ou seja, a porta de entrada e saída formam a célula de memória. As conexões auto-recorrentes indicam feedback com atraso de um intervalo de tempo. O LSTM é adequado para classificar, processar e prever séries temporais com intervalos de tempo de duração desconhecida. A insensibilidade relativa do comprimento da lacuna dá ao LSTM uma vantagem sobre os modelos de RNNs tradicionais, modelos ocultos de Markov (MOM) e outros métodos de aprendizado de sequência.

3.3.3. Gated Recurrent Unit (GRU)

A GRU é a nova geração de redes neurais recorrentes e é bastante semelhante a uma LSTM. As GRUs se livraram do estado da célula e usaram o estado oculto para trans-

ferir informações. Essa arquitetura possui apenas duas portas, uma de redefinição (*reset gate*) e uma de atualização (*update date*), que são utilizadas para resolver o problema da dissipação do gradiente de uma RNN padrão. Basicamente, elas são dois vetores que decidem quais informações devem ser passadas para a saída. O que há de especial é que elas podem ser treinadas para manter informações de muito tempo atrás, sem dissipá-las com o tempo ou remover informações irrelevantes para a predição. Enquanto as LSTMs possuem dois estados diferentes passados entre as células, o estado da célula e o estado oculto, que carregam a memória de longo e curto prazo, respectivamente, as GRUs têm apenas um estado oculto transferido entre as etapas do tempo. Esse estado oculto é capaz de manter as dependências de longo e curto prazo ao mesmo tempo, devido aos mecanismos de restrição e cálculos pelos quais o estado oculto e os dados de entrada passam.

4. Experimentos

Esta seção irá descrever a forma como os experimentos realizados foram configurados (Subseção 4.1) e discutir os resultados obtidos a partir desse conjunto de experimentos realizados, utilizando dados de medição de rede reais (Subseção 4.2).

4.1. Configuração dos Experimentos

A fim de realizar experimentos usando dados do mundo real, foram utilizados dados da RNP, através do Serviço de Monitoramento da Rede Ipê (MonIPÊ). O MonIPÊ usa o padrão de monitoramento internacional perfSONAR, onde as medições de Vazão ocorrem a cada 4 horas, enquanto as medições de Perda e Atraso ocorrem a cada 5 minutos. A partir deste cenário, foi decidido focar na vazão de rede, devido a maior representatividade do mesmo em relação a capacidade de transmissão das aplicações sobre a rede, e pela oportunidade de avaliar o comportamento de diferentes tipos de controle de congestionamento para TCP (*Cubic* e BBR) em relação ao processo de predição.

É válido ressaltar que a comunicação fim-a-fim entre os diversos PoPs da RNP possuem comportamentos distintos ao longo do dia e durante a semana, visto que a utilização da infraestrutura de rede da RNP segue um comportamento social dos usuários. Portanto, foram considerados quatro pontos de comunicação: Ceará (CE) - São Paulo (SP); Minas Gerais (MG) - Rio Grande do Sul (RS); Pará (PA) - Bahia (BA); E por fim, Paraná (PR) - Amazonas (AM). Estes pontos foram escolhidos devido: (i) Heterogeneidade em relação a localização geográfica, a qual influência no número de enlaces utilizados no caminho fim-a-fim e na carga total gerada sobre a infraestrutura de rede; (ii) Capacidade da infraestrutura de rede (enlaces variam entre 200 Gbps a 1 Gbps), este fato faz com que a comunicação originada de um determinado PoP tenha um comportamento muito diferente em relação a outros dois que possuam uma variação de capacidade fim-a-fim distintas; e, (iii) Por conterem os pontos de comunicação internacionais da RNP (nos estados do Ceará, São Paulo e Rio Grande do Sul), fazendo com que parte do tráfego de rede seja direcionado a esses pontos devido a comunicação com servidores internacionais na Europa, Estados Unidos, África e demais países da América Latina. Estes fatos podem ser observados na Figura 3.

