Vitor Fontana Zanotelli

Caracterização e Previsão de Falhas em Serviços de Conectividade à Internet

Vitória, ES

Maio, 2022

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Zanotelli, Vitor, 1987-
Z33c Caracterização e Previsão de Falhas em Serviços de Conectividade à Internet / Vitor Zanotelli. - 2022.
89 f. : il.

Orientador: Magnos Martinello.
Coorientador: Giovanni Comarela.
Dissertação (Mestrado em Informática) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Redes de computadores. 2. Aprendizado de máquinas. I. Martinello, Magnos. II. Comarela, Giovanni. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 004

Caracterização e Previsão de Falhas em Serviços de Conectividade à Internet

Vitor Fontana Zanotelli

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Informática.

Aprovada em 20 de maio de 2022.

Prof. Dr. Magnos Martinello
Orientador, participação presencial

Prof. Dr. Giovanni Ventorim Comarela
Coorientador, participação presencial

Prof. Dr. Vinícius Fernandes Soares Mota
Membro Interno, participação presencial

Prof. Dr. Antônio Augusto de Aragão Rocha
Membro Externo, participação presencial

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Vitória/ES, 20 de maio de 2022

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

PROTOCOLO DE ASSINATURA

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
VINICIUS FERNANDES SOARES MOTA - SIAPE 1331743
Departamento de Informática - DI/CT
Em 15/08/2022 às 15:52

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/538236?tipoArquivo=O

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

PROTOCOLO DE ASSINATURA

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
GIOVANNI VENTORIM COMARELA - SIAPE 1998739
Departamento de Informática - DI/CT
Em 15/08/2022 às 16:42

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/538345?tipoArquivo=O

Vitor Fontana Zanotelli

Caracterização e Previsão de Falhas em Serviços de Conectividade à Internet

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Mestre em Informática.

Universidade Federal do Espírito Santo – UFES

Centro Tecnológico

Programa de Pós-Graduação em Informática

Orientador: Prof. Dr. Magnos Martinello

Coorientador: Prof. Dr. Giovanni Comarela

Vitória, ES

Maio, 2022

Dedicado a minha mãe.

Agradecimentos

Agradeço a todos que me acompanharam nessa jornada. Especialmente, agradeço aos meus pais, Riviane e Marcos, que sempre me deram apoio e incentivo em todos os momentos da minha vida, sem eles nada do que fiz seria possível.

Sou grato aos meus orientadores, Magnos e Giovanni, pela oportunidade de aprendizado, mentoria e compreensão durante esse período. Ao amigo Renato, pela amizade desenvolvida em meio a códigos e modelos. Aos demais membros do NERDS, professores e servidores da UFES pela participação nesse processo de formação.

Agradeço aos meus amigos e amigas pelo suporte, companhia, cafés e cervejas. Também quero agradecer aos amigos do OLJ, presentes nas vitórias e nas derrotas dos diversos campos de batalha virtuais.

"I'm Commander Shepard, and this is my favorite dissertation on applications of Machine Learning in Computer Networks!"
(Commander Shepard of the SSV Normandy)

Resumo

O serviço de conectividade à Internet oferecido pela Rede Ipê da RNP é fundamental para a comunidade científica brasileira por interconectar universidades e centros de pesquisa em todo o país. A rede também apresenta ligações internacionais, permitindo a cooperação brasileira com entidades de pesquisa estrangeiras. É uma rede de grandes dimensões, que produz um alto volume de dados e apresenta desafios complexos relacionados ao seu funcionamento. O objetivo deste trabalho divide-se em duas partes: a primeira compreende a apresentação de uma análise da rede por meio da caracterização do comportamento de suas falhas; a segunda consiste na construção de modelos de aprendizado de máquina capazes de prever a ocorrência de falhas, possibilitando a mitigação dos problemas ocasionados por essas ocorrências. Os dados são coletados através da ferramenta Via Ipê e correspondem ao período de novembro de 2020 a novembro de 2021. Trata-se de um problema de aprendizado supervisionado, abordado como uma tarefa de classificação binária com o uso de redes neurais do tipo LSTM. A Rede Ipê apresenta comportamento heterogêneo, manifestando grande variedade na dependabilidade dos serviços de conectividade em seus diferentes PoPs. Para tratar esse cenário, são explorados diferentes modelos, dos mais gerais aos mais específicos, considerando-se as características da rede. A avaliação dos modelos revela a ocorrência de diferentes tipos de falhas, complementando a análise inicial dos dados. O problema mostrou-se complexo, mas, apesar das dificuldades encontradas, o modelo proposto mostra-se promissor e apresenta bons resultados para vários casos.

Palavras-chave: Rede Ipê. Aprendizado de Máquinas. Redes Neurais. Detecção de falhas.

Abstract

The Ipê Network is fundamental to the Brazilian scientific community, being responsible for interconnecting universities and research centers throughout the country. The network also presents international connections, allowing Brazilian cooperation with foreign research entities. It is an extensive network, producing a high volume of data and presenting challenges related to its operation. This work is divided into two parts, the first being responsible for presenting an analysis of the network through the characterization of failure behavior. The second attempt consists of constructing learning models to predict the occurrence of failures, allowing for planning on how to mitigate the problems caused by the occurrence of failures. Data is collected through the Via Ipê web app and corresponds to the period of November 2020 through November 2021. The problem is modeled as supervised learning for binary classification and recurrent neural networks (LSTMs) are used. The Ipê Network presents heterogeneous behavior, manifesting great variety on the dependability of its connectivity services in its different PoPs. Different models considering the network's characteristics are proposed to deal with this scenario, from more general to more restricted models. The models' performance metrics reveal different types of failures, complementing the initial analysis of the data. The problem is shown to be difficult, but the proposed methodology shows promise, with acceptable results in some cases.

Keywords: Rede Ipê. Machine Learning. Neural Networks. Failure Detection.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CCDF	Complementary Cumulative Distribution Function
CPU	Central Processing Unit
GPU	Graphical Processing Unit
JSON	JavaScript Object Notation
LSTM	Long short-term memory
NERDS	Núcleo de Estudos em Redes Definidas por Software
PoP	Point of Presence
QoS	Quality of Service
RNN	Recurrent Neural Network
RNR	Rede Neural Recorrente
RNP	Rede Nacional de Pesquisa
RTT	Round-Trip Time
SDN	Software Defined Network
UFES	Universidade Federal do Espírito Santo

Sumário

1 Introdução

As redes modernas estão em constante processo de evolução, tanto em sua arquitetura quanto em suas perspectivas de operação. Nesse processo de mudança, as redes deixam de ser caracterizadas apenas como dispositivos físicos provedores de conectividade. O modelo clássico representado por uma coleção de *appliances* de *hardware* conectados por enlaces muda para um novo paradigma em que redes virtualizadas podem ser provisionadas sob demanda. Com essa evolução, ocorre um aumento significativo no volume de dados gerados, transmitidos e armazenados a todo instante. A maior facilidade de acesso e aquisição de dados de monitoramento, aliada ao avanço da programabilidade de redes, coloca em foco e viabiliza o uso de aprendizado de máquinas nas operações dessas redes (BOUTABA et al., 2018).

Com relação à infraestrutura de redes, dependências adicionais e novas demandas são criadas em função da natureza distribuída e virtualizada das aplicações. Cria-se, então, uma necessidade de se oferecer garantias das propriedades de dependabilidade nos serviços de conectividade, principalmente no que se refere a disponibilidade, confiabilidade, manutenibilidade e desempenho (MARTINELLO, 2005). Para que tais garantias sejam oferecidas, é preciso que se entendam as falhas e seu comportamento em múltiplas dimensões: localização, causas, duração, frequência e tipos. Para se alcançar esse entendimento, dados de monitoramento da rede devem ser coletados, analisados e sumarizados, de forma que seja possível tomar decisões que garantam a dependabilidade do sistema. As análises devem ser capazes de gerar *insights* sobre o funcionamento da rede, suas características e particularidades. Por conta do enorme volume de dados que esses compostos de informações relativas às falhas e métricas de rede apresentam, a análise deve ser acompanhada de sumarizações capazes de auxiliar no planejamento de operação da rede e no tratamento de falhas.

O presente trabalho divide-se em duas partes. A primeira consiste na análise e caracterização dos dados coletados através da ferramenta Via Ipê referentes ao sistema de conectividade da Rede Ipê da Rede Nacional de Ensino e Pesquisa (RNP). Num primeiro momento, os dados em sua forma bruta são coletados e armazenados, e medidas de monitoramento de rede são descritas e caracterizadas por suas distribuições de probabilidade, de forma que haja um entendimento inicial do funcionamento da rede. Em seguida, o conceito de falha é definido não como o momento de completa interrupção de prestação de um serviço de conectividade, mas como o momento em que a prestação do serviço degrada além do aceitável. Considera-se a falha como o momento em que a taxa de perda de pacotes ultrapassa o limiar de 3%, este definido com base em critérios da própria Rede Ipê sobre a qualidade do serviço. Após a escolha desse limiar de tolerância, o conjunto de

dados passa por uma nova etapa de caracterização, agora considerando-se as distribuições de probabilidade relacionadas às falhas, como o tamanho das falhas, o tempo entre falhas e o número de falhas em um determinado local. Diferentes comportamentos são revelados: enquanto alguns locais são caracterizados por falhas longas e infrequentes, outros apresentam falhas curtas e frequentes. Em algumas ocasiões, as falhas apresentam correlação espacial; algumas são repentinas, enquanto outras passam por um processo gradual de degradação de qualidade. Com base nessas características, é possível realizar uma análise mais específica quanto à dependabilidade do serviço, que se mostra heterogênea nos Pontos de Presença do país. O conhecimento gerado nesse processo também permite que sejam propostas ações para o melhoramento do serviço, além de servir de base para auxiliar na construção dos modelos de aprendizado de máquina utilizados na etapa de predição.

A segunda parte do trabalho propõe uma metodologia para a predição de falhas. A abordagem proposta diferencia-se das encontradas na literatura pelo seu foco na Rede Ipê, pela forma como as falhas são definidas e pela especificação do problema de previsão. São propostos diferentes modelos de aprendizagem de máquina baseados em redes neurais recorrentes do tipo LSTM (*Long Short-Term Memory*). Essas redes têm a capacidade de tratar dados de natureza sequencial ou temporal, mostrando-se adequadas ao tipo de dado coletado na primeira etapa do trabalho. Pela dimensão da Rede Ipê e com base nos *insights* obtidos na fase inicial do trabalho, um modelo único que servisse para todo o território nacional mostrou-se improvável. Modelos são então propostos de modo a considerar diferentes níveis geográficos: do mais geral, que abrange todo o território nacional, ao mais específico, em que cada modelo abrange apenas um estado. O comportamento dos modelos também passa por uma etapa de análise, em que a performance de um modelo treinado em um estado é medida em outros locais, de forma que seja possível verificar a capacidade de generalização e identificar similaridades e diferenças no comportamento de falhas em localidades distintas. A capacidade de um modelo manter seu desempenho ao longo do tempo também é medida. Modelos são treinados no primeiro mês do conjunto de dados e seus desempenhos são registrados nos demais meses do ano. Sendo assim, é possível verificar quando um modelo precisa ser treinado novamente e identificar como a mudança no comportamento das falhas num mesmo local influencia nos resultados.

Embora a metodologia proposta ainda não apresente resultados satisfatórios no cenário mais geral (um modelo para todo o território nacional), bons resultados são alcançados em casos mais restritos, mostrando o potencial da estratégia adotada. Melhorias podem ser feitas para mitigar as dificuldades identificadas, abrindo espaço para o estudo de novas técnicas e aprimoramentos em trabalhos futuros.

1.1 Problema de pesquisa e hipóteses

A Rede Ipê disponibiliza métricas de seu funcionamento. Atualmente, no entanto, falhas são tratadas apenas de forma reativa. A existência de um sistema capaz de prever possíveis falhas com alguma antecedência pode permitir que sejam tomadas medidas para mitigar os danos ou até evitá-los. O processo de caracterização das falhas permite a identificação de diferentes comportamentos presentes no conjunto de dados, auxiliando a criação dos modelos e gerando novos *insights* sobre as características e individualidades da rede. Este trabalho propõe, então, a seguinte pergunta de pesquisa: *é possível, através do uso de algoritmos de aprendizado de máquina, criar um sistema que consiga antecipar falhas em um enlace de rede com alguma confiabilidade?*

Considerando o contexto escolhido da Rede Ipê, este trabalho apresenta as seguintes hipóteses:

- H1: O uso de algoritmos de aprendizado de máquina é adequado para a previsão de falhas em enlaces de rede;
- H2: Falhas apresentam diferentes dificuldades de predição, podendo ser fáceis, difíceis ou até impossíveis;
- H3: O comportamento das falhas não é uniforme no território nacional.

1.2 Objetivos

Este trabalho tem como objetivo principal a criação de um modelo de predição de falhas para serviços de conectividade à Internet utilizando algoritmos de aprendizado de máquina. Para atingir esse objetivo e responder às hipóteses de pesquisa, apresentam-se os seguintes objetivos específicos:

- Realizar a caracterização dos dados coletados em relação ao comportamento das falhas;
- Criar diferentes modelos de aprendizado de máquina considerando o conhecimento adquirido no processo de caracterização dos dados;
- Avaliar o comportamento dos modelos em diferentes contextos, como o tempo e o espaço.

1.3 Organização do trabalho

Este trabalho organiza-se da seguinte forma: o Capítulo 2 explica os conceitos de redes de computadores e aprendizado de máquina no contexto deste trabalho. O Capítulo 3 situa o trabalho na literatura, apresentando estudos relacionados. O Capítulo 4 apresenta a Rede Ipê, bem como informações relacionadas ao conjunto de dados coletados. O capítulo 5 define falha no contexto do trabalho e apresenta a caracterização das falhas em diferentes contextos. O capítulo 6 define o problema de previsão e apresenta os modelos de aprendizado de máquina criados. Por fim, o Capítulo 7 apresenta as conclusões e discute trabalhos futuros.

2 Referencial Teórico

O objetivo deste Capítulo é apresentar ao leitor a teoria necessária para a compreensão do trabalho. A Seção 2.1 apresenta os conceitos referentes à dependabilidade em redes de computadores. A Seção 2.2 discute os conceitos relacionados ao Aprendizado de Máquinas, métricas de avaliação e arquiteturas de modelos utilizadas.

2.1 Dependabilidade em Serviços de Conectividade

A dependabilidade pode ser definida como a capacidade de um sistema de computação (ou comunicação) de prover um serviço com um nível de confiança que seja justificado. Pode-se dizer que um sistema possui dependabilidade quando é capaz de evitar falhas em seu serviço, tanto em frequência quanto em magnitude, acima de um nível aceitável definido por seus usuários. A dependabilidade apresenta três componentes: **atributos (*attributes*)**, **ameaças (*threats*)** e **meios para sua obtenção (*means*)** (MARTINELLO, 2005; AVIžIENIS; LAPRIE; RANDELL, 2004).

As **ameaças** possuem três conceitos relacionados: *defeito (faults)*, *erros (errors)* e *falhas (failures)*. Um serviço está funcionando de maneira correta quando seu funcionamento está de acordo com sua implementação e função desejada. Caso contrário, uma *falha* ocorre quando o estado do serviço difere desse comportamento. A diferença entre o funcionamento correto e incorreto é denominada *erro* e a causa de sua ocorrência é chamada de *defeito*.

A dependabilidade possui seis **atributos** em sua composição: *confiabilidade (reliability)*, *disponibilidade (availability)*, *proteção (safety)*, *confidencialidade (confidentiality)*, *integridade (integrity)*, *manutenibilidade (maintainability)* e *segurança (security)*. Um sistema de conectividade pode apresentar todos os atributos ou apenas um subconjunto destes, de acordo com sua especificação:

- Confiabilidade: continuidade do serviço provido de maneira correta;
- Disponibilidade: prontidão do serviço provido de maneira correta;
- Proteção: ausência de ocorrências catastróficas durante a operação do serviço;
- Confidencialidade: ausência de acesso não autorizado a informações consideradas sigilosas;
- Integridade: ausência de alterações impróprias ao sistema;
- Manutenibilidade: capacidade de o sistema ser modificado ou reparado;

- Segurança: ocorrência de *disponibilidade* para usuários autorizados, *confidencialidade* dos dados e *integridade*; impróprio, nesse caso, significa não autorizado.

Os **meios** para obtenção da dependabilidade podem ser divididos em quatro categorias: *prevenção de defeitos (fault prevention)*, *tolerância de defeitos (fault tolerance)*, *remoção de defeitos (fault removal)* e *predição de defeitos (fault forecasting)*:

- Prevenção de defeitos: formas de prevenir a ocorrência ou introdução de defeitos;
- Tolerância de defeitos: formas de evitar a ocorrência de falhas em meio à presença de defeitos;
- Remoção de defeitos: formas de reduzir a frequência e a magnitude de defeitos;
- Predição de defeitos: formas de inferir o número atual, as ocorrências futuras e as possíveis consequências de defeitos.

Para este trabalho, a característica de interesse nas **ameaças** são as *falhas*, os **atributos** de interesse são a *confiabilidade*, *disponibilidade* e *manutenibilidade* e o **meio** estudado é o de **predição**, mas este relacionado às *falhas* e não aos *defeitos*. Uma descrição mais aprofundada do contexto do estudo de falhas encontra-se na Seção 5.

2.2 Aprendizado de Máquinas

O aprendizado de máquina (*machine learning*) baseia-se na criação de sistemas capazes de adquirir conhecimento próprio através da extração de padrão dos dados, sendo, portanto, capazes de resolver tarefas difíceis de serem descritas formalmente e que, em alguns casos, parecem intuitivas ou mesmo automáticas para seres humanos (GOODFELLOW; BENGIO; COURVILLE, 2016).

O campo apresenta três conceitos principais: **dados**, **modelo** e **aprendizado**. Modelos relacionados a conjuntos de dados são criados e passam por um processo de aprendizado. Esse processo pode ser entendido como a forma automática de se encontrarem padrões e estrutura nos dados através da otimização dos parâmetros do modelo, com o objetivo de generalizar o que foi aprendido para dados ainda não vistos (DEISENROTH; FAISAL; ONG, 2020).