Por fim, faz-se necessário compreender a diferença entre os TCPs BBR e *Cubic*, visto que ambos possuem comportamentos diferentes [Miyazawa et al. 2020, Kanaya et al. 2020]. O TCP-BBR ao invés de responder diretamente a perda de paco-

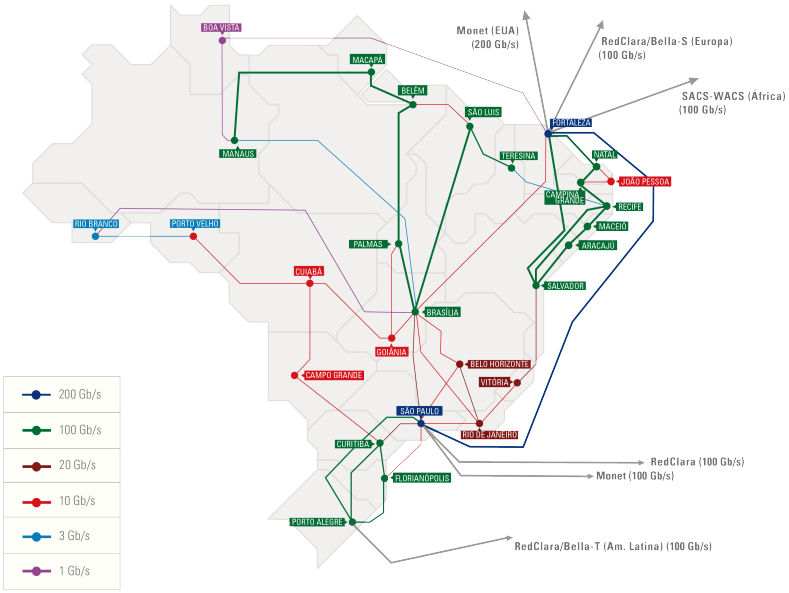


Figura 3. Rede Ipê.

tes, ele também leva em consideração largura de banda, RTT e outras métricas, calculando o "*pacing gain*" para definir os valores entre envio de pacotes e a janela de congestionamento. Esta abordagem do BBR traz maior vazão e mais equilíbrio ao algoritmo [Cardwell et al. 2017]. Por outro lado, o TCP *Cubic* utiliza uma função cúbica no tempo para o aumento da janela de congestionamento em tempo real, frente a perda de pacotes, fazendo assim com que o *Cubic* responda com mais qualidade à perda de pacotes, sem limitar consideravelmente o crescimento da janela quando não acontece a perda [Ha et al. 2008].

4.1.1. Treinamento do Modelo de Predição

Para cada série gerada, separamos cerca de 70% da série para treinamento e 30% para teste, ou seja, 105 ciclos e 45 ciclos, respectivamente. As séries foram submetidas a dois fluxos diferentes. No primeiro, realizamos o processo clássico de predição, onde treinamos com as três redes neurais citadas anteriormente e fazemos a predição de 45 ciclos no futuro. Já no segundo, usando as mesmas redes neurais do primeiro fluxo, mas antes aplicamos o modelo de predição proposto, com destaque para o tratamento de *outliers*. O tratamento desses valores discrepantes foi feito utilizando interpolação linear e a Decomposição STL.

4.1.2. Definição de Faixas de Desempenho

Usualmente as empresas e ISPs definem faixas de desempenho para classificar as medições, visando suavizar as variações de desempenho durante o processo de avaliação, bem como destacar os limiares que são definidos a partir dos requisitos dos serviços que utilizam a rede [Bayne et al. 2017]. Assim, definimos cinco faixas de valores de vazão

de rede, baseados na organização usada no MonIpe da RNP[2]: *Red*, para valores menores que 0.2 Gbit/s; *Orange*, para valores entre [0.2; 0.5] Gbit/s; *Yellow*, para valores entre [0.5; 0.8] Gbit/s; *Blue*, para valores entre [0.8; 1.0] Gbit/s; E por fim, *Green*, para valores maiores ou iguais a 1.0 Gbit/s. Com essas faixas criadas, adotamos o seguinte critério: As predições acertam, se o valor da predição e o valor de teste respectivo estão dentro da mesma faixa, e erram, caso contrário.

4.1.3. Métricas de Avaliação

Com as predições e definição das faixas, obtemos a Acurácia para cada um dos modelos de predição (sem e com PDR), para cada ponto de comunicação e cada tipo de TCP mencionado anteriormente. Sendo assim, temos os casos de $PDR + NNAR$, $PDR + LSTM$ e $PDR + GRU$ com o uso da proposta deste artigo e $NNAR$, $LSTM$ e GRU usando apenas os modelos de predição tradicionais existentes. Além disso, usaremos a Raiz do Erro Quadrático Médio (RMSE) por um período T, seguindo a Equação 2, onde $\hat{y}_t$ é o valor estimado e y_t é o valor real do uso da medição no momento t.

$$\mathrm{RMSE}(T) = \frac{1}{\sqrt{T}} \left(\sum_{t=1}^{T} (\hat{y}_t - y_t)^2 \right)^{\frac{1}{2}} \tag{2}$$

4.2. Resultados

Esta seção apresenta os resultados obtidos a partir da análise dos experimentos realizados com um conjunto de dados reais. Os resultados do RMSE são apresentados nas Figuras 4 e 5, enquanto que os dados de Acurácia são mostrados nas Figuras 6 e 7.

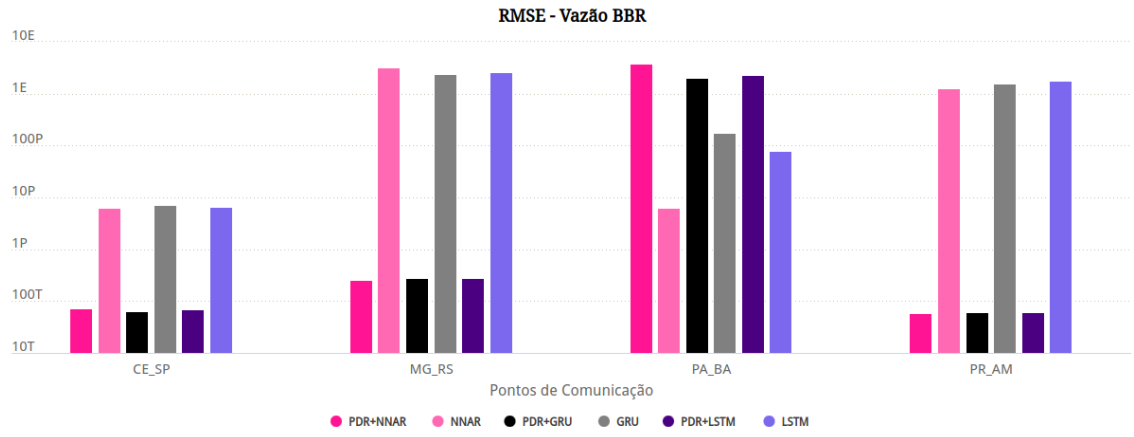


Figura 4. Resultados do RMSE para as Predições de Vazão com TCP BBR.

Analisando as Figuras 4 e 5 é possível notar que, de forma geral, o erro das predições em que o PDR foi aplicado é menor quando comparado com o método clássico. A única exceção, é para os pontos PA-BA, onde todos os modelos de rede neural adicionadas do modelo PDR, obtiveram um RMSE maior. Isso ocorreu devido a estes pontos em específico, ainda possuírem muitos *outliers* em relação a tendência da série temporal,

[2]monipe-central.rnp.br

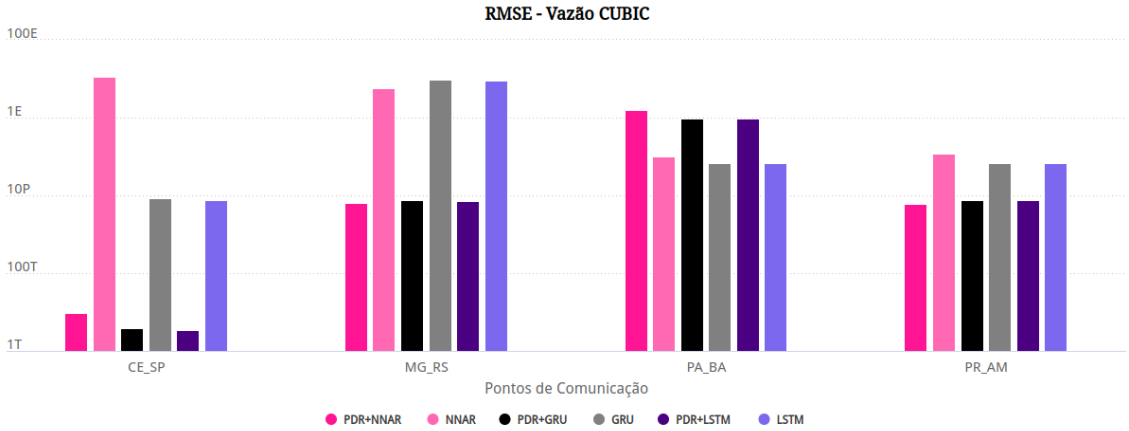


Figura 5. Resultados do RMSE para as Predições de Vazão com TCP Cubic

mesmo após a aplicação de técnicas para suavização, causando distorções no processo de predição e, consequentemente, impactando os resultados finais.