Uma definição formal proposta por (MITCHELL, 1997) para o aprendizado é: *Podemos dizer que um programa de computador aprendeu através de uma experiência* ***E*** *em relação a uma classe de tarefas* ***T*** *e métricas de performance* ***P*** *se sua performance em tarefas em* ***T*** *aumenta com a experiência* ***E***. Um exemplo clássico é o de um modelo usado para identificar a presença de gatos em fotos que apresenta uma taxa de acertos maior

conforme a experiência de ser exposto a mais fotos de animais.

Algoritmos de aprendizado de máquinas, em geral, encontram-se na literatura divididos em três tipos: *aprendizado supervisionado*, *aprendizado não supervisionado* e *aprendizado por reforço*:

- Aprendizado supervisionado: o conjunto de dados é rotulado e, durante o processo de treinamento, o modelo aprende a mapear uma observação ao seu rótulo correto;
- Aprendizado não supervisionado: o conjunto de dados não é rotulado e, durante o processo de treinamento, o modelo aprende a dividir o conjunto de dados de forma que observações semelhantes encontrem-se próximas e observações dessemelhantes possuam uma distância maior entre si;
- Aprendizado por reforço: o algoritmo seleciona e executa ações que recebem um *feedback* positivo ou negativo do ambiente e, com base nesses resultados, seus parâmetros são otimizados, de forma que consiga alcançar a melhor estratégia possível quando seus resultados são majoritariamente positivos.

Com relação ao tipo de problema ou tarefa na utilização dos modelos, os mais comuns são:

- Problemas de classificação: o algoritmo realiza um mapeamento entre os dados de interesse e possíveis classes ou categorias;
- Problemas de regressão: o algoritmo realiza um mapeamento entre os dados de interesse e possíveis valores numéricos.

Para este trabalho, os modelos escolhidos são os de aprendizagem supervisionada para resolver um problema de classificação. O conjunto de dados utilizado é descrito na Seção 4 e os modelos encontram-se na Seção 6.

2.2.1 Métricas de Avaliação

Ao se propor uma solução de aprendizado de máquinas para um problema, utilizam-se métricas de avaliação para validar a escolha de modelos. Dessa forma, é possível medir a performance e verificar a capacidade de generalização dos modelos propostos. Métricas também são empregadas para realizar a comparação entre diferentes modelos e propostas de solução, de forma que seja possível selecionar a melhor maneira de se resolver o problema e justificar as escolhas feitas durante o processo de modelagem.

Para um problema de classificação binária, um modelo associa um rótulo positivo ou um negativo a cada observação do conjunto de dados (1 e 0 respectivamente), tornando

possível a ocorrência de quatro tipos de resultado: verdadeiro positivo (*true positive ou TP*), verdadeiro negativo (*true negative ou TN*), falso positivo (*false positive ou FP*) e falso negativo (*false negative ou FN*), sendo:

- Verdadeiro Positivo: o modelo associa corretamente o rótulo positivo a uma observação pertencente à classe positiva;
- Verdadeiro Negativo: o modelo associa corretamente o rótulo negativo a uma observação pertencente à classe negativa;
- Falso Positivo: o modelo associa erroneamente o rótulo positivo a uma observação pertencente à classe negativa;
- Falso Negativo: o modelo associa erroneamente o rótulo negativo a uma observação pertencente à classe positiva.

A partir desses resultados, é possível definir as métricas de avaliação para um modelo. Para este trabalho, são utilizadas: **acurácia (*accuracy*)**, **precisão (*precision*)** e **revocação (*recall*)**.

Acurácia

Descreve o quanto o modelo acertou considerando todas as classes, representando a fração de acertos com relação ao todo. Pode ser vista na Equação 2.1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

É importante observar que, nos casos em que o conjunto de dados é muito desbalanceado (i.e. existe uma quantidade muito maior de representantes de uma classe com relação a outra), a acurácia pode não ser uma métrica adequada para se julgar a performance do modelo. Esta particularidade é discutida em seções seguintes do trabalho.

Precisão

É calculada pela razão entre o número de classificações corretas da classe positiva e o número total de classificações da classe positiva. Mede a exatidão que um modelo tem ao inserir corretamente uma observação na classe positiva. Pode ser vista na Equação 2.2.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

Revocação

É calculada pela razão entre o número de classificações corretas da classe positivas e o número total de observações no conjunto de dados pertencentes à classe positiva. Mede a capacidade que o modelo tem de detectar observações pertences à classe positiva. Pode ser vista na Equação 2.3.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

2.2.2 Arquitetura de Modelos

O campo de Aprendizado de Máquinas apresenta uma variedade de propostas de algoritmos e arquiteturas. A escolha de uma solução adequada está relacionada ao tipo de problema a ser resolvido e aos dados disponíveis. Pela natureza temporal dos dados utilizados neste trabalho, as redes neurais do tipo *Long Short-Term Memory* (LSTM) foram escolhidas para tratar o problema de predição. Foi utilizado, também, um modelo mais simples de rede neural para comparar o resultado, o *Multilayer Perceptron* (MLP).

Multilayer Perceptron (MLP)

O *Multilayer Perceptron* é também conhecido como uma rede neural de alimentação direta (*feedforward neural network*). Como definido em (GOODFELLOW; BENGIO; COURVILLE, 2016), dada a existência de uma função f que mapeia vetores $\mathbf{x}$ de um conjunto de dados ao seu respectivo rótulo y, o MLP tem como objetivo encontrar uma função f' que se aproxime de f. Sua descrição é dada pela Equação 2.4, onde $\boldsymbol{\theta}$ é um vetor de parâmetros do modelo que torne a função f' o mais próxima de f.

$$y = f'(\boldsymbol{x}, \boldsymbol{\theta}) \tag{2.4}$$

O MLP é considerado uma rede *feedforward* porque a informação segue em apenas um sentido, ou seja, não existe um mecanismo de *feedback* em que a saída do modelo possui ligação com sua entrada. A inclusão deste tipo de mecanismo é uma característica das redes neurais recorrentes, como é o caso das redes neurais LSTM descritas na Seção 2.2.2.

Um modelo MLP pode apresentar múltiplas camadas e, nesse caso, é representado por um encadeamento de funções, como mostra a Equação 2.5, onde $f^{(1)}$ representa a primeira camada, $f^{(2)}$ a segunda e assim por diante, até a n-ésima camada $f^{(n)}$.

$$f'(\boldsymbol{x}) = f^{(n)} \circ f^{(n-1)} \circ ... \circ f^{(1)} \tag{2.5}$$

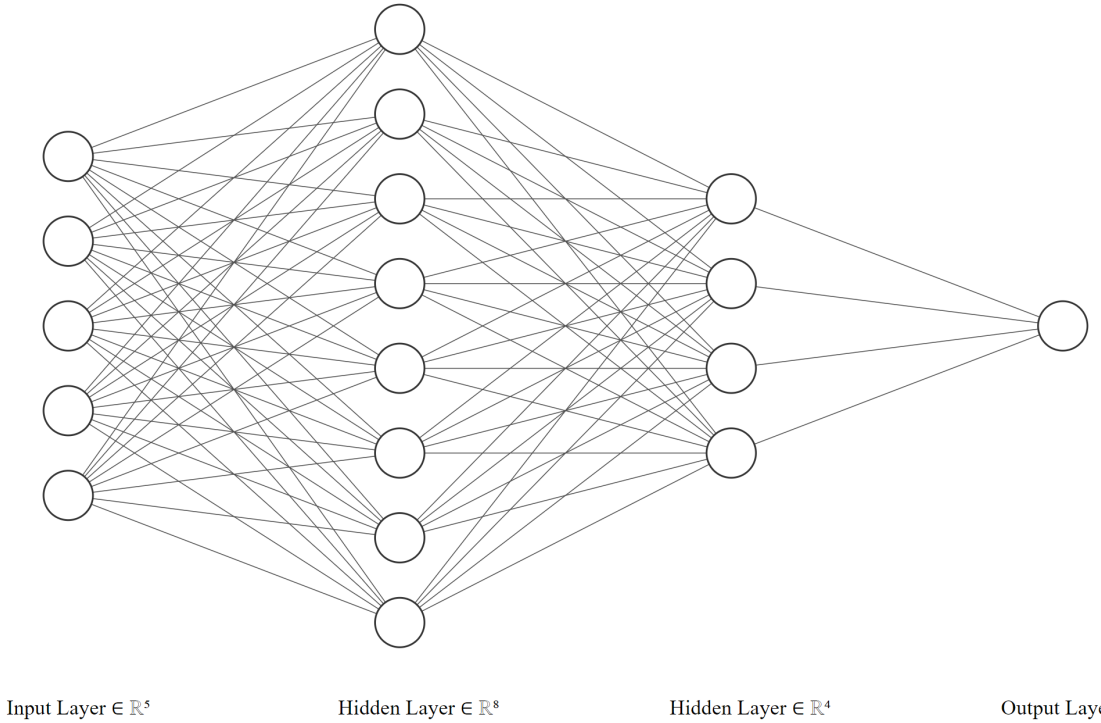

Figura 1 – *Multilayer Perceptron* de 4 camadas.

Cada camada de uma rede neural é composta por uma ou mais unidades, ou neurônio, e cada unidade de uma camada está ligada a todas as unidades da camada seguinte. Uma unidade realiza uma combinação linear de suas entradas, ponderada por pesos relacionados a cada ligação, e aplica uma função de ativação ao resultado. A Equação 2.6 descreve como é calculada a saída y de uma unidade onde a entrada é um vetor $\boldsymbol{x}$ de tamanho n, w_i representa o peso associado à ligação i e a função de ativação é ϕ.

$$y = \phi\left(\sum_{i=1}^{n} x_i w_i\right) \tag{2.6}$$

A Figura 1, criada pela ferramenta desenvolvida em (LENAIL, 2019), apresenta como exemplo uma representação visual de um modelo MLP. Essa rede neural apresenta uma camada de entrada formada por cinco unidades, seguida por duas camadas ocultas com oito e quatro unidades respectivamente e, por fim, uma camada de saída com uma unidade.

Redes Neurais Recorrentes (RNR)

As redes neurais recorrentes funcionam de forma similar às redes *feedforward*, mas apresentam a adição de *feedback*, uma ligação da saída de uma camada com a entrada de uma camada anterior. Essa adição permite a persistência de informação no modelo, tornando seu uso adequado ao tratamento de sequências. Entradas usuais para esse tipo de modelo são frases em linguagem natural ou sequências numéricas, como em séries

temporais.

A unidade de uma rede neural recorrente opera de forma similar à de uma unidade pertencente a uma rede neural com apenas *feedforward*, mas apresenta a adição da saída do instante no tempo anterior na sequência recebida como entrada. Dessa forma, num instante t o resultado y_t para uma unidade é uma combinação linear das entradas nesse instante e as saídas do instante $t-1$, como mostra a Equação 2.7, onde $\boldsymbol{W_x}$ é a matriz de pesos associada às ligações de entrada no instante atual, $\boldsymbol{W_y}$ a matriz de pesos associada às ligações com os resultados do instante anterior, $\boldsymbol{x_t}$ o vetor de entrada no instante atual e $\boldsymbol{y_{(t-1)}}$ o vetor com as saídas do instante anterior.

$$y_t = \phi\left(\boldsymbol{W_x x_t} + \boldsymbol{W_y y_{(t-1)}}\right) \tag{2.7}$$

As RNRs apresentam bons resultados quando a informação relevante a um item na sequência não se encontra muito distante, porque a retenção de informação limita-se ao que se chama de *short-term memory*. Uma solução para esse problema é o uso das redes neurais recorrentes do tipo *Long Short-Term Memory* (LSTM) (HOCHREITER; SCHMIDHUBER, 1997). As redes LSTM possuem um *long-term state*, onde a informação é armazenada e, com o uso de um *forget gate* e um *input gate*, selecionam quando adicionar ou remover informações dessa memória de longo prazo. Com essa capacidade de persistir e atualizar informação na rede quando necessário, as LSTMs conseguem capturar padrões longos em sequências.

3 Trabalhos Relacionados

O objetivo desta Seção é posicionar o presente trabalho na literatura da área. Serão apresentas publicações relacionadas à área de redes, caracterização de falhas e uso de aprendizado de máquinas para predição.

Redes de computadores encontram-se cada vez mais presentes na sociedade, apresentando um crescimento que não deve desacelerar nos próximos anos, considerando-se a quantidade de usuários com acesso à Internet, o uso de celulares e a presença equipamentos IoT[1]. Nesse contexto, grandes quantidades de dados são transmitidas e geradas, tanto pelos usuários quanto em telemetria de rede.

As análises desses dados eram, tradicionalmente, realizadas de forma *offline* e ações subsequentes dificilmente poderiam ser efetuadas em tempo real. Uma mudança desse paradigma tornou-se possível na última década com o avanço e maior disseminação de tecnologias relacionadas às redes programáveis (BOSSHART et al., 2014; Liberato et al., 2018). Tornaram-se comuns, atualmente, a análise em tempo real e a implementação de modelos de aprendizado de máquinas em ambientes de produção. A literatura evidencia o interesse da comunidade científica na utilização de modelos de aprendizado de máquina, e trabalhos que empregam essas técnicas encontram-se em diversas áreas, como classificação de tráfego, roteamento e predição de tráfego, controle de congestionamento, segurança de redes e gestão de controle, recursos, falhas, QoS e QoE (BOUTABA et al., 2018).

Dados originários de redes de computadores são comumente encontrados na forma de séries temporais. como sequências de medições, métricas ou *logs* do estado da rede. O tratamento de séries temporais é reconhecido na literatura como um problema desafiador e presente na interseção da área de redes de computadores com outras disciplinas como a mineração de dados (*data mining*) (YANG; WU, 2006). Algoritmos capazes de tratar esse problema de forma satisfatória são alvo de pesquisa e centenas já se encontram em uso. Além disso, é possível encontrar comparações entre o desempenho dos principais tipos de algoritmos, como *Time Series Forest* (TSF), *Wighted Dynamic Time Warping* (WDTW), *Collective Of Transformations-based Ensemble* (COTE) e *Elastic Ensemble* (BAGNALL et al., 2017).

Uma abordagem que, nos últimos anos, apresentou sucesso em incrementar o estado da arte é o uso de aprendizado de máquinas, em especial o Aprendizado Profundo (*Deep Learning*), com aplicações em tarefas como processamento de imagens, vídeos e tratamento de dados sequenciais através das Redes Neurais Recorrentes (LECUN; BENGIO; HINTON,

[1]Cisco Annual Internet Report: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

2015). A tarefa de classificação em séries temporais também encontra propostas de solução no *Deep Learning* (WANG; YAN; OATES, 2017). O modelo mais simples é o de *Multilayer Perceptron* (MLP), mas, nesse tipo de modelo, a informação das relações temporais é perdida e os elementos da série temporal são tratados como independentes entre si, sem a sua dependência temporal. As Redes Neurais Recorrentes (RNN) conseguem capturar essa dependência sequencial, mas sofrem o problema de dissipação do gradiente (*Vanishing Gradient*), que dificulta o aprendizado de sequências longas. Como discutido na Seção 2, as redes LSTM são uma forma tratar esse problema. A literatura também apresenta outras soluções de *Deep Learning* para tratar problemas de natureza temporal, como *Time Le-Net*, *Multi Channel Deep Convolutional Neural Network* (MCDCNN), *Time Convolutional Neural Network* (Time-CNN) e *Echo State Networks* (ESNs) (FAWAZ et al., 2019).

A caracterização de falhas em redes de computadores pode ser tratada como um problema de classificação de séries temporais em que, após uma sequência de estados da rede, o modelo informa se ocorrerá ou não uma falha. É um problema que se encontra consolidado na literatura de redes há décadas, e conta com a presença de estudos renomados que realizam a análise e a categorização das falhas por meio de diversas metodologias. Em (Markopoulou et al., 2008) os autores utilizam uma abordagem estatística para caracterizar as falhas nos enlaces de rede de uma rede IP. Falhas são divididas em classes e, a partir de propriedades identificadas como relevantes, (e.g. tempo entre falhas e tempo até o reparo) modelos probabilísticos são utilizados para sua predição. O estudo também revela que uma pequena parte dos links é responsável pela maioria das falhas. Esse tipo de comportamento também é encontrado na Rede Ipê, e será apresentado na Seção 5. Encontra-se em (GILL; JAIN; NAGAPPAN, 2011) uma abordagem similar, mas no contexto de *Data Centers*. Análises estatísticas são realizadas para caracterizar as falhas e seu impacto.

Trabalhos que incorporam o aprendizado de máquinas ao ferramental utilizado para resolver problemas de rede tornaram-se comuns nos últimos anos e o uso de tais técnicas na tarefa de predição pode ser exemplificado por (Azzouni; Pujolle, 2018), (OLIVEIRA; BARBAR; SOARES, 2016) e (Lens Shiang et al., 2020), que tratam a previsão de tráfego em diferentes contextos. O primeiro, assim como o trabalho proposto por esta pesquisa, utiliza redes neurais recorrentes do tipo LSTM, mas com o intuito de prever valores em matrizes de tráfego utilizando dados reais provenientes da rede GÉANT. O segundo faz uma comparação entre modelos do tipo MLP, RNN e SAE (*Stacked Autoencoder*) para a previsão em dados de trafégo provenientes de uma ISP (*Internet Service Provider*) presente em 11 países europeus. Os resultados mostram que modelos mais simples como o MLP e RNN podem apresentar resultados superiores a modelos mais complexos com o SAE, além da vantagem de se usarem modelos de rede neural recorrente com relação aos modelos MLP para tratar dados de natureza temporal. Por fim, o terceiro trabalho utiliza redes neurais recorrentes do tipo GRUs (*Gated Recurrent Units* — um modelo mais 'simples'

quando comparado às LSTMs) para a previsão de tráfego em redes 5G, usando dados coletados e disponibilizados por um provedor de serviços telefônicos europeu. Ambos os trabalhos apresentam resultados com alta acurácia, ilustrando a capacidade dos modelos de aprendizado de máquina em tratar problemas de rede.