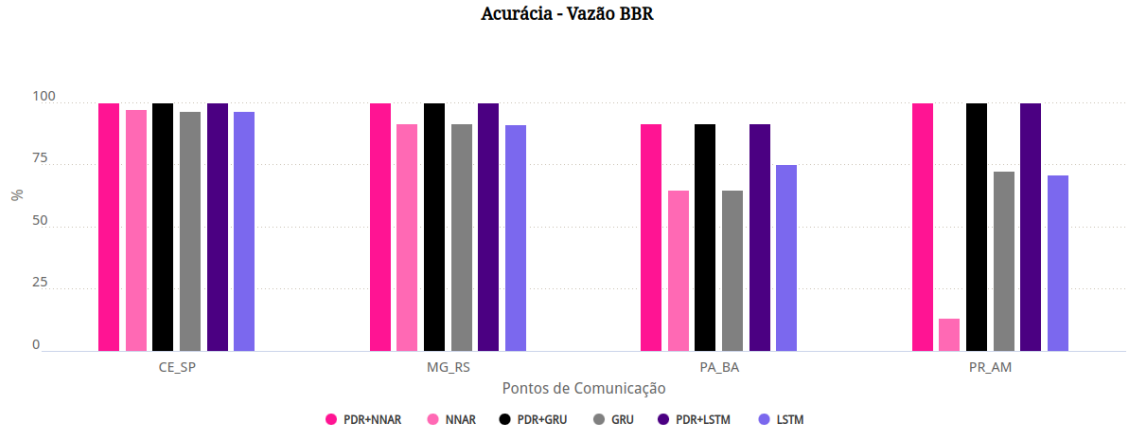


Figura 6. Resultados de Acurácia para as Predições de Vazão com TCP BBR.

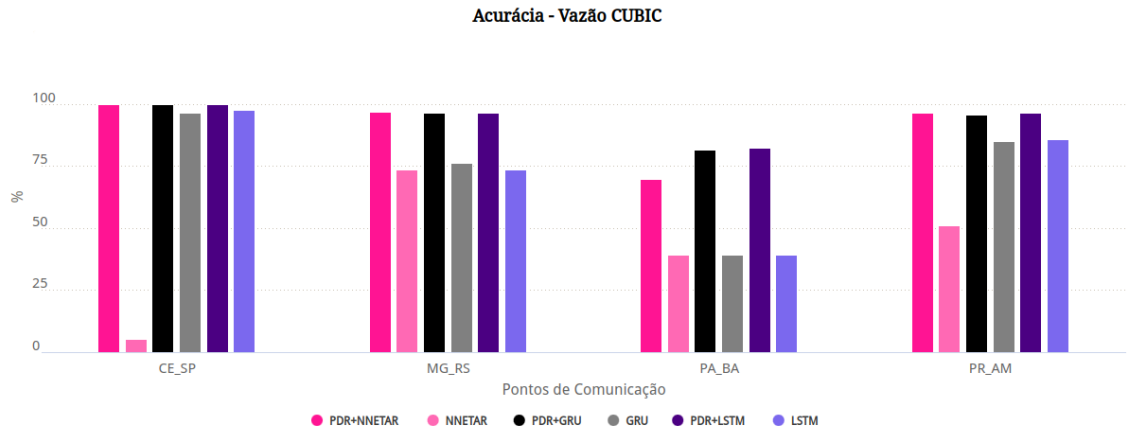


Figura 7. Resultados de Acurácia para as Predições de Vazão com TCP Cubic.

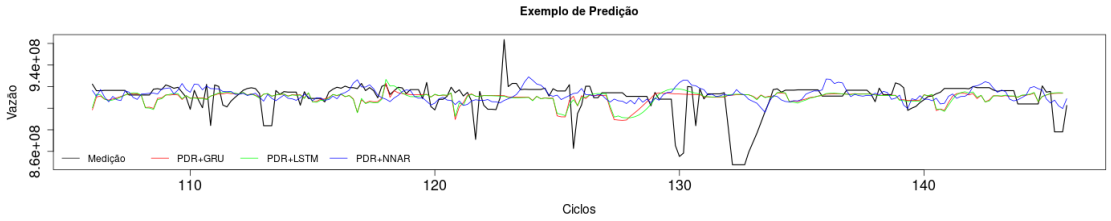
Percebe-se que para todos os pontos de comunicação, na Figura 6 e 7 a acurácia melhorou com a aplicação do modelo PDR. Para o BBR, a acurácia média com PDR foi