Com relação à previsão de falhas, também são encontradas diferentes propostas de solução para aplicações em redes. Utilizando dados provenientes da rede acadêmica NERSC (*National Energy Research Scientific Computing Center*)[2], em (GIANNAKOU; DWIVEDI; PEISERT, 2020) o problema de previsão da retransmissão de pacotes TCP é tratado com o uso de modelos de Floresta Aleatória (*Random Forest*). É realizada uma seleção de parâmetros e, de 52 medidas inerentes ao fluxo IP, apenas sete são utilizadas (e.g. tamanho do arquivo e duração do fluxo), alcançando acurácia próxima de 99%. Em (Zhou; Zhang, 2018), diversos algoritmos de classificação são utilizados para detectar a perda de pacotes em dados de telemetria pertencentes a uma rede 4G LTE e coletados a partir de um ambiente de virtualização de funções de rede vEPC (*Virtual Evolved Packet Core*). Os dados selecionados passam por um processo de redução de dimensionalidade e algoritmos como Árvore de Decisão (*Decision Tree*), Floresta Aleatória, SGD e Redes Neurais MLP alcançam alta acurácia, próxima ou superior a 95%.

Para a previsão de falhas baseada em logs, em (Zhong; Guo; Wang, 2016), dados originários de uma rede metropolitana passam por um processo de extração de características para servirem de entrada em modelos, usando os algoritmos RIPPER, BayesNet, Floresta Aleatória e a distribuição de Weibull para a ocorrência de falhas na rede em janelas de tempo de tamanhos variados. Os resultados são bem distintos e o algoritmo de RIPPER destaca-se apresentando acurácia superior a 60% e revocação superior a 80% em alguns casos. Em (ZHANG et al., 2016), são usados logs reais provenientes de dois sistemas empresariais, um *Web Server Cluster* (WSC) e um *Mailer Server Cluster* (MSC). Um processo de mineração de texto, com auxílio de uma técnica inspirada em TF-IDF (*Term Frequency-Inverse Document Frequency*), extrai *features* para serem usadas em redes LSTM na predição das falhas, apresentando resultados com precisão e revocação superiores a 70 e 80% respectivamente.

O uso de aprendizado de máquinas e redes LSTM para a predição de falhas não é novo, mas este trabalho difere-se dos demais:

- Na natureza dos dados coletados, originários da Rede Ipê. Utiliza-se, portanto, um conjunto de dados com características distintas e de importância para a pesquisa brasileira;
- Na disponibilização dos dados coletados para a comunidade;

[2]NERSC: <https://www.nersc.gov/>

- Na definição de falha e na forma como o problema de previsão é apresentado: falhas são tratadas como uma forma de quebra de SLA (*Service Level Agreement*) e uma falha ocorre quando o serviço deixa de apresentar um desempenho dentro dos níveis estipulados.

4 Conjunto de Dados

O objetivo desta seção é apresentar ao leitor o conjunto de dados utilizado neste trabalho. A Seção 4.1 apresenta a origem dos dados, descrevendo a Rede Ipê e a ferramenta Via Ipê. Em seguida, a Seção 4.2 descreve como foi efetuada a coleta e o armazenamento dos dados. A Seção 4.3 apresenta o pré-processamento e a descrição dos dados coletados. A Seção 4.4 caracteriza o conjunto de dados e, por fim, a Seção 4.5 apresenta o ambiente usado para programação, a linguagem e as questões relativas à reprodutibilidade do trabalho.

4.1 A Rede Ipê

A crescente demanda por conectividade traz grandes desafios aos provedores que precisam atender a certas expectativas de disponibilidade e qualidade na transferência de dados em seus serviços. O Brasil possui uma vasta extensão territorial e configura-se como um dos maiores países em área, o que torna ainda mais complexo o problema de gerir e garantir a prestação desse tipo de serviço. Para prover serviço de conectividade no território nacional, a Rede Nacional de Pesquisa (RNP)[1] oferece o serviço da Rede Ipê.

A Rede Ipê abrange todo o território brasileiro e possui um Ponto de Presença (PoP - *Point of Presence*) em cada um dos 26 estados e do Distrito Federal. Cada um desses pontos opera o serviço de conectividade em sua respectiva localidade. Trata-se de uma rede de caráter acadêmico e, conforme a topologia nacional apresentada na Figura 2,

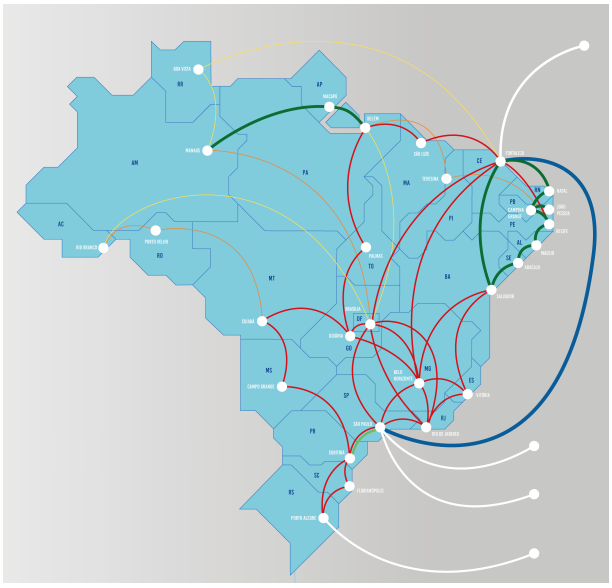

Figura 2 – Topologia Rede Ipê em 2020. Fonte: https://www.rnp.br/en/ipe-network

[1] RNP e Rede Ipê: <https://www.rnp.br/en/ipe-network>

possui tanto ligações entre PoPs quanto ligações externas ao país. Cada PoP apresenta também uma topologia local e estão ligados às instituições de ensino e pesquisa.

Por meio de suas conexões internacionais, a Rede Ipê está conectada com os demais países da América Latina através da RedCLARA[2], com a Europa através da rede Géant[3] e com os Estados Unidos pela AmLight[4]. Dessa forma, um importante aspecto da Rede Ipê para a pesquisa nacional é a presença de uma infraestrutura que facilita a colaboração entre instituições brasileiras e estrangeiras.

Com o intuito de garantir um bom funcionamento e atender às necessidades de seus usuários, a RNP disponibiliza o Via Ipê[5], uma ferramenta de monitoramento constante, de fácil acesso e uso. Disponibilizada através de um aplicativo web, a ferramenta traz transparência ao serviço prestado e possibilita que os usuários tenham acesso às estatísticas de monitoramento da Rede Ipê.

A ferramenta Via Ipê permite que qualquer usuário com acesso à internet verifique o estado de conectividade de uma instituição conectada a um PoP. É possível fazer a busca pelo nome da instituição ou navegar pelo mapa oferecido pelo aplicativo web. As informações fornecidas pela ferramenta compreendem tanto a qualidade do serviço no instante da pesquisa quanto informações mais detalhadas e gráficos históricos de seu funcionamento, como exemplificado na Figura 3.

A plataforma Via Ipê possui uma infraestrutura distribuída de coleta de dados em que, a cada minuto, dados referentes à qualidade de todos os enlaces de rede são coletados e sumarizados para gerar uma visualização na ferramenta gráfica oferecida. Esses dados encontram-se em seu estado bruto armazenados como arquivos no formato JSON (*JavaScript Object Notation*) contendo toda a informação referente ao estado da rede em determinado momento. Embora não seja objetivo da ferramenta o fornecimento desses arquivos em seu estado bruto, eles se encontram também disponíveis para download. Foram esses os arquivos coletados para a realização deste trabalho. O processo de coleta e descrição dos dados encontram-se nas Seções 4.2 e 4.3.

A disponibilização e visualização dos dados não tem como objetivo principal o gerenciamento e a tomada de decisões da Rede Ipê. Entretanto, por possuir uma interface pública, monitoramento constante e grande volume de dados, a ferramenta Via Ipê torna-se ideal no auxílio da criação de um conjunto de dados para pesquisa e experimentação relacionadas a serviços de conectividade. A importância da Rede Ipê no cenário nacional de pesquisa justifica ainda mais a escolha desse conjunto, além da possibilidade deste trabalho ou de trabalhos futuros contribuírem para seu melhor funcionamento.

[2]RedCLARA: <https://www.redclara.net/>
[3]Rede Géant: <https://network.geant.org/>
[4]Rede AmLight: <https://www.amlight.net/>
[5]Ferramenta Via Ipê: <https://viaipe.rnp.br/>

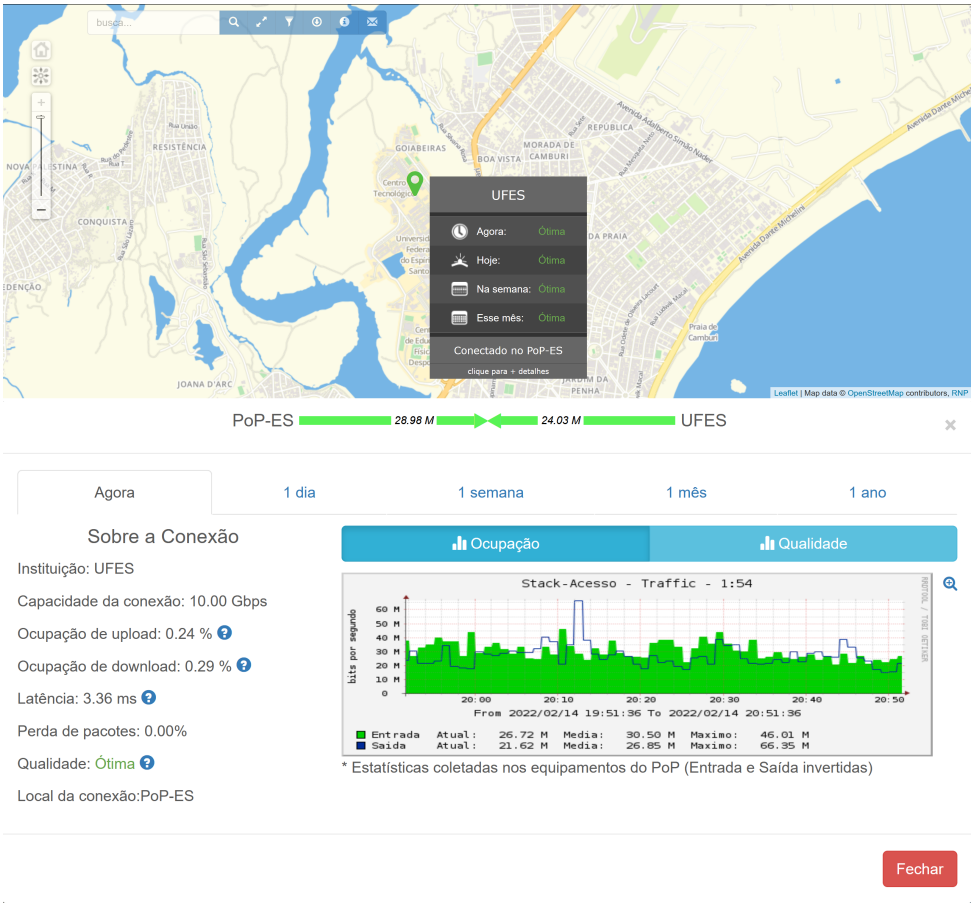

Figura 3 – Ferramenta Via Ipê. Fonte: https://viaipe.rnp.br

4.2 Coleta e armazenamento

A interface gráfica do Via Ipê é ideal para o usuário comum verificar o estado da rede, mas, para uma análise completa de toda a região, são usados os dados brutos. Uma das etapas do trabalho é, então, a coleta dos dados brutos disponibilizados pela ferramenta. Apesar de apresentar informação histórica nos gráficos, não são disponibilizados os dados brutos antigos e consta apenas o JSON do minuto atual para download. A solução encontrada para esse problema foi a criação de um coletor implementado por um *script* Python que, a cada minuto, acessa a interface do Via Ipê via protocolo HTTP, faz o download do arquivo JSON correspondente ao minuto atual e armazena em um servidor no laboratório NERDS [6] (Núcleo de Estudos em Redes Definidas por Software) da UFES. Os arquivos são organizados hierarquicamente em pastas de acordo com sua *timestamp* na forma: *ano/dia/hora/minuto.JSON*. Esse conjunto de dados brutos é sempre mantido sem alteração para garantir a validação de resultados obtidos e a reprodutibilidade do trabalho.

Para a realização do trabalho, foram selecionados dois subconjuntos de tempo de funcionamento do coletor. O primeiro subconjunto, que compreende apenas duas semanas

[6] Laboratório NERDS: <http://nerds.ufes.br/>

Tabela 1 – Minutos faltantes por mês.

Mês	Minutos	% do mês	Mês	Minutos	% do mês
11/2020	55	0.12	06/2021	194	0.45
12/2020	93	0.21	07/2021	598	1.33
01/2021	170	0.38	08/2021	1371	3.07
02/2021	162	0.40	09/2021	5247	12.15
03/2021	242	0.54	10/2021	1172	2.62
04/2021	28	0.06	11/2021	338	0.78
05/2021	215	0.48			

do mês de abril de 2020, não será descrito em mais detalhes, porque foi utilizado apenas para a criação das ferramentas e do código utilizados neste trabalho e sua validação. Também serviu para gerar alguns *insights* sobre o conjunto de dados, de forma a evitar que se criasse algum viés. O segundo conjunto de dados, que constitui de fato o objeto de análise nesse trabalho, compreende o mês de novembro de 2020 até novembro de 2021. Dessa forma, é possível observar efeitos sazonais e possíveis diferenças de comportamentos da rede no tempo (e.g. férias, inicial e final de período letivo, feriados etc.).

Um fato importante torna esse período atípico em relação ao histórico da Rede Ipê. Esse período foi marcado pela pandemia de COVID-19[7], que pode ter causado uma diferença no comportamento da rede em relação aos demais anos de sua existência. Não é objeto de pesquisa deste trabalho relacionar a pandemia aos resultados e às análises obtidas; tais efeitos e qualquer correlação ou causa relacionados à pandemia encontram-se fora de escopo. No entanto, é importante registrar que esse evento alterou o funcionamento de universidades, com muitas instituições realizando o ensino remotamente e com um calendário distinto do tradicional. Também ocorre que o modo de trabalho de operadores de rede, como outras profissões, foi afetado, assim como o de pesquisadores e demais usuários ou mantenedores da Rede Ipê.

Como anteriormente mencionado, os dados são baixados de minuto em minuto e, dessa forma, o conjunto de dados contém todos os minutos coletados do primeiro dia de novembro de 2020 até o último minuto de novembro de 2021, totalizando um ano e um mês. Nesse período, alguns minutos estão ausentes. Esses dados faltantes decorrem primariamente de falhas no coletor (e.g. queda de energia, problemas com a internet nos servidores etc.), ou de algum erro na página do Via Ipê durante a coleta, ou ainda de algum arquivo baixado que estava corrompido. A causa não é registrada e todos esses minutos são tratados da mesma maneira.

O número de minutos faltantes para cada mês encontra-se na Tabela 1. Pode-se observar que, no geral, são poucos: nove dos meses em questão apresentam menos que

[7] Linha do tempo COVID-19: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline>

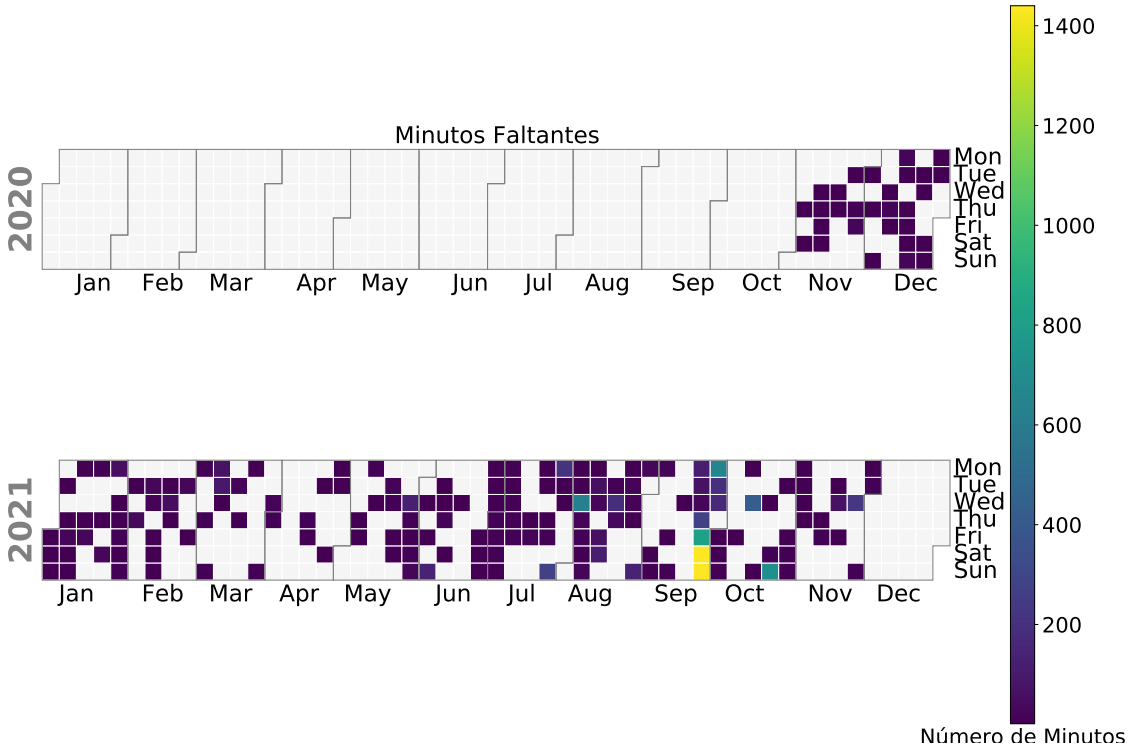

Figura 4 – Calendário de minutos faltantes

1.5% dos minutos ausentes. Os meses de agosto, setembro e outubro de 2021 concentram a maior parte dos casos, representando 78.9% dos casos. O pior caso encontra-se no mês de setembro de 2021, com mais de cinco mil minutos, ou seja, cerca de três dias e meio. Embora a ocorrência de perdas não seja o cenário ideal, ao analisar em proporção a quantidade de minutos de um mês, do melhor caso com 0.06% (apenas 28 minutos em abril de 2021) ao pior caso com 12.15%, conclui-se que a quantidade de perda não impossibilita a análise a ser realizada.

Outra informação importante para se avaliar o impacto dos minutos ausentes é a forma como estes se encontram distribuídos nos meses: se estão concentrados em algum dia ou espalhados pelo mês, se estão concentrados em dias de semana ou finais de semana, no começo ou no fim de mês. Para essa análise, a Figura 4 mostra um mapa de calor em um calendário correspondente ao intervalo de tempo em questão.

Para a maioria dos meses, a perda é baixa e os minutos encontram-se espalhados em diversos dias, sem algum padrão recorrente. Apenas os três piores casos apresentam alguns dias com uma concentração maior de perda, com aproximadamente 10 horas em falta. O grande número de perda do mês de setembro está concentrado em apenas dois dias consecutivos, que estão completamente ausentes e constituem o último final de semana do mês. A visualização da distribuição reforça a afirmação anterior de que, mesmo não sendo um cenário ideal, se tratados adequadamente, os minutos faltantes não impedem a realização das análises.

time	estado	local	interface	client_side	packet_loss	rtt	download	upload
2021-02-01 13:01:00	ES	CETEM-ES	e9147	True	0.00	3.02	1.302409e+06	180336.15
2021-02-01 13:01:00	ES	EBSERH-HUCAM	e9194	True	0.41	1.01	1.149780e+08	13600047.41
2021-02-01 13:01:00	ES	EMESCAM	e9123	False	0.00	0.95	2.479452e+07	9565954.35
2021-02-01 13:01:00	ES	FAPES	e9149	True	0.00	0.90	2.024947e+06	7087227.37
2021-02-01 13:01:00	ES	HUCAM	e9128	False	0.00	1.01	9.877460e+07	8503881.56
...	...	...	...	...	...	...	...	...
2021-02-28 23:59:00	ES	SECT	e9687	False	0.00	0.93	8.889083e+04	14603.29
2021-02-28 23:59:00	ES	UFES	e9619	False	0.00	7.16	2.249391e+07	24285465.18
2021-02-28 23:59:00	ES	UFES-Alegre	e9683	True	0.00	3.55	1.981134e+05	1461444.06
2021-02-28 23:59:00	ES	UFES-Jeronimo-Monteiro	e9694	True	0.00	3.17	1.093273e+06	950911.12
2021-02-28 23:59:00	ES	UFES-SaoMateus	e9684	True	0.00	3.41	3.532023e+05	3162457.56

Figura 5 – Exemplo de *Data Frame*

4.3 Pré-processamento e descrição dos dados

Para o início do processo de análise os dados, os arquivos originais JSON devem passar por uma etapa de pré-processamento. O primeiro passo é transformar os dados em uma estrutura de dados mais adequada e armazená-los em um local diferente, de forma que modificações sejam feitas sem comprometer os arquivos originais.

O conteúdo dos arquivos JSON é transformado em estruturas de dados do tipo *Data Frame* da biblioteca Pandas[8] da linguagem Python[9], seguindo o modelo de *tidy data* proposto por Hadley Wickham (WICKHAM, 2014). Facilita-se, assim, o processo seguinte de análise exploratória dos dados, possibilitando tanto o uso de ferramentas padronizadas já existentes quanto o aproveitamento de ferramentas desenvolvidas neste trabalho para uso futuro em casos semelhantes. A Figura 5 apresenta, como exemplo, o recorte de um *Data Frame* usado no trabalho.

Data Frame é um formato tabular em que cada linha é considerada uma *observação* ou um *objeto de interesse*, e as colunas representam uma variável[10]. Esse formato tem equivalência direta com uma notação matemática utilizando a seguinte matriz $\boldsymbol{D}_{m,n}$:

$$\boldsymbol{D} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} \text{———} & \boldsymbol{x_1^T} & \text{———} \\ \text{———} & \boldsymbol{x_2^T} & \text{———} \\ & \vdots & \\ \text{———} & \boldsymbol{x_n^T} & \text{———} \end{bmatrix}$$

O conjunto de dados pode ser representado formalmente por essa matriz ou por um

[8] Biblioteca Pandas: <https://pandas.pydata.org/>

[9] Mais informações sobre a linguagem e bibliotecas utilizadas encontram-se na Seção 4.5

[10] Variável, feature e característica são utilizadas com o mesmo significado no texto deste trabalho.

Tabela 2 – Descrição das variáveis do conjunto de dados.

Variável	**Descrição**
Tempo	Data e hora da medição. Com precisão em minutos.
Estado	Sigla representando o estado em que a medição ocorreu.
Local	Local ligado ao PoP a que a medição se refere.
Interface	Identificador da interface de rede do local.
Client Side	Indica se a medição foi realizada do lado do *client* (local) ou PoP.
Packet Loss	Valor de perda de pacotes. Em percentual.
RTT	*Round Trip Time*. Medido em milissegundos (ms).
Download	Taxa de download. Medida em bits por segundo (bps).
Upload	Taxa de upload. Medida em bits por segundo (bps).

conjunto $\mathbf{C}$ de $\mathbf{m}$ vetores de observações $\boldsymbol{C} = \{\boldsymbol{x_1}, ..., \boldsymbol{x_m}\}$ onde $\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}$ é um vetor com $\mathbf{n}$ variáveis.

Os arquivos JSON possuem uma diversidade de informações coletadas nos enlaces durante o seu momento de criação. Entretanto, para este trabalho, apenas um subconjunto dessas informações é utilizado. As variáveis de interesse e sua descrição encontram-se na Tabela 2, sendo estas as colunas dos *Data Frames* utilizados. Com a descrição das variáveis, interpreta-se uma observação do conjunto de dados. Por exemplo, a primeira linha do *Data Frame* apresentado na Figura 5 é interpretada da seguinte forma: *No dia primeiro de fevereiro de 2021, as 13:01 pm, no estado do Espírito Santo, o Centro de Tecnologia Mineral (CETEM), em sua interface de rede denominada e9147, com medições feitas a partir do cliente, apresentou perda de pacotes de 0%, um RTT de 3.02 milissegundos, uma taxa de download de* $1.3 \cdot 10^6$ *bps e upload de* $1.8 \cdot 10^5$ *bits por segundo.*

O conjunto de dados é então composto por nove variáveis: cinco delas individualizam uma observação (tempo, estado, local, interface, *client side*) e quatro são variáveis numéricas que quantificam a observação (*packet loss*, rtt, download e upload). Embora, num primeiro momento, pareça que utilizar apenas o tempo e o identificador de interface tornaria possível individualizar uma observação, apenas essas duas variáveis não são suficientes. Os identificadores não existem durante todo o período de tempo, e uma interface pertencente a um local pode, por exemplo, trocar seu nome de *e0001* para *e1002*. Também ocorre que esses nomes podem se repetir, como se verificou no conjunto de dados, em que identificadores utilizados durante um período no Rio Grande do Sul foram repetidos posteriormente em São Paulo. É necessário, portanto, adicionar as informações de estado e local e, com as cinco variáveis, individualiza-se uma observação.

Para cada mês, é criado um *data frame* que é primeiro sanitizado, de forma que linhas contendo alguma variável com informação corrompida ou ausente sejam eliminadas,

Tabela 3 – Número de locais e interfaces por estado.

		Locais	Interfaces
Norte	AC	39	152
	AM	33	287
	AP	12	71
	PA	107	2725
	RO	24	669
	RR	16	122
	TO	28	510
Sudeste	ES	36	273
	MG	198	21101
	RJ	134	7175
	SP	98	4665
Sul	PR	70	2558
	RS	124	4850
	SC	257	23519

		Locais	Interfaces
Nordeste	AL	20	120
	BA	150	5985
	CE	106	959
	MA	61	2821
	PB	42	775
	PE	112	4550
	PI	41	720
	RN	70	394
	SE	21	169
Centro-Oeste	DF	211	5420
	GO	39	677
	MS	56	1200
	MT	32	527

para então o *dataframe* ser salvo no formato parquet[11], usando compressão gzip[12]. Dessa forma, o conjunto de dados é armazenado de maneira eficiente e pode ser carregado em apenas alguns segundos.

4.4 Caracterização dos dados

O conjunto de dados corresponde a mais de 500 milhões de observações, compreendendo o tempo de primeiro de novembro de 2020 até 30 de novembro de 2021 e contendo informação dos 26 estados brasileiros e do Distrito Federal. São 27 PoPs apresentando conexão com um total de 2137 locais em todo o território nacional. A Tabela 3 mostra como esses locais encontram-se distribuídos pelos PoPs, bem como a quantidade de identificadores únicos de interfaces que cada PoP apresentou durante o período de estudo.

Há uma grande variedade no número de locais que um determinado PoP se conecta. Alguns possuem poucas ligações, como Roraima e Rondônia, com 24 e 16 locais respectivamente. Outros apresentam muitas ligações, como Minas Gerais, com 198, e Santa Catarina, com 257. As regiões também apresentam grande diferença entre si, como Santa Catarina, no sul, que possui número similar à totalidade da região norte.

Com relação às interfaces, observou-se que mais de 90% dos locais possuem apenas uma em determinado momento no tempo. A quantidade de nomes para interfaces que um PoP apresenta durante um período mostra que é comum a mudança de nomes, e em alguns estados a frequência é muito maior que nos demais. Minas Gerais e Santa Catarina apresentaram mais de 20 mil identificadores de interface durante o período analisado. O

[11] Apache Parquet: <https://parquet.apache.org/>
[12] Gzip: <https://www.gnu.org/software/gzip/>

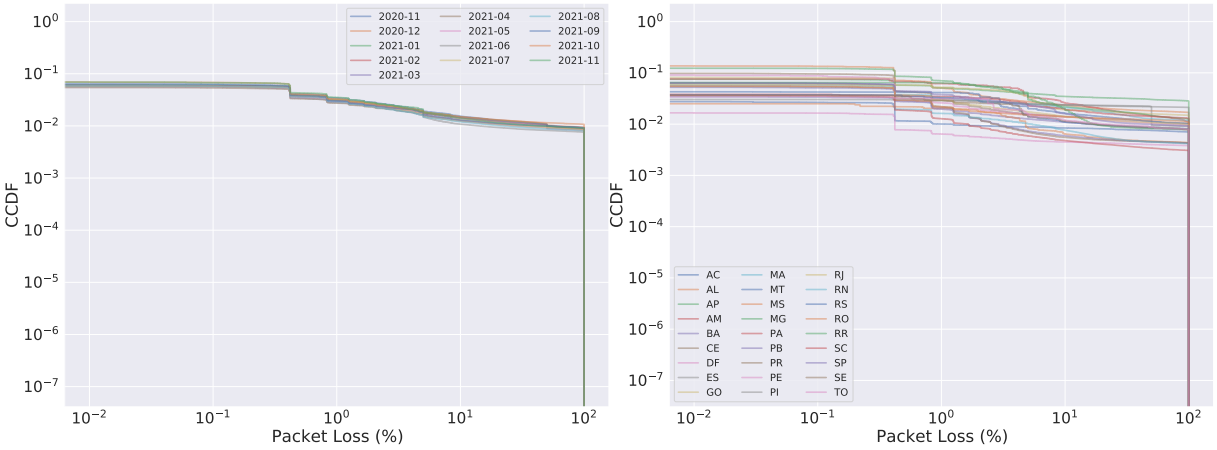

Figura 6 – Distribuições empíricas da variável *packet loss* ao longo do tempo e por unidade federativa.

motivo para essas mudanças não é conhecido e também não se sabe se são periódicas e planejadas. Tais indagações ainda precisam ser elucidadas em trabalhos futuros.

A Figura 6 apresenta as CCDFs (*Complementary Cumulative Distribution Function*)[13] da variável *packet loss* considerando todo o território nacional ao longo do tempo e considerando todo o tempo, mas apresentando a divisão geográfica. Quando observadas ao longo do tempo, as CCDFs não apresentam muita diferença, mantendo o mesmo padrão durante todo o período em análise. A ocorrência de perda de pacotes é rara: não há perda em mais de 90% do tempo. Perdas superiores a 1% ocorrem em apenas 5% das observações e perdas de 100% em apenas 1%.

Quando se considera a divisão geográfica, os estados brasileiros ainda apresentam um padrão próximo em suas CCDFs, mas suas probabilidades apresentam uma faixa de variação maior. Para a maioria dos estados, a existência de *packet loss* ainda é rara, com probabilidade inferior a 10%, mas, enquanto essa ocorrência é ainda mais rara em alguns estados, podendo chega a quase 1%, evidenciam-se também estados onde a ocorrência é de 20%.

Com relação ao RTT, a figura 7 apresenta a divisão da variável em relação ao tempo e sua divisão geográfica. As distribuições não aparentam diferenças significativas quando se considera seu comportamento no tempo. Apenas nos casos de valores superiores a 10^3 ms, nota-se uma diferença em suas probabilidades, mas essa diferença está presente apenas em probabilidades baixas, inferiores a 0.1%. Os valores de RTT em 90% dos casos são inferiores a 10 ms e em 99% dos casos inferiores a 100 ms.

Quando se realiza a divisão por estados, as distribuições de RTT deixam de

[13] Em português: Função de Distribuição Acumulada Complementar. A CCDF de uma variável aleatória X é dada por $P(X > x)$. Para um ponto (x_i, y_i) no gráfico, y_i representa a probabilidade de se encontrarem valores maiores que x_i no conjunto de dados. Nos gráficos apresentados neste trabalho, ambos os eixos encontram-se na escala logarítmica.

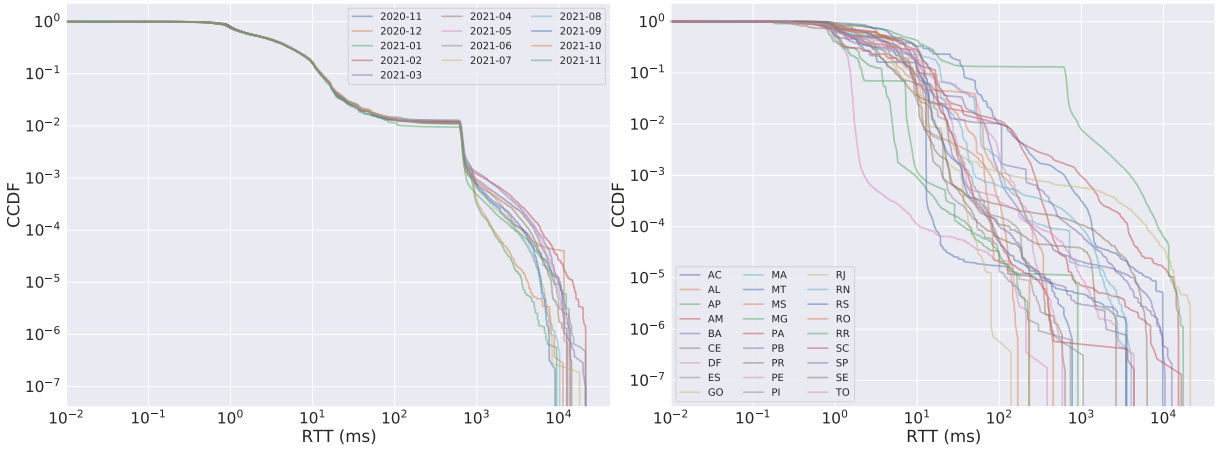

Figura 7 – Distribuições empíricas da variável RTT ao longo do tempo e por unidade federativa.

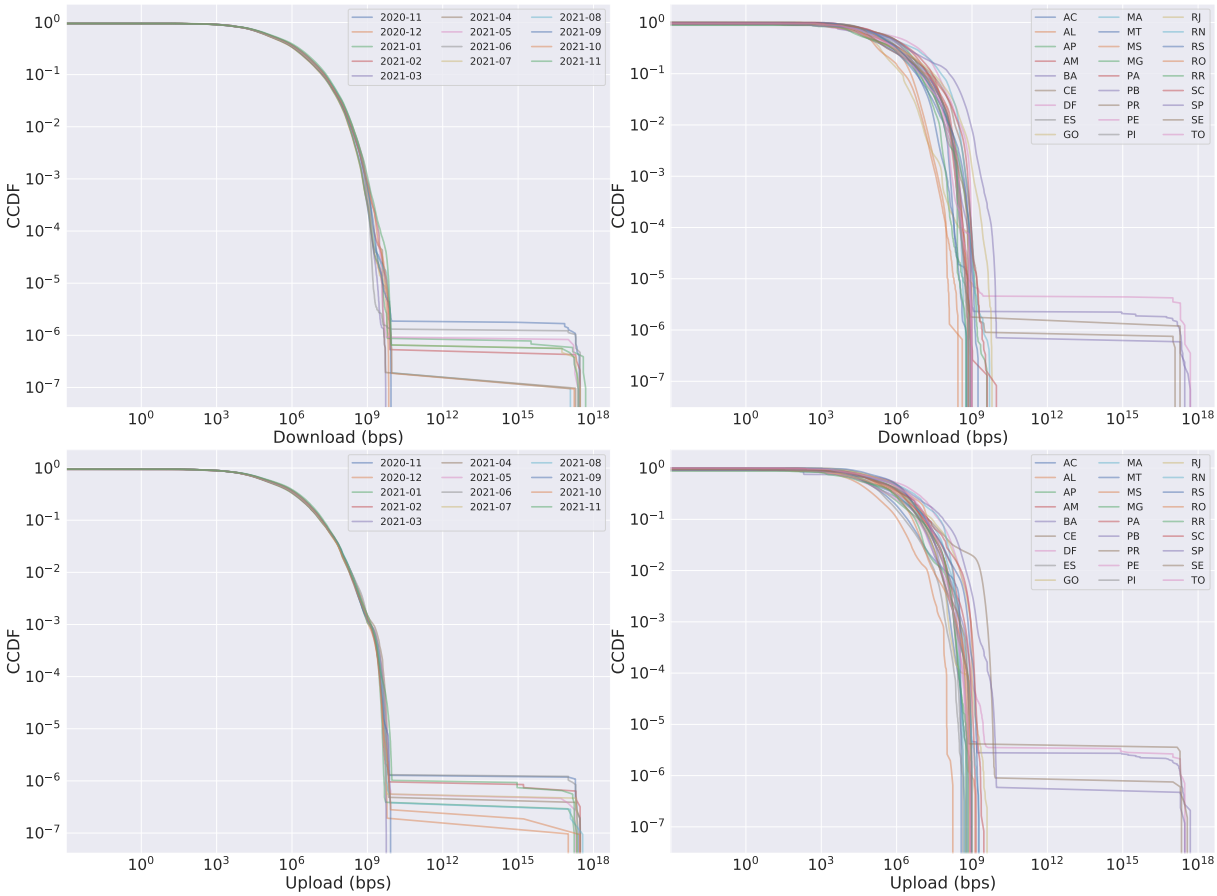

Figura 8 – Distribuições empíricas das variáveis download e upload ao longo do tempo e por unidade federativa.

apresentar um padrão uniforme em suas distribuições e cada localidade passa a apresentar uma curva distinta das demais. Valores de RTT superiores a 10 ms podem ser extremamente raros, com probabilidade de 0.001%, ou podem ser bem comuns, apresentando probabilidade próxima de 80%. Os valores mais comuns, com probabilidade de 90%, encontram-se então numa faixa de 1 a 100 ms.