de 97.91%, mas para o processo tradicional, foi de 79.81%. Já para o *Cubic*, utilizando o PDR, a acurácia média foi de 92.85%, enquanto que para os procedimentos usuais, foi de apenas 61.37%. De forma geral, o PDR teve uma acurácia média de 95.38% e o fluxo usual, obteve uma média de 70.59%. Portanto, podemos perceber que o PDR melhorou a acurácia em aproximadamente 25%.

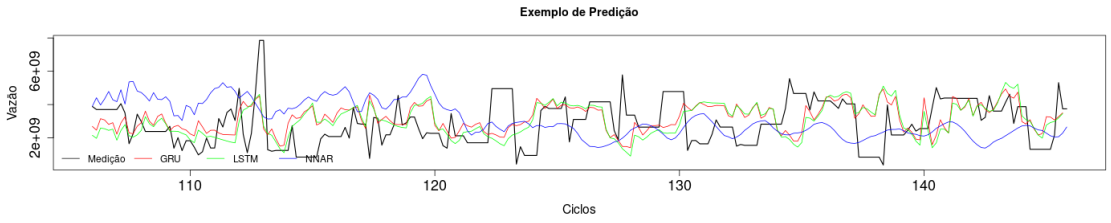
Importante ressaltar também que tanto para o TCP Cubic quanto BBR os modelos de predição conseguiram atingir resultados satisfatórios de acurácia e RMSE, o que pode indicar que os algoritmos de controle de congestionamento aplicado aos protocolos tem pouco impacto no processo de predição.

Apesar do uso das faixas de vazão para compreender melhor o comportamento dos dados, ao comparamos algumas observações com o resultado produzido pelos preditores, observa-se um bom ajuste aos dados originais, indicando que, mesmo sem utilizar o critério das faixas de vazão, as predições produzidas utilizando o PDR se mostram visualmente satisfatórias.

A fim de prover um feedback visual, a Figura 8 ilustra alguns casos de predição em relação às medições, considerando tanto o uso e não uso do PDR para as três técnicas de predição analisadas nestes artigo. É válido ressaltar que a proporção utilizada no eixo y é baixa a fim de mostrar os dados com menor granularidade, ou seja, pegou-se um intervalo entre 0.8 e 0.9 Gbit/s ao invés de usar o intervalo total (entre 0 e 1 Gbit/s).



(a) Com o PDR.



(b) Sem o PDR.

Figura 8. Modelos de Predição Comparados As Medições.

A partir da análise experimental realizada neste trabalho, concluímos que o modelo PDR traz uma grande vantagem para o contexto de predição sobre desempenho de rede. Também é possível notar que apesar de sua eficiência, o passo da remoção de erros de ciclo pode ser revisado, com o objetivo maximizar sua eficiência sobre conjuntos de dados com muitos *outliers*.

5. Conclusão

A medição do desempenho de rede é um iniciativa para obter informações para um planejamento estratégico das empresas e ISPs, onde a predição deste desempenho é uma funcionalidade crucial para garantir uma execução dos serviços que executam sobre a rede e da experiência dos usuários neste uso. Todavia, o processo de predição de desempenho é complexo, ainda mais diante da realidade atual de demanda elástica. A fim de lidar com esta situação, este artigo apresentou o PDR, um modelo adaptativo de predição de desempenho de rede baseado em redes neurais e análise de séries temporais. Os experimentos realizados, usando dados da RNP, mostram que o modelo proposto consegue atingir altos níveis de acurácia na predição, bem como supera o uso dos modelos de predição existentes em cerca de 25%. Como trabalhos futuros, pretende-se verificar a eficiência do modelo trazido, sobre dados de outros tipos de medição de rede, como atraso e perda. Adicionalmente, é importante verificar o comportamento de outros modelos de aprendizado de máquina combinados com o PDR.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) do Brasil (N^o 303877/2021-9) e Rede Nacional de Pesquisa (RNP) através do Programa de Monitoramento de Redes (PMON 2022) pelo apoio financeiro.

Referências

Baig, S., Iqbal, W., Berral, J. L., Erradi, A., and Carrera, D. (2019). Adaptive prediction models for data center resources utilization estimation. *IEEE Transactions on Network and Service Management*, 16(4):1681–1693.

Bayne, L., Schepis, D., and Purchase, S. (2017). A framework for understanding strategic network performance: Exploring efficiency and effectiveness at the network level. *Industrial Marketing Management*, 67:134–147.

Blu, T., Thévenaz, P., and Unser, M. (2004). Linear interpolation revitalized. *IEEE Transactions on Image Processing*, 13(5):710–719.

Cardwell, N., Cheng, Y., Gunn, C. S., Yeganeh, S. H., and Jacobson, V. (2017). Bbr: congestion-based congestion control. *Communications of the ACM*, 60(2):58–66.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition. *J. Off. Stat*, 6(1):3–73.

Costarelli, D., Seracini, M., and Vinti, G. (2020). A comparison between the sampling kantorovich algorithm for digital image processing with some interpolation and quasi-interpolation methods. *Applied Mathematics and Computation*, 374:125046.

Gomes, R. L., Bittencourt, L. F., and Madeira, E. R. M. (2020). Reliability-aware network slicing in elastic demand scenarios. *IEEE Communications Magazine*, 58(10):29–34.

Ha, S., Rhee, I., and Xu, L. (2008). Cubic: a new tcp-friendly high-speed tcp variant. *ACM SIGOPS operating systems review*, 42(5):64–74.

Hewamalage, H., Bergmeir, C., and Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427.

Kanaya, T., Tabata, N., and Yamaguchi, S. (2020). A study on performance of cubic tcp and tcp bbr in 5g environment. In *2020 IEEE 3rd 5G World Forum (5GWF)*, pages 508–513.

Katris, C. and Daskalaki, S. (2019). Dynamic bandwidth allocation for video traffic using farima-based forecasting models. *Journal of Network and Systems Management*, 27(1):39–65.

Lara-Benítez, P., Carranza-García, M., and Riquelme, J. C. (2021). An experimental review on deep learning architectures for time series forecasting. *International Journal of Neural Systems*, 31(03):2130001. PMID: 33588711.

Maleki, A., Nasseri, S., Aminabad, M. S., and Hadi, M. (2018). Comparison of ARIMA and NNAR Models for Forecasting Water Treatment Plant's Influent Characteristics. *KSCE Journal of Civil Engineering*, 22(9):3233–3245.

Menezes, R., Oliveira, D., and Gomes, R. (2021). O impacto da pandemia do covid-19 no comportamento do tráfego de rede e no processo de predição. In *Anais do XII Workshop de Pesquisa Experimental da Internet do Futuro*, pages 25–30, Porto Alegre, RS, Brasil. SBC.

Miyazawa, K., Yamaguchi, S., and Kobayashi, A. (2020). Mechanism of cyclic performance fluctuation of tcp bbr and cubic tcp communications. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1139–1144.

Mok, R. K. P., Zou, H., Yang, R., Koch, T., Katz-Bassett, E., and Claffy, K. C. (2021). Measuring the network performance of google cloud platform. In *Proceedings of the 21st ACM Internet Measurement Conference*, IMC '21, page 54–61, New York, NY, USA. Association for Computing Machinery.

Oliveira, D. H. L., de Araujo, T. P., and Gomes, R. L. (2021). An adaptive forecasting model for slice allocation in softwarized networks. *IEEE Transactions on Network and Service Management*, 18(1):94–103.

Rafi, S. H., Nahid-Al-Masood, Deeba, S. R., and Hossain, E. (2021). A short-term load forecasting method using integrated cnn and lstm network. *IEEE Access*, 9:32436–32448.

Sone, S. P., Lehtomäki, J. J., and Khan, Z. (2020). Wireless traffic usage forecasting using real enterprise network data: Analysis and methods. *IEEE Open Journal of the Communications Society*, 1:777–797.

Syu, Y., Wang, C., and Fanjiang, Y. (2019). Modeling and forecasting of time-aware dynamic qos attributes for cloud services. *IEEE Transactions on Network and Service Management*, 16(1):56–71.

Tomic, I., Bleakley, E., and Ivanis, P. (2022). Predictive capacity planning for mobile networksmdash;ml supported prediction of network performance and user experience evolution. *Electronics*, 11(4).

Yang, H., Li, X., Qiang, W., Zhao, Y., Zhang, W., and Tang, C. (2021). A network traffic forecasting method based on sa optimized arima–bp neural network. *Computer Networks*, 193:108102.