Na Figura 8, encontramos as CCDFs das variáveis download e upload ao longo do tempo, considerando-se a localização geográfica. As duas variáveis apresentam distribuições semelhantes, com as taxas de transferência na faixa de Mbps e raros picos alcançando Gbps. Quando se considera apenas seu comportamento no tempo, as distribuições encontram-se muito próximas, apresentando apenas uma pequena variação nas probabilidades para valores superiores a 1 Gbps, sendo a diferença mais acentuada no caso de download. A divisão geográfica, por outro lado, torna explícita a diferença das distribuições nos estados brasileiros. Taxas superiores a 10^8 bps podem ser raras em alguns estados, com probabilidades próximas de 1%, ou podem ser comuns, com probabilidade superior a 90%.

Os dados também apresentam valores irreais, na ordem de 10^{18} bps. Esses valores são considerados anomalias e possuem causa desconhecida. Sua ocorrência é provavelmente relacionada a algum problema de medição ou processamento dos dados. Apesar de não ocorrerem durante todos os meses, são encontrados na maioria do período selecionado, e se limitam a uma pequena parte dos estados. Por apresentarem probabilidade extremamente baixa de ocorrência, da ordem de 10^{-6}, e por sua limitação geográfica, tais anomalias não são removidas do conjunto de dados e não devem apresentar problemas para os resultados futuros.

4.5 Ambiente de programação e reprodutibilidade

As ferramentas e os códigos produzidos para a realização deste trabalho utilizam a linguagem Python em sua versão 3.7. Foram utilizados Jupyter Notebooks[14] para a realização das análises, criação de gráficos e modelos. Os Notebooks utilizam a plataforma Google Colab[15] para sua execução, visando a facilitar a reprodução do trabalho. Outras opções para a execução dos notebooks criados são o Kaggle[16] e o Microsoft Azure[17].

O uso de notebooks com o Colab também permite a utilização gratuita (com algumas restrições de tempo de uso) de GPUs (*Graphical Processing Units*). Em (CARNEIRO et al., 2018), modelos de aprendizado de máquina com redes neurais profundas obtiveram resultados favoráveis quando comparados a modelos treinados utilizando apenas CPUs (*Central Processing Units*). Nesse trabalho, verificou-se que a utilização de GPUs chegou a alcançar ganho de até 5 vezes na diminuição do tempo de treinamento para alguns modelos.

Os notebooks e scripts utilizados encontram-se disponibilizados no GitHub[18]. Os

[14]Jupyter Notebook: <https://jupyter.org/>
[15]Google Colab: <https://colab.research.google.com/>
[16]Kaggle: <https://www.kaggle.com/>
[17]Microsoft Azure: <https://visualstudio.microsoft.com/pt-br/vs/features/notebooks-at-microsoft/>
[18]GitHub do trabalho: <https://github.com/VitorSpa/ViaIpe-Tools>

dados brutos referentes ao mês de novembro de 2021 encontram-se no Zenodo[19]. Documentação e informações adicionais estão disponíveis nos links apresentados.

[19]Zenodo do trabalho: <https://zenodo.org/record/5042650>

5 Caracterização de Falhas

Esta seção caracteriza as falhas ocorridas na Rede Ipê durante o período de análise de novembro de 2020 a novembro de 2021. A caracterização ocorre primariamente através da análise de distribuições de probabilidade, mas outros tipos de visualização também são utilizados para auxiliar o entendimento do fenômeno. A definição do que é uma falha no contexto deste trabalho encontra-se na Seção 5.1. Em seguida, a Seção 5.2 apresenta a caracterização e a análise do comportamento das falhas, começando pelo nível nacional (5.2.1), seguido do nível regional (5.2.2) e, finalmente, apresentando a caracterização de alguns estados selecionados (5.2.3).

5.1 Definição de falha

Em (MARTINELLO, 2005), uma falha num serviço de conectividade é definida como uma transição do estado de um serviço que está funcionando de forma correta (ou esperada) para um estado em que não mais esteja. No caso de um serviço de conectividade entre entidades x e y, uma falha ocorre quando a comunicação entre x e y degrada além de um limiar de qualidade pré-estabelecido ou, no pior caso, fica totalmente impossibilitada.

Formalmente, seja $s_t^T = [v_1, v_2, ..., v_n]$ um vetor representando o estado do serviço de conectividade entre x e y no tempo t. A composição de s_t contém variáveis v_i que podem ser referentes tanto a informações sobre o serviço em si (e.g. operadora, tecnologia e localização) quanto a métricas de desempenho e qualidade. Assim, é possível afirmar que ocorreu uma falha no serviço de conectividade entre x e y no tempo t quando: s_t apresenta alguma violação em suas variáveis conforme os requisitos pré-estabelecidos (e.g. interrupção total da transmissão de dados ou perda de pacotes suficientemente alta) e s_{t-1} apresentava o sistema funcionando de maneira correta. É importante notar que a falha ocorre apenas na mudança do estado do serviço. Dois momentos em sequência, s_t e s_{t+1}, ambos violando os limites estabelecidos, não representam duas falhas, mas sim uma falha que se estende ao longo do tempo durante as duas medições.

A definição de falha num serviço de conectividade neste trabalho segue, então, diretamente a caracterização da qualidade de serviço utilizada pela ferramenta Via Ipê para estabelecer o limite de degradação aceitável. Para a ferramenta, a qualidade é dividida em quatro categorias, definidas em função do valor da variável *packet loss*. Para um estado de serviço s_t com valor de *packet loss* v_i, a qualidade $Q(s_t)$ em um dado momento é definida como:

$$Q(s_t) \Rightarrow \begin{cases} \text{Ótima, caso } v_i \leq 0.01\% & (5.1a) \\ \text{Boa, caso } 0.01\% < v_i \leq 1\% & (5.1b) \\ \text{Regular, caso } 1\% < v_i \leq 3\% & (5.1c) \\ \text{Ruim, caso } v_i > 3\% & (5.1d) \end{cases}$$

Para transformar em uma classificação binária, este trabalho define o estado de falha[1] $F(s_t)$ como o período em que o *packet loss* é superior a 3%, de forma que:

$$F(s_t) \Rightarrow \begin{cases} \text{Não-falha, caso } v_i \leq 3\% & (5.2a) \\ \text{Falha, caso } v_i > 3\% & (5.2b) \end{cases}$$

A representação dos dados descrita na seção anterior (Seção 4) está relacionada com a definição apresentada da seguinte forma: considera-se uma observação x_i do conjunto de dados, essa observação representa um estado s_t entre um PoP e um local num determinado momento no tempo. Dessa forma, basta que se verifique o valor de *packet loss* de cada linha de um *Data Frame* para se criar uma nova variável indicando se é falha ou não.

Outros limiares também poderiam ser escolhidos. Por exemplo, os casos extremos de perda de pacotes, considerando o caso mais restritivo, em que uma falha ocorre com a presença de qualquer perda de pacote, ou o caso menos restritivo em que uma falha ocorre apenas nos casos de 100% de perda. Essa variação dos limites de degradação apresenta-se como objeto de estudo para trabalhos futuros.

5.2 Comportamento das falhas

Ainda seguindo a definição de qualidade de um serviço de conectividade em 5.1, a Figura 9 apresenta a variação do serviço ao longo do tempo em interfaces conectadas a um mesmo PoP durante a primeira semana de 2020. É possível verificar diferentes aspectos sobre a dinâmica em uma mesma interface e em relação a interfaces distintas. Observa-se que há vários perfis de interfaces: a Interface 1 raramente chega ao estado de falha, mas costuma variar sua qualidade nos valores intermediários com certa periodicidade; a Interface 2 apresenta poucas falhas, mas uma delas é bem longa se comparada às demais; a Interface 3 mantém-se estável na categoria ótima; a Interface 4, por sua vez, falha frequentemente; e, por fim, a Interface 5 apresenta poucas falhas, falhas curtas e pequenas variações intermediárias de qualidade, sem um padrão aparente.

Com relação às falhas, além do tamanho e da frequência variável já apresentados, percebe-se também que elas podem apresentar correlação espacial, como evidenciado pela

[1] Neste trabalho, o estado de falha pode ser também expresso como 1 para falha e 0 para não-falha.

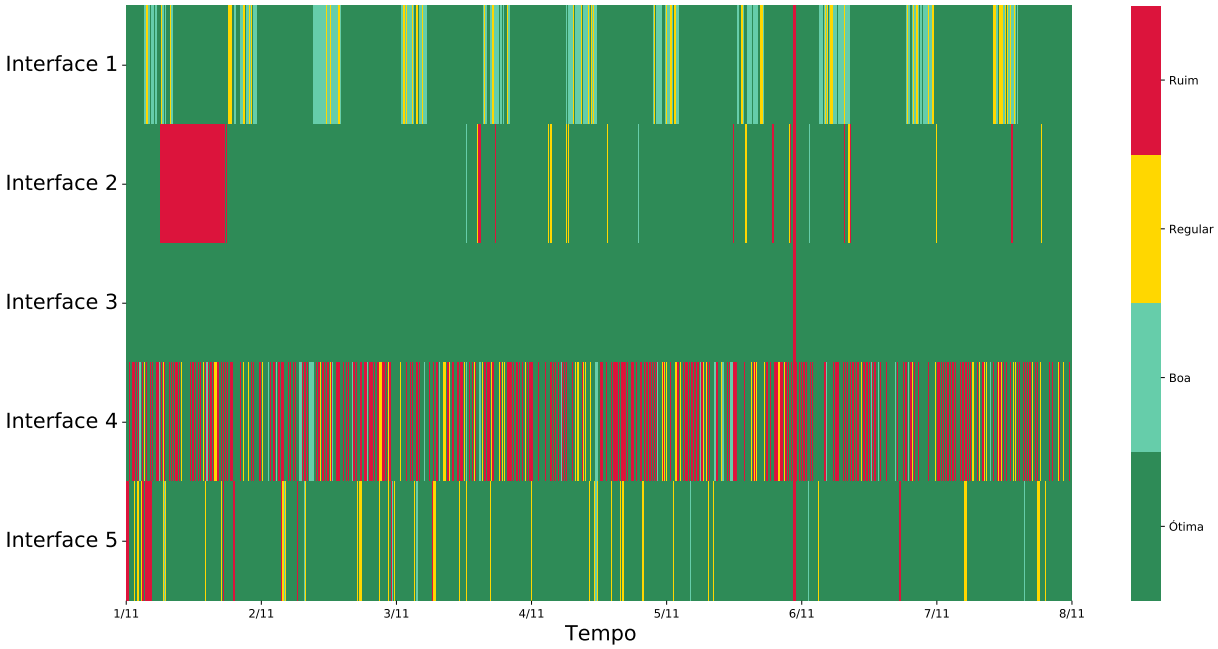

Figura 9 – Qualidade do serviço de conectividade em interfaces ao longo do tempo.

linha vermelha presente em todas as interfaces próximas ao dia 6 de novembro. Algumas falhas ocorrem de forma repentina, quando a mudança é direta de um estado ótimo para um estado ruim. Outras falhas ocorrem de maneira gradual, com a qualidade degradando com o tempo e passando pelos valores intermediários (bom e regular, em verde-claro e amarelo, respectivamente) até chegar ao estado de falha, quando a qualidade é classificada como ruim. Tais diferenças entre comportamentos indicam que a tarefa de predição apresentada na seção seguinte pode ser simples em alguns casos e inviável em outros.

Essa diferença no comportamento das interfaces e a diferença na distribuição das variáveis do conjunto de dados descrita na seção anterior motivam a divisão realizada neste capítulo, que apresenta o estudo das falhas considerando todo o território brasileiro, a divisão regional do país e, por fim, alguns estados selecionados individualmente.

A caracterização do comportamento das falhas dá-se novamente através do uso de CCDFs para a visualização das distribuições de probabilidades de quatro variáveis de interesse: tempo entre falhas, duração da falha, número de falhas e tempo em falha por interface. Com isso, busca-se responder questões relativas às seguintes métricas de dependabilidade dos serviços de conectividade: confiabilidade, disponibilidade e manutenibilidade. As distribuições apresentadas consideram as interfaces em que ocorreram ao menos uma falha.

5.2.1 Comportamento nacional

A Figura 10 mostra as distribuições de falhas num contexto nacional durante todo o período em análise. Com relação à duração, 90% das falhas em território nacional são

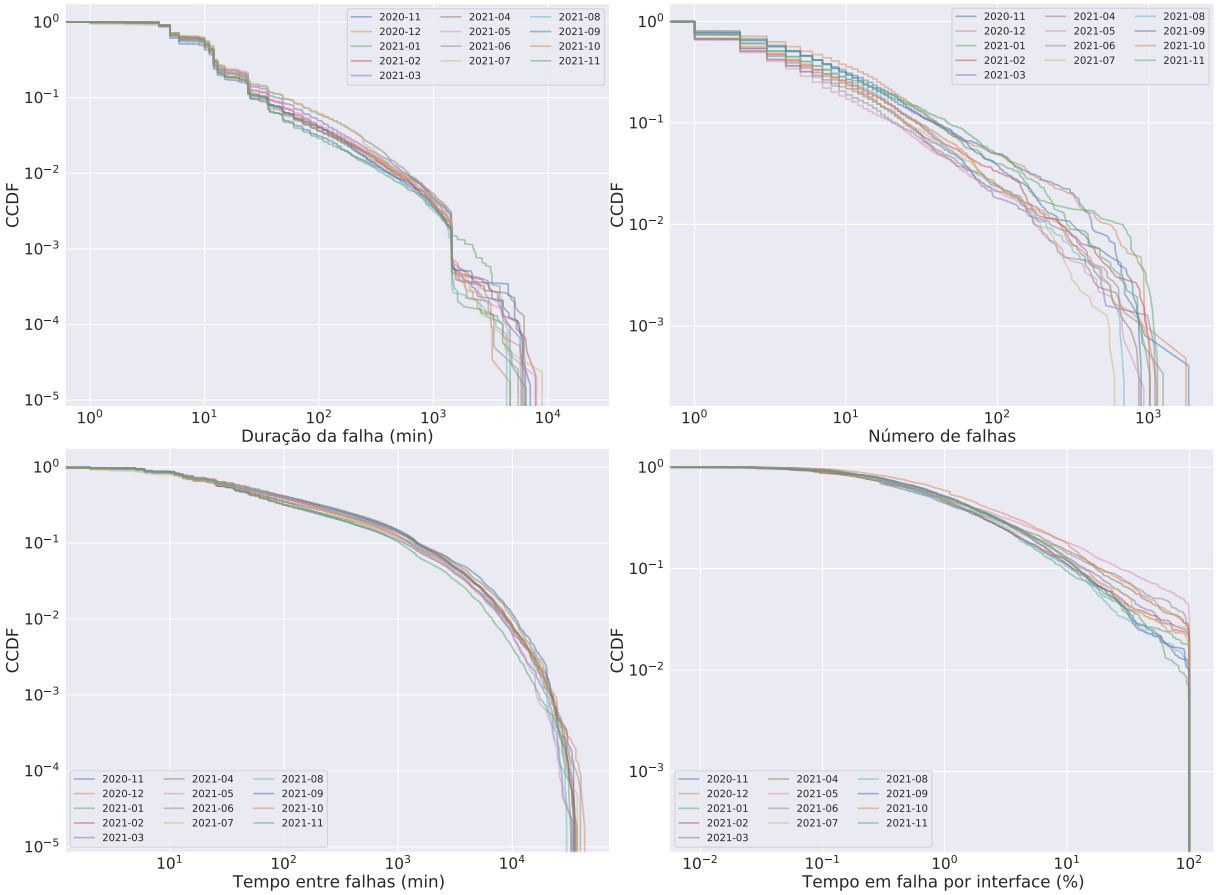

Figura 10 – Distribuições empíricas das falhas ao longo do tempo considerando todo o território nacional.

inferiores a 20 minutos e, para falhas com essa duração, a distribuição de probabilidade não apresenta muita mudança ao longo do tempo. Falhas superiores a tamanhos próximos de 100 minutos apresentam diferenças mais significativas em suas probabilidades dependendo do mês, e sua ocorrência varia numa faixa de 1 a 10%. Para valores acima de 1000 minutos, a diferença entre os meses mostra-se ainda maior, mas suas probabilidades de ocorrência são, em geral, inferiores a 0.1%.

Com relação ao número de falhas, a figura mostra que, em interfaces onde ocorrem falhas, houve uma variação considerável dependendo do mês: interfaces com até 10 falhas, com uma representação que pode variar de 40 a 80%. A diferença em probabilidades também aumenta com o número de falhas, e as distribuições para interfaces com número de falhas superior a 100 são ainda mais distintas.

O tempo entre falhas é a variável que apresentou menor influência com relação às demais. As distribuições só apresentam variação significativa para valores acima de 100 minutos. A ocorrência de tempo entre falhas, ou seja, o tempo de funcionamento normal ininterrupto, superior a 1000 minutos (aproximadamente 17 horas), apresentou uma variação na faixa de 10 a 30%.

A proporção do tempo que uma interface fica em estado de falha é alta, 60 a

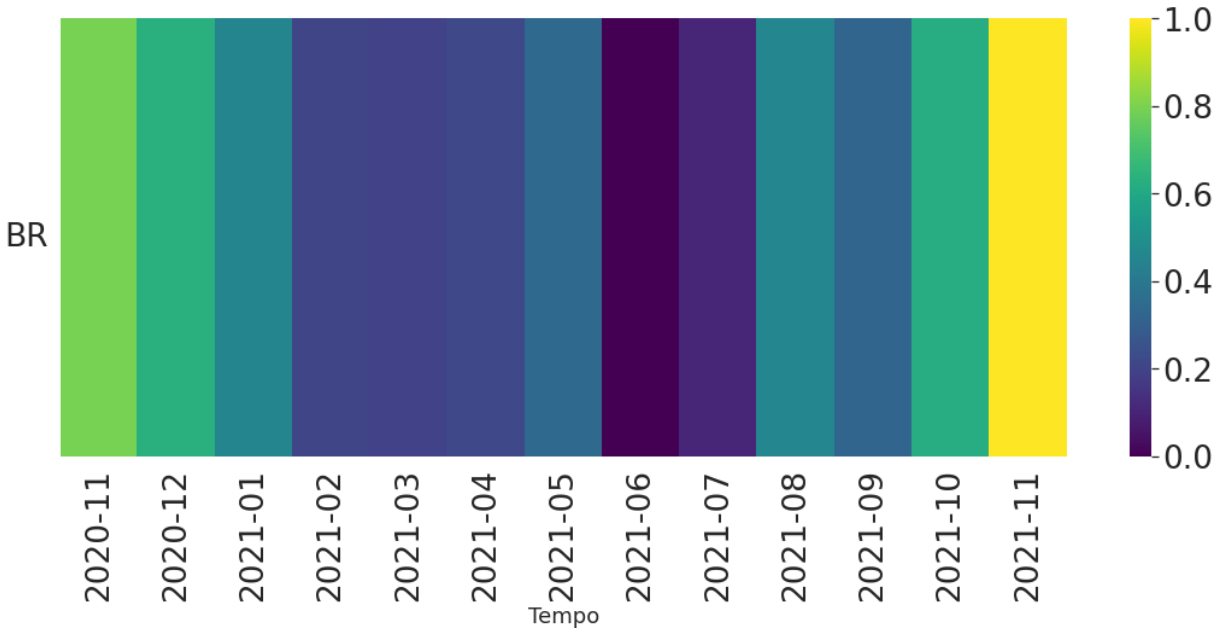

Figura 11 – Número de falhas ao longo do tempo considerando todo o território nacional.

70% das interfaces num mês apresentaram uma disponibilidade superior a 99%. Variações significativas ocorrem ao longo do tempo para valores superiores a 0.5% de tempo em falha. Apenas as regiões Sudeste e Sul apresentaram interfaces que ficaram em estado de falha 100% de sua existência.

Resumindo: falhas, quando ocorrem, costumam ser curtas; poucas interfaces apresentam um grande número de falhas; o tempo ininterrupto de operação normal numa interface é baixo; e o percentual do tempo em falha costuma ser alto.

Com relação ao número absoluto de falhas ocorridas no território nacional, a Figura 11 apresenta um mapa de calor exibindo o valor normalizado ao longo do tempo. São identificados dois picos de falha: o primeiro ocorre em novembro de 2020 e, gradativamente, a ocorrência de falhas diminui até julho, chegando ao mínimo; em seguida, ocorre um aumento que leva ao segundo pico, em novembro de 2021.

5.2.2 Comportamento regional

Após a identificação dos meses com comportamento extremo, as CCDFs para as regiões consideram apenas os meses de novembro de 2020 e junho de 2021, para evitar ruído nos gráficos e facilitar a compreensão.

A Figura 12 apresenta as CCDFs para o mês de Novembro de 2020, um dos picos de falhas no Brasil. Para a duração das falhas, as regiões Centro-Oeste, Norte e Sul apresentam valores menores que a curva de probabilidade nacional. Falhas nessas regiões costumam ser curtas, com apenas 1% das falhas apresentando duração superior a 30 minutos. As falhas mais longas estão presentes nas regiões Nordeste e Sudeste, onde a probabilidade

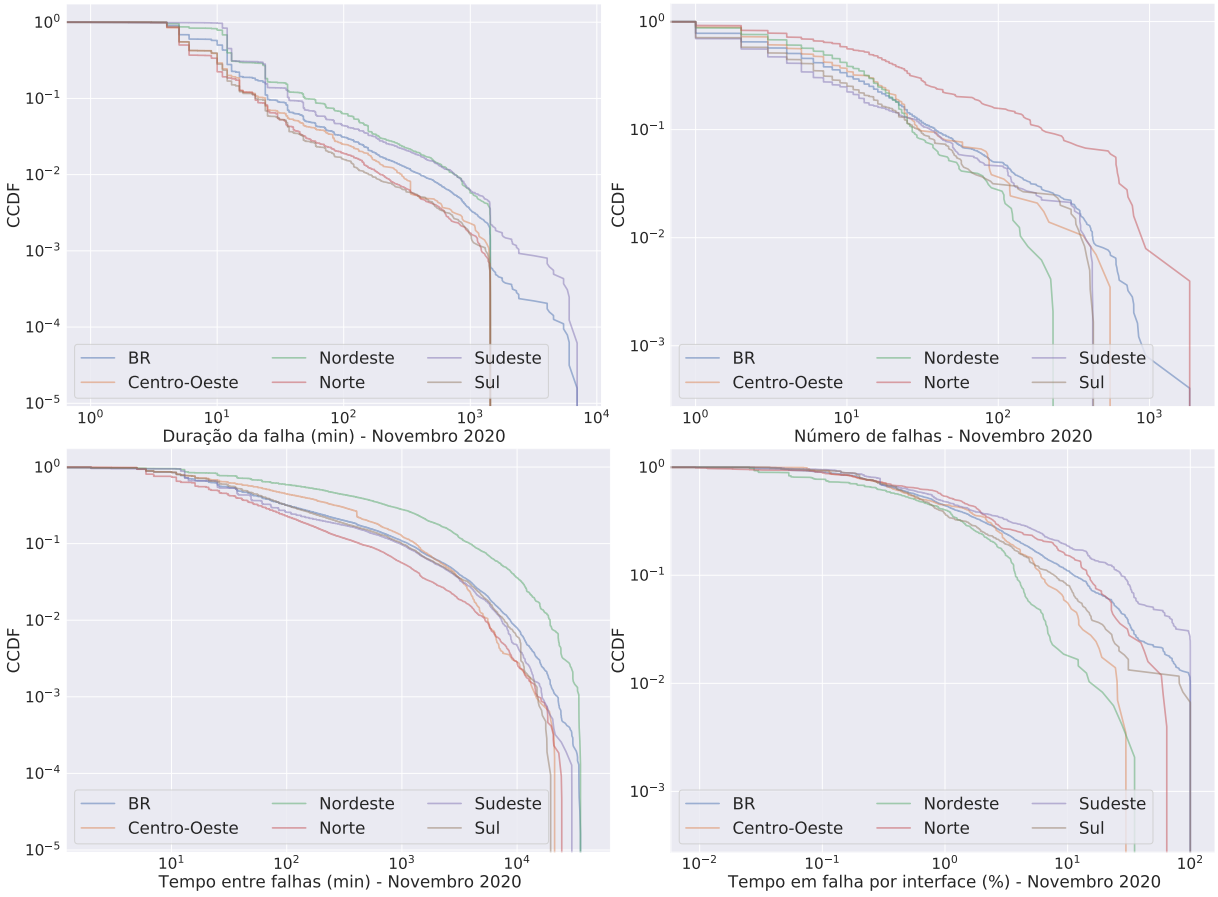

Figura 12 – Distribuições empíricas das falhas no mês de Novembro de 2020 divididas por região.

de ocorrência de falhas superiores a 50 minutos é superior a 50%. Falhas próximas a 10^4 minutos (quase 7 dias) ocorrem apenas na região Sudeste, com frequência baixa e valores próximos a 0.01%. O fato de todas as regiões, exceto a Sudeste, apresentarem igual duração máxima sugere que o evento causador possa ser do mesmo tipo, ou até o mesmo evento.

Com relação ao número de falhas, o Norte se destaca, apresentando uma curva com significativa dissimilaridade das demais regiões. A probabilidade de se encontrarem interfaces com mais de 100 falhas no Norte é de aproximadamente 20%, enquanto nas demais regiões fica próxima de 1%. O Norte e Sudeste apresentam as interfaces que superam mil falhas, enquanto nas demais regiões os maiores valores de falhas por interface variam de 300 a 700.

Quanto ao tempo entre falhas, apenas a região Nordeste apresentou valores que podem ser considerados bons. Para essa região, 70% das interfaces apresentam tempo de operação superior a mil minutos. Nas demais regiões, esse valor fica próximo de 30%, e no pior caso está a região Norte, com apenas 1%. Conclui-se que, no geral, nas interfaces em que ocorrem falhas, o tempo normal de operação é baixo.

Considerando o percentual de tempo em falha por interface, para valores inferiores a 1%, as curvas se mantêm próximas, com probabilidades entre 60 a 70%. Para valores

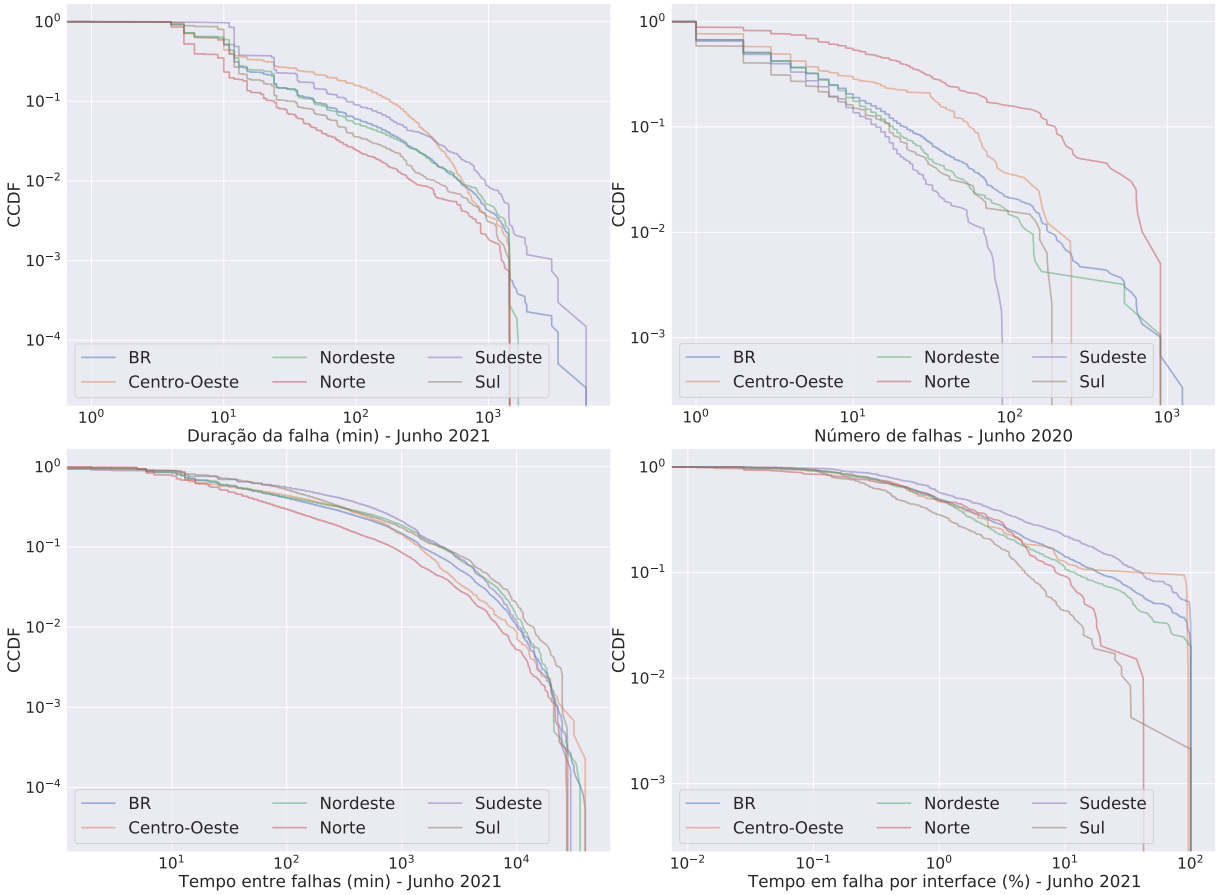

Figura 13 – Distribuições empíricas das falhas no mês de Junho de 2020 divididas por região.

superiores, nas regiões Norte e Sudeste o percentual de tempo é maior, ocorrendo valores acima de 10%, com probabilidade de 10 a 30%. A região Nordeste apresenta resultados melhores: valores acima de 10% são raros na região, e a probabilidade é próxima de 0.5%.

A Figura 12 apresenta as CCDFs para Junho de 2021, mês com a menor quantidade de falhas no período em análise. A região Norte continua apresentando falhas curtas, com apenas 10% de falhas superiores a 10 minutos. A região Centro-Oeste passa a apresentar falhas mais longas, com maiores probabilidades de ocorrência para falhas de 10 a 1000 minutos. A região Sudeste manteve-se como a única com falhas próximas de 10^{14} minutos. O mesmo tipo de ocorrência do mês em análise anterior aparenta apresentar-se novamente. A maioria das regiões demonstram o mesmo valor de duração, com a região Nordeste contendo ocorrências um pouco maiores.

A região Norte continua apresentando alto número de falhas por interface. Em 10% de suas interfaces, o número e falhas é superior a 500. A região Centro-Oeste agora apresenta valores acima da curva nacional, com 10% de suas interfaces demonstrando valores superiores a 100 falhas. Para o tempo entre falhas, a região Norte continua apresentando baixo tempo contínuo de operação normal. Apenas 20% de suas interfaces apresentam tempo contínuo de operação superior a 1000 minutos. As curvas das demais

Tabela 4 – Comparação entre desempenhos das regiões em relação à curva nacional.

	Novembro/2020		**Junho/2021**	
	Melhor Caso	Pior Caso	Melhor Caso	Pior Caso
Duração da Falha	Norte	Sudeste, Nordeste	Norte	Sudeste
Número de Falhas	Nordeste	Norte	Sudeste	Norte
Tempo entre Falhas	Nordeste	Norte	Sudeste	Norte
Tempo em Falha	Nordeste	Sudeste, Norte	Sul	Sudeste

Figura 14 – Número de falhas por local ao longo do tempo para cada região.

regiões encontram-se bem próximas, variando de 50 a 70% das interfaces com tempo contínuo superior a 1000 minutos. Ainda assim, são valores baixos, mas é importante lembrar que essas são as interfaces em que ocorrem falhas.

Durante esse período, a região sul apresenta baixo percentual de tempo em falha por interface. Metade de suas interfaces apresentam percentual superior a 1%, e menos de 1% de suas interfaces apresentam percentual superior a 10%. As demais regiões têm 60 a 80% de suas interfaces apresentando percentual superior a 1%, e 10 a 50% de suas interfaces apresentando percentual superior a 10%.

Um resumo da análise realizada nos dois meses, comparando os melhores e piores casos das regiões, encontra-se na Tabela 4. Apesar de aparecer como o melhor caso na duração de falhas, a região Norte apresenta alto número de falhas, baixo tempo contínuo de operação normal e elevado percentual de tempo em falha, o que a torna uma região seja de grande interesse para o estudo das falhas. A região Nordeste, por sua vez, apresenta falhas longas no mês de novembro, mas costuma ter um desempenho acima da média nacional nas métricas estudadas.

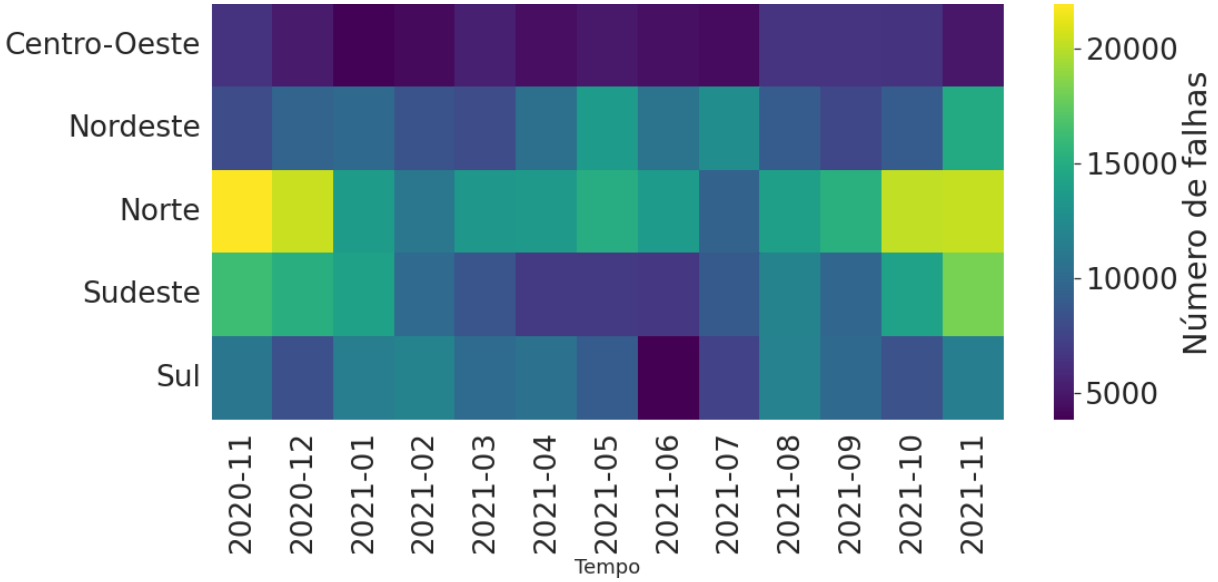

Figura 15 – Número de falhas ao longo do tempo para cada região.

A Figura 14 apresenta uma comparação do número absoluto de falhas por região ao longo do tempo, normalizado pela quantidade de locais de uma região (Tabela 3). Quando se considera a quantidade de locais, e não apenas o número de falhas, a única região que se destaca é a Norte, por se tratar de uma região com poucos locais e muitas falhas. Nas demais regiões, os meses não tiveram grande influência na fração de falhas por local, mas a região Norte apresenta o padrão detectado quando se observa o território nacional, com picos de falha nos meses próximos a Novembro.

A Figura 15 apresenta o número absoluto de falhas por região. É possível observar que, em números absolutos, outras regiões também apresentaram os picos próximos ao mês de novembro e, portanto, o padrão não ocorreu exclusivamente na região Norte. Percebe-se também que a região Sudeste é responsável por grande parte das falhas. A região Centro-Oeste manteve um número baixo de falhas durante todo o período, enquanto a região Sul apresentou baixo número de falhas apenas no mês de Junho de 2021.

5.2.3 Comportamento estadual

A figura 16 apresenta a fração de falhas em escala por estado, considerando todo o período de tempo em análise no mapa do Brasil. Na Figura, é possível identificar os estados do Amazonas, Roraima, Mato Grosso e Sergipe como os principais casos de falha proporcional (considerando o número de locais de um estado) do país. O resto do país não apresenta tantas falhas por interface, com algumas exceções aparecendo em cada região.

A Figura 17 apresenta a fração de falhas por local em cada estado brasileiro. A partir da imagem, é possível observar alguns estados que se destacam, como Roraima, na região Norte, que apresenta um número elevado de falhas nos primeiros meses de análise, de novembro de 2020 até março de 2021. O estado do Amazonas, também na região norte,

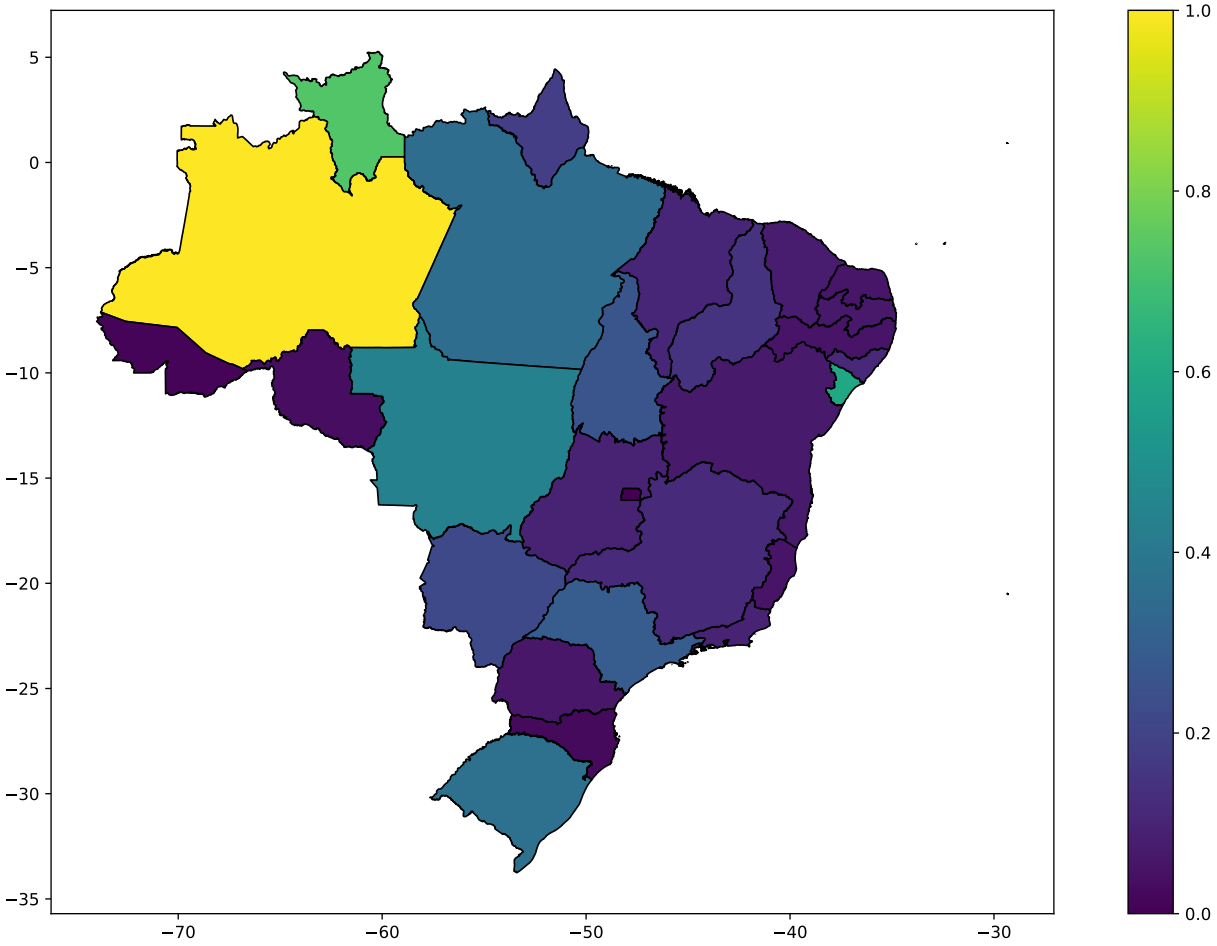

Figura 16 – Fração de falhas por estado, em escala.

manteve um valor alto de fração de falhas por local durante todo o período, mas o seu pico foi no mês de novembro de 2021. Sergipe, ao contrário dos demais estados, concentra suas falhas nos meses de maio e junho de 2021.

É possível perceber que, em geral, as regiões não são homogêneas. No Centro-Oeste, as falhas concentram-se nos estados de Mato Grosso do Sul e Mato Grosso. A região Nordeste é a mais homogênea: os estados apresentam pouca fração de falha, com exceção de Sergipe e seu período de falhas no meio do ano. A região Norte apresenta estados com muitas falhas, como Roraima, Amazonas, Amapá e Pará, mas também estados com poucas falhas, como Acre, Rondônia e Tocantins. Na região Sudeste, o estado de São Paulo apresenta uma fração de falhas mais elevada durante os períodos de pico do país, e os demais estados da região mantêm níveis baixos. O Rio Grande do Sul é o único estado da região Sul com frações mais elevadas, seguindo o padrão nacional anteriormente encontrado.

O Brasil é um país de grandes dimensões, por isso a análise de distribuições de falhas através de CCDFs será limitada a apenas alguns estados durante os meses de interesse: novembro de 2020 e 2021 e junho de 2021. Busca-se, dessa forma, cobrir diferentes perfis estaduais para, em conjunto com as análises anteriores, atingir um entendimento mais

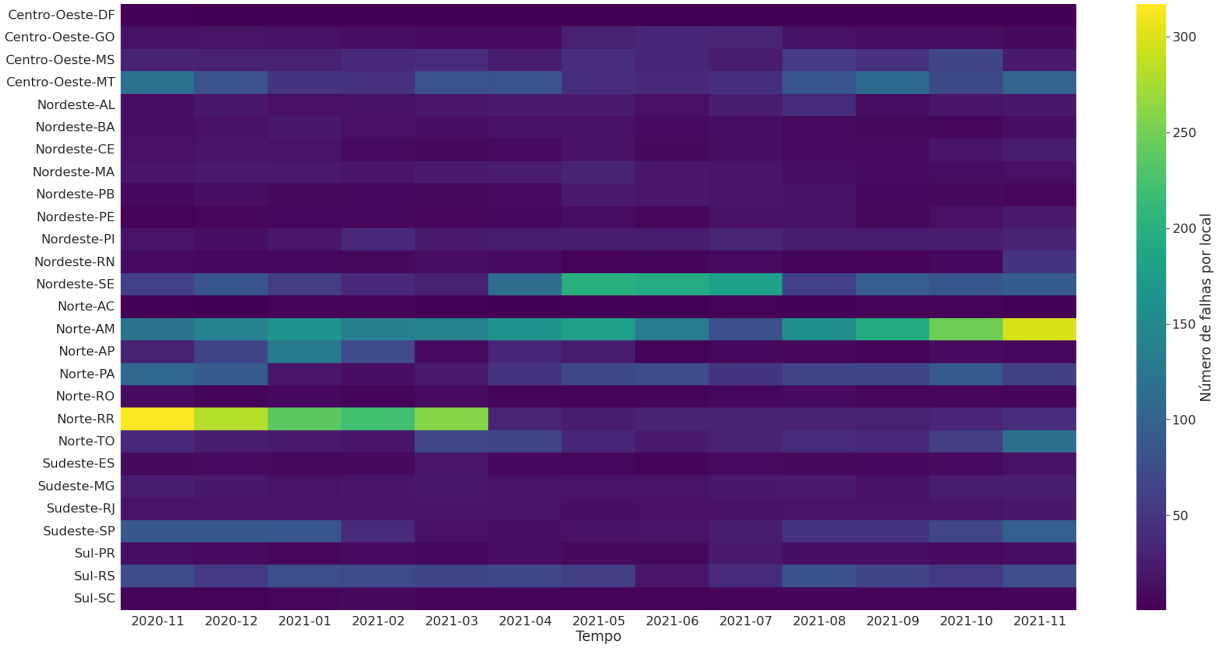

Figura 17 – Número de falhas ao longo do tempo para cada estado.

completo das falhas no território nacional. A escolha dos estados segue os critérios de número de falhas e dimensão (número de locais e interfaces). Os estados selecionados são, então: São Paulo, Minas Gerais, Roraima, Amazonas e Pará.

São Paulo

São Paulo é um estado que apresenta número alto de locais e interfaces, 98 e 4665, respectivamente. Na Figura 15, observa-se que o Sudeste apresenta comportamento semelhante ao nacional e, na Figura 17, observa-se que o estado de São Paulo também apresenta comportamento semelhante, com picos (embora menores) da fração de falha nos meses de novembro de 2020 e 2021 e valores menores no mês de junho de 2021.

A Figura 18 apresenta as CCDFs relativas às suas falhas. A duração das falhas não apresenta grande diferença em probabilidades nos meses analisados para valores inferiores a 50 minutos. Os meses com maior número de falhas apresentam probabilidades menores para valores altos. Falhas com duração superior a 100 minutos são menos frequentes nesses meses, com 1% de probabilidade de ocorrência. Já no mês de junho, a ocorrência é de cerca de 7%.

O número de falhas por interface é mais alto no mês de novembro de 2021, como previsto, apresentando 60% de interfaces com 10 ou mais falhas. O mês de novembro de 2021 curiosamente apresenta chances próximas ao mês de junho, quando ocorreram menos falhas. Isso sugere que, nesse período, as falhas, em geral, encontram-se mais espalhadas pelas interfaces do estado. Interfaces com mais de 10 falhas representam apenas 20% dos casos. O mês, em conjunto do mês de novembro de 2021, ainda apresenta algumas interfaces com muitas falhas, com um número próximo a 100 ocorrências. Já o mês de

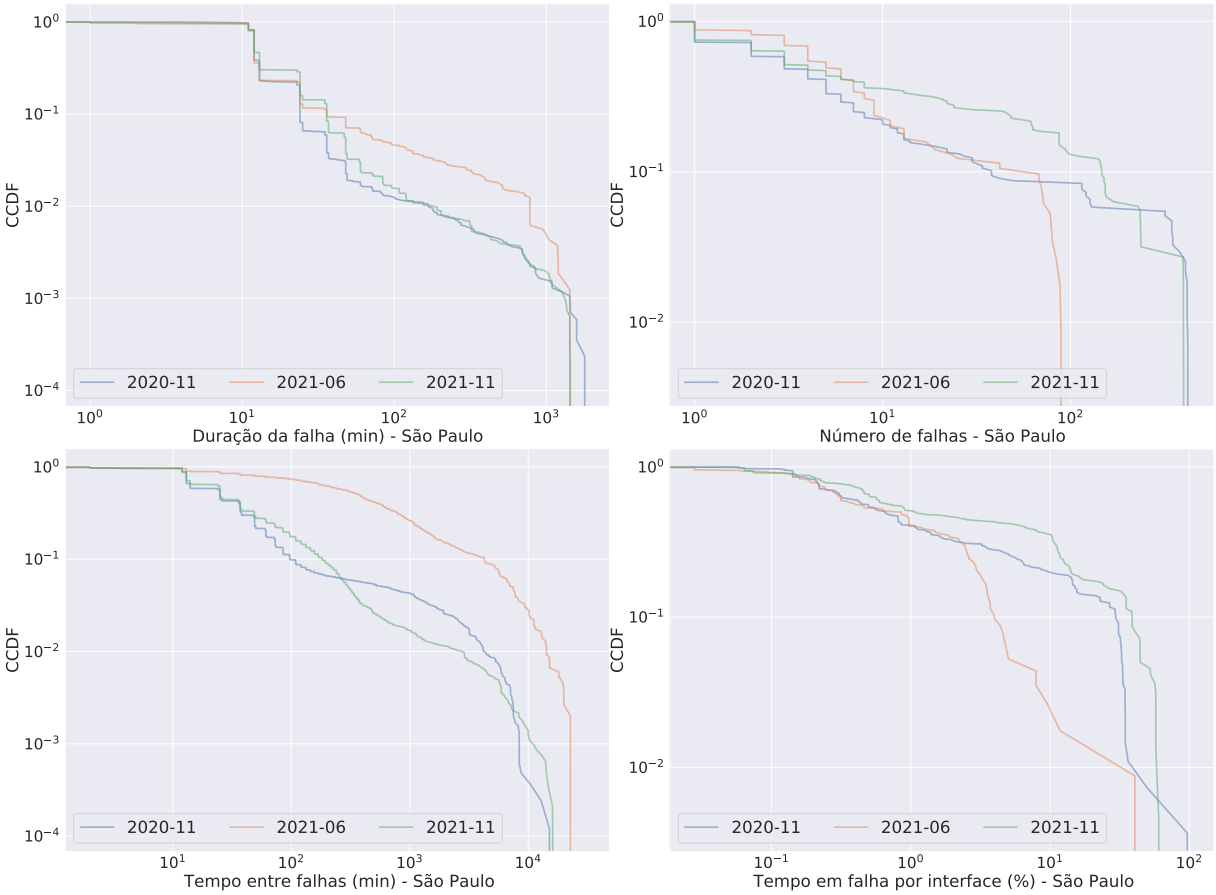

Figura 18 – Distribuições empíricas de São Paulo

junho limita-se a falhas com valor máximo próximo a 100 minutos.

Com relação ao tempo entre falhas, os valores para os meses de novembro são baixos, o percentual de interfaces com tempo contínuo de operação normal superior a 1000 minutos é de 7% para 2020 e de 3% para 2021. Esse valor é bem mais alto para junho, que atingiu 50%, mas ainda é um valor considerado baixo do ponto de vista operacional.

A fração do tempo em falha por interface é próxima nos três meses para valores inferiores a 1%. Para uma fração de 10%, as curvas divergem, com o mês de junho apresentando um valor baixo, de apenas 3% de ocorrência, e os meses de novembro de 2020 e 2021 apresentando valores altos, de 50 e 70%, respectivamente.

Minas Gerais

Minas Gerais é um dos estados com o maior número de locais e interfaces, 198 e 21101, respectivamente. A Figura 17 mostra que o estado não apresenta grande variação em sua fração de falhas no período em análise, mantendo o valor da fração de falhas por local próximo de 50, o que o difere da região Sudeste, que segue o padrão nacional demonstrado na Figura 11.

A Figura 19 apresenta as CCDFs relativas às suas falhas. Com relação à duração

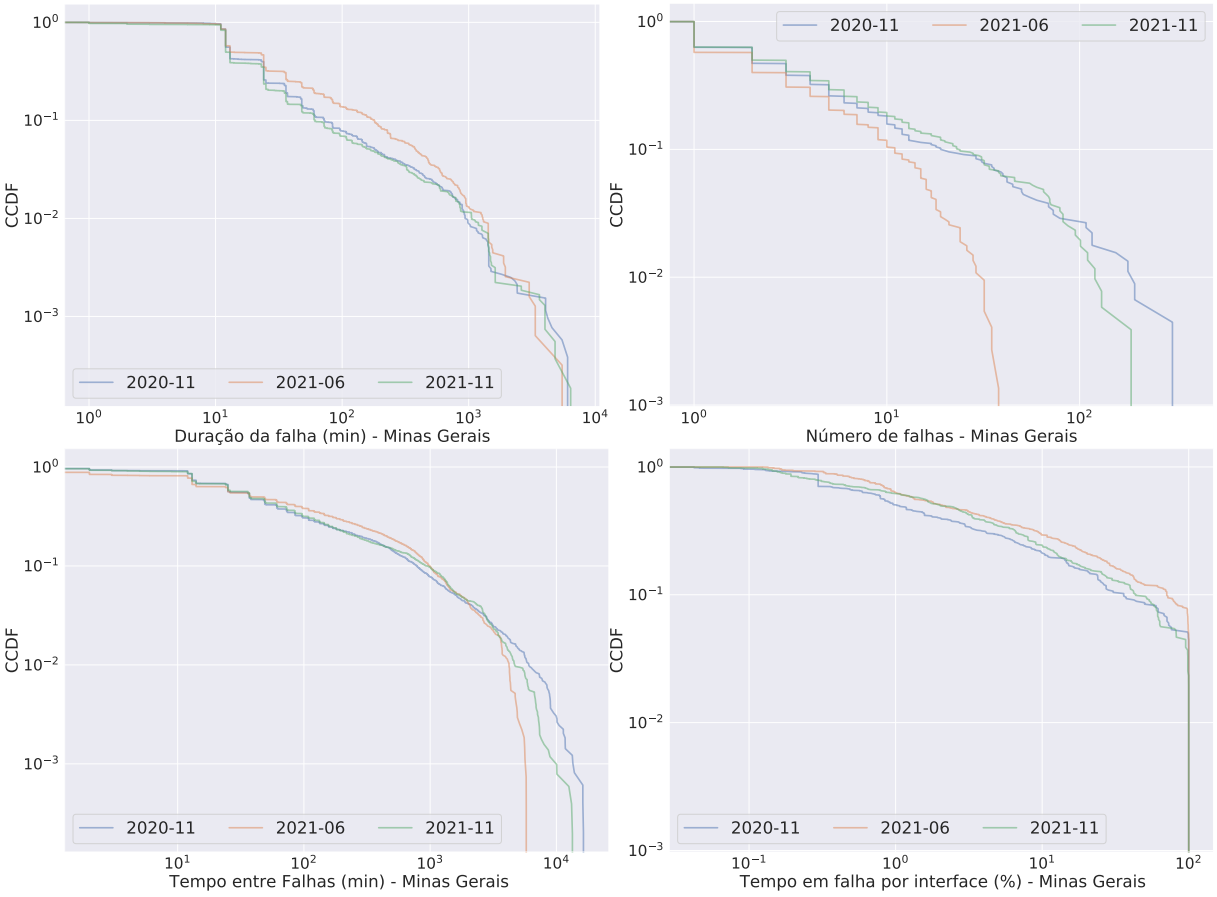

Figura 19 – Distribuições empíricas de Minas Gerais.

das falhas, as curvas são próximas, exceto no mês de junho, que apresenta probabilidades para valores entre 50 e 100 minutos. Enquanto nos meses de novembro a probabilidade de uma falha superior a 100 minutos é de aproximadamente 10%, no mês de junho ela se encontra próxima de 30%.

O Mês de junho apresenta um número máximo de falhas por interface menor que os demais, com o máximo de 70, enquanto os meses de novembro apresentam valores superiores a 300 e 600, respectivamente.

As curvas para o tempo entre falhas do estado são próximas, divergindo consideravelmente apenas para valores superiores a 9000 minutos (aproximadamente 6 dias). A probabilidade para o tempo de operação contínuo superior a 1000 minutos fica próxima de 7%, o que demonstra um desempenho melhor que o nacional, que apresentou valores próximos de 30%.

Com relação ao tempo percentual em falha por interface, as curvas também são próximas, com o mês de novembro de 2020 apresentando probabilidades pouco menores e o mês de junho probabilidades maiores. Percentuais superiores a 1% variam de 60 a 70%, e superiores a 10% variam de 30% a 50%.

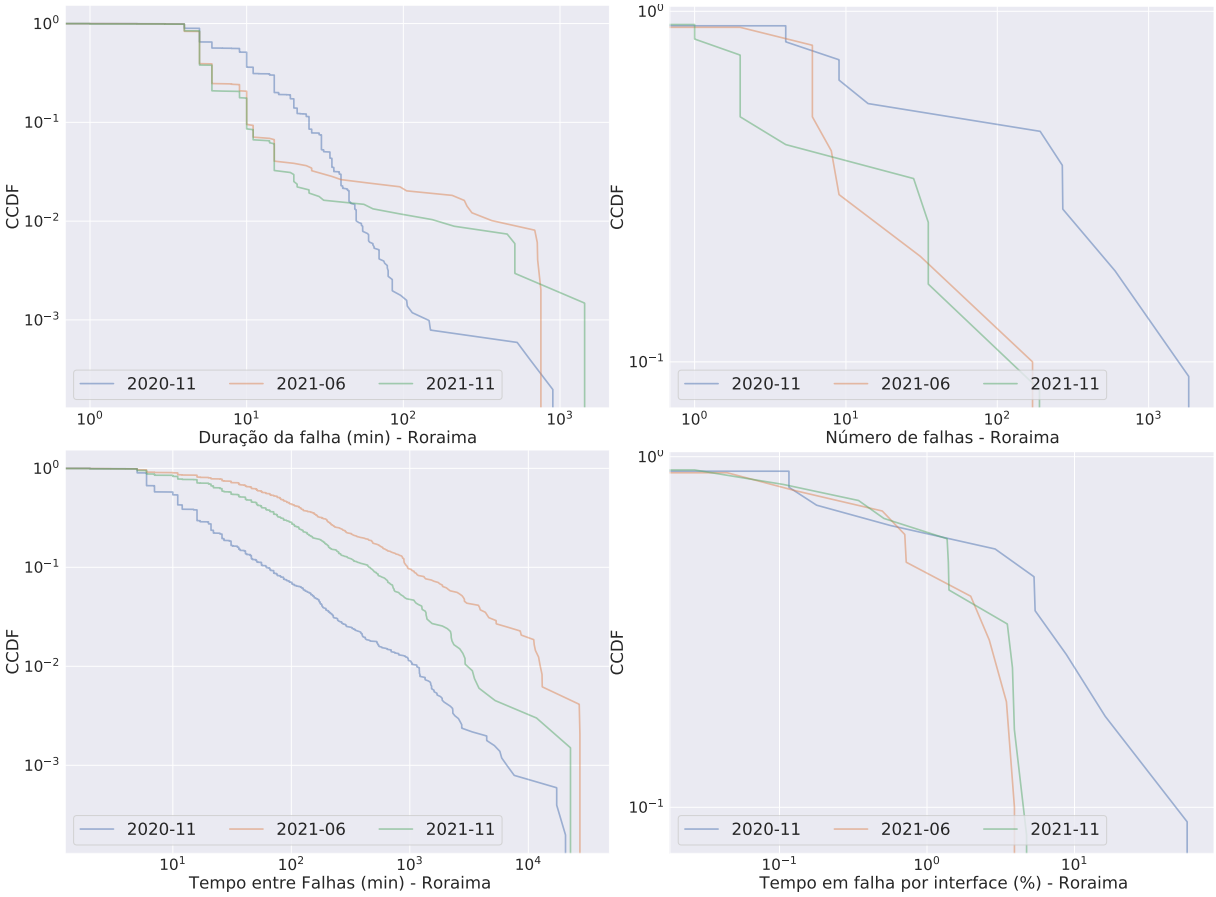

Figura 20 – Distribuições empíricas de Roraima

Roraima

Roraima é um dos estados com o menor número de locais e interfaces, 16 e 122, respectivamente. São poucos os locais e suas interfaces variam pouco no tempo. A Figura 17 mostra que o estado apresenta variação em sua fração de falhas no período em análise, atingindo um alto valor o final do ano de 2020 e, em seguida, diminuindo ao longo do tempo. O estado segue parte do padrão da região Norte, como mostra a Figura 11, apresentando o primeiro pico de falhas.

A Figura 20 apresenta as CCDFs relativas às suas falhas. As falhas do estado costumam ser curtas. Falhas superiores a 10 minutos ocorrem com probabilidade inferior a 10% nos meses do ano de 2021 e com mais frequência no mês de novembro de 2020, com 50% de probabilidade. Em todos os casos, falhas superiores a 30 minutos ocorrem com apenas 10% de probabilidade. Para falhas superiores a 100 minutos, as probabilidades são inferiores a 5% nos meses de 2021 e para o mês de 2020 são ainda mais raras, com probabilidade próxima de 0.1%

O número de falhas por interface no estado é alto. Em novembro de 2020, 80% das interfaces apresentam mais que 10 falhas e, para os meses de 2021, essa probabilidade varia entre 40 e 60%. Nos meses de junho e novembro de 2021, 10% das interfaces apresentam

mais de 200 falhas e, em novembro de 2020, 10% das interfaces apresentam mais de 2 mil falhas.

O tempo entre falhas no estado costuma ser baixo, com o mês de novembro de 2011 apresentando os menores valores. Para esse mês, apenas 1% das interfaces apresentaram tempo de operação contínuo sem falhas superior a 1000 minutos. No mês de junho, esse valor aumenta para 30% mas, em novembro de 2020, volta a diminuir, chegando a 10%. Em todos os casos, o valor pode ser considerado baixo.

O percentual de tempo em falha é alto no estado para o mês de novembro de 2020. Valores acima de 1% do tempo de operação ocorrem em aproximadamente 90% dos casos e valores acima de 10% ocorrem em 50% dos casos. Os meses de 2021 apresentam probabilidade alta, de 90%, da ocorrência valores superiores a 1%. Apesar disso, não apresentam valores máximos altos, nenhuma interface chega a apresentar 10% do seu tempo em estado de falha.

Amazonas

O Amazonas é um estado que apresenta número baixo de locais e interfaces, 33 e 287, respectivamente. São poucos os locais e suas interfaces variam pouco no tempo quando se compara ao resto do país. Na Figura 17, o estado apresenta variação em sua fração de falhas no período em análise, apresentando um alto valor durante o ano, que aumenta consideravelmente no final do ano de 2021. O estado segue parte do padrão da região Norte, apresentando o segundo pico de falhas, como mostra a Figura 11.

A Figura 21 apresenta as CCDFs relativas às suas falhas. As curvas para a duração das falhas são próximas para valores inferiores a 10 minutos. Valores superiores ocorrem com probabilidade de 10%. Falhas com duração maior que 10 minutos começam a apresentar divergência. Falhas que duram mais que 100 minutos têm ocorrência inferior a 1% nos meses junho e novembro de 2021, e ocorrência de 5% no mês de novembro de 2020.

O número de falhas, como demonstrado pela Figura 17, é muito maior no mês de novembro de 2021, e os outros meses apresentam curvas semelhantes. Para os meses de novembro de 2020 e junho de 2021, a ocorrência de interfaces com mais de 100 falhas é alta, 30 e 40% respectivamente, mas, no mês de novembro de 2021, ela é ainda maior, com aproximadamente 95%.

O tempo de operação ininterrupta das interfaces nesse estado é baixo. Nos meses de novembro de 2020 e junho de 2021, apenas 10% das interfaces apresentaram valores superiores a 1000 minutos e, em novembro de 2021, o número é menor, chegando apenas a 5%.

O tempo que uma interface permanece em estado de falhas é alto. Interfaces com valores superior a 1% ocorrem com 50% de probabilidade e, nos meses de novembro de 2020

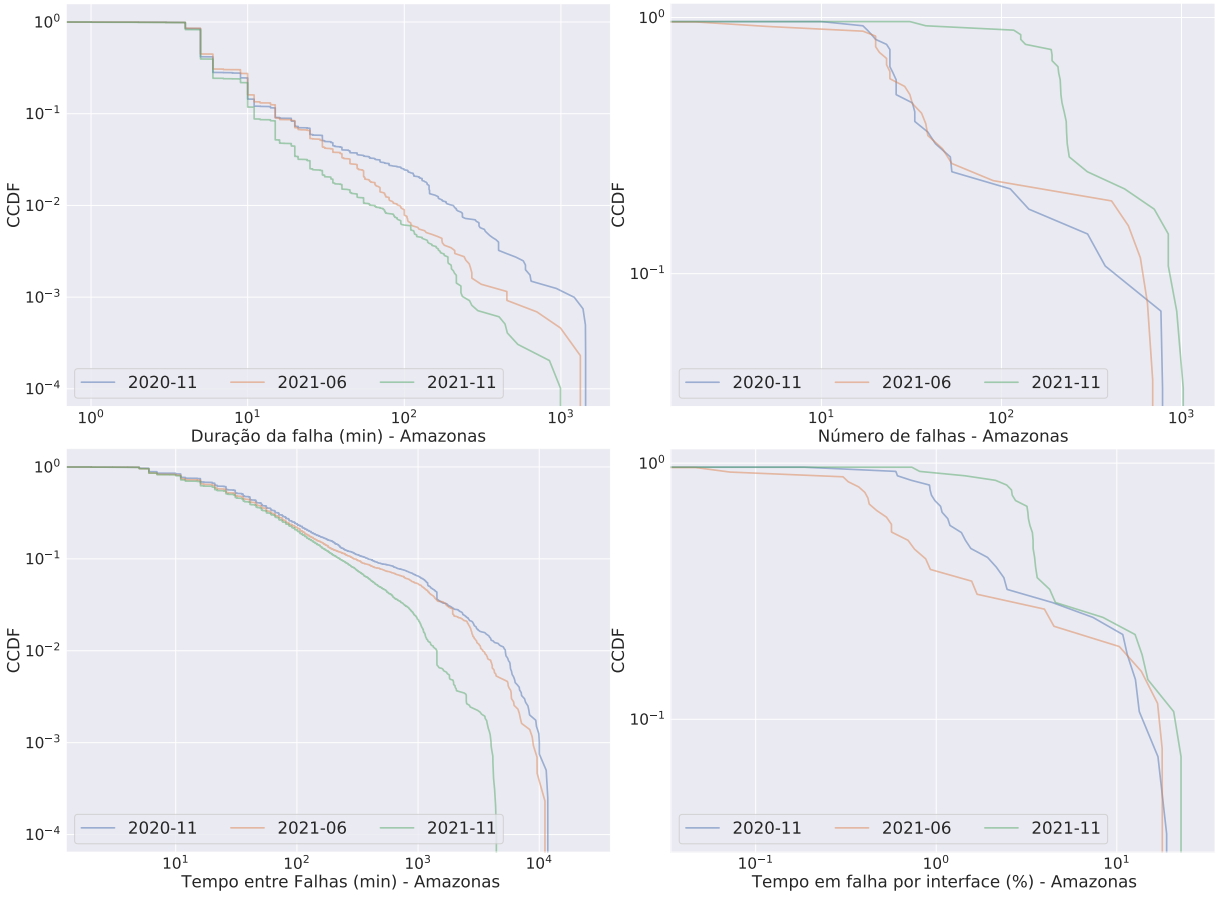

Figura 21 – Distribuições empíricas do Amazonas.

e 2021, ocorrem com probabilidade próxima de 99%. Valores superiores a 10% também ocorrem com frequência, variando de 40% no mês de junho a 50% nos meses de novembro.

Pará

O Pará é um dos estados com número alto de locais e interfaces, 107 e 2725, respectivamente. Como mostra a Figura 17, o estado apresenta um número mediano de falhas no começo do período em análise e mantém números semelhantes durante o resto do ano. Não chega a apresentar um padrão bimodal como o resto da sua região, como se vê na Figura 14, e nem houve algum mês que se destacou em relação às suas falhas.

A Figura 22 apresenta as CCDFs relativas às suas falhas. A duração das falhas costuma ser baixa. O melhor caso é o mês de novembro de 2020, em que apenas 7% das falhas apresentam valores superiores a 30 minutos, e o pior caso é o mês de junho de 2021, em que esse valor é de 20%. Falhas com valor superior a 100 minutos ocorrem apenas em 3% dos casos no mês de novembro de 2021 e em 5% nos demais meses selecionados.

Com relação à frequência das falhas, o mês de novembro de 2021 apresenta o melhor comportamento, mas a ocorrência de interfaces com muitas falhas é alta. Para o referido mês, a ocorrência é de 10%, enquanto que, nos demais, a probabilidade é próxima de 50%.

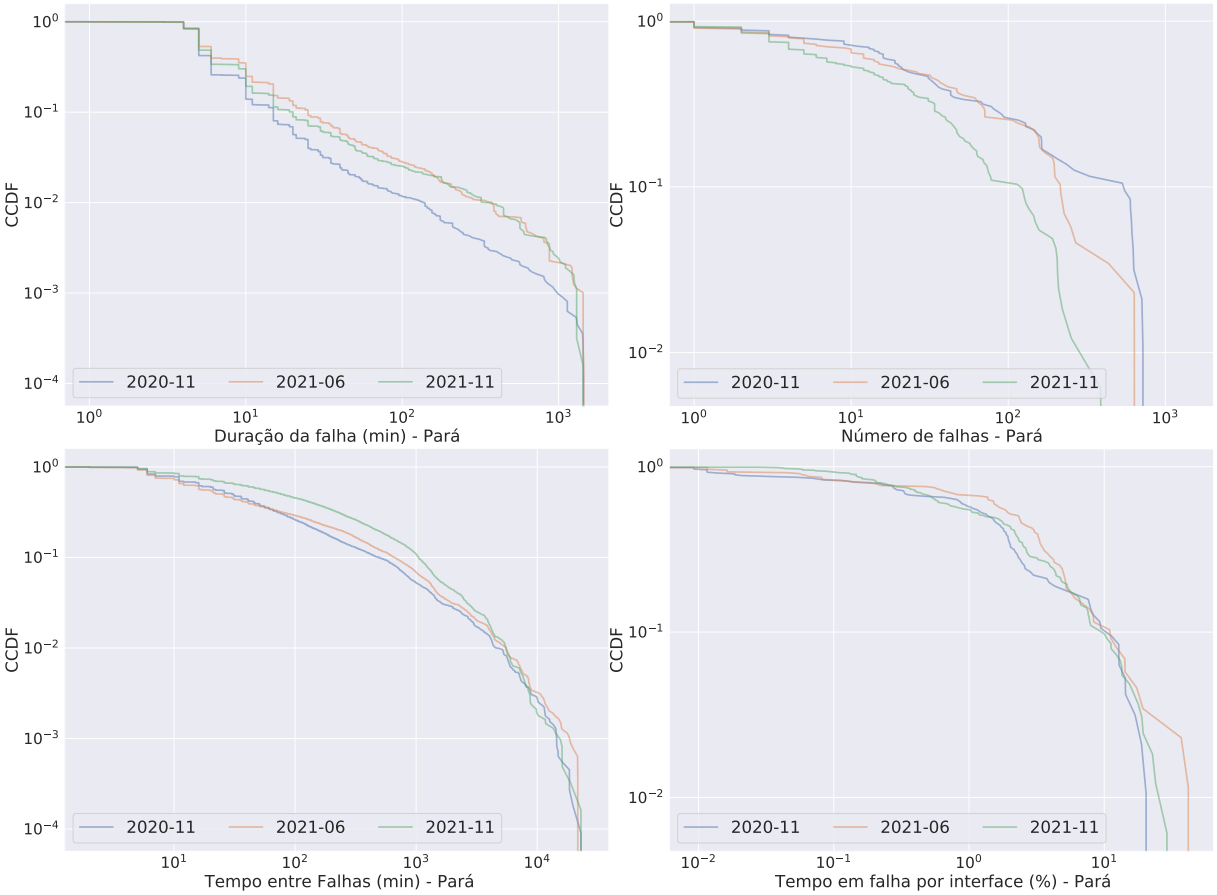

Figura 22 – Distribuições empíricas do Pará.

O tempo ininterrupto de funcionamento das interfaces pode ser considerado bom quando se compara aos demais estados. Períodos superiores a 1000 minutos ocorrem com probabilidade próxima de 5% nos meses de novembro de 2020 e junho de 2021, e com mais frequência no mês de novembro de 2021, com probabilidade de 10%.

Percentuais de tempo em falha superiores a 1% são comuns, ocorrendo em mais de 70% dos casos. Valores superiores a 1% são mais raros, ocorrendo em apenas 10% dos casos.

Uma comparação entre os estados pode também ser feita para ajudar na compreensão do comportamento das falhas em território nacional. A Figura 23 apresenta as CCDFs para os estados anteriormente analisados em relação ao mês de novembro de 2020. A imagem mostra que as falhas foram curtas nesse período: 90% das falhas possuem menos que 30 minutos. Em Roraima, são bem mais curtas que nos demais estados, e Minas Gerais costuma apresentar falhas maiores. Os estados do Norte apresentam interfaces com número maior de falhas que os estados do Sudeste. Apesar de haver variação no tempo entre falhas, com Roraima apresentando tempos mais baixos de operação contínua e Minas Gerais apresentando tempos mais altos, as probabilidades para tempos longos são baixas do ponto de vista operacional. A disponibilidade, em geral, é baixa: 70% das interfaces apresentam pelo menos 1% do seu tempo em falha. A figura deixa claro que o serviço de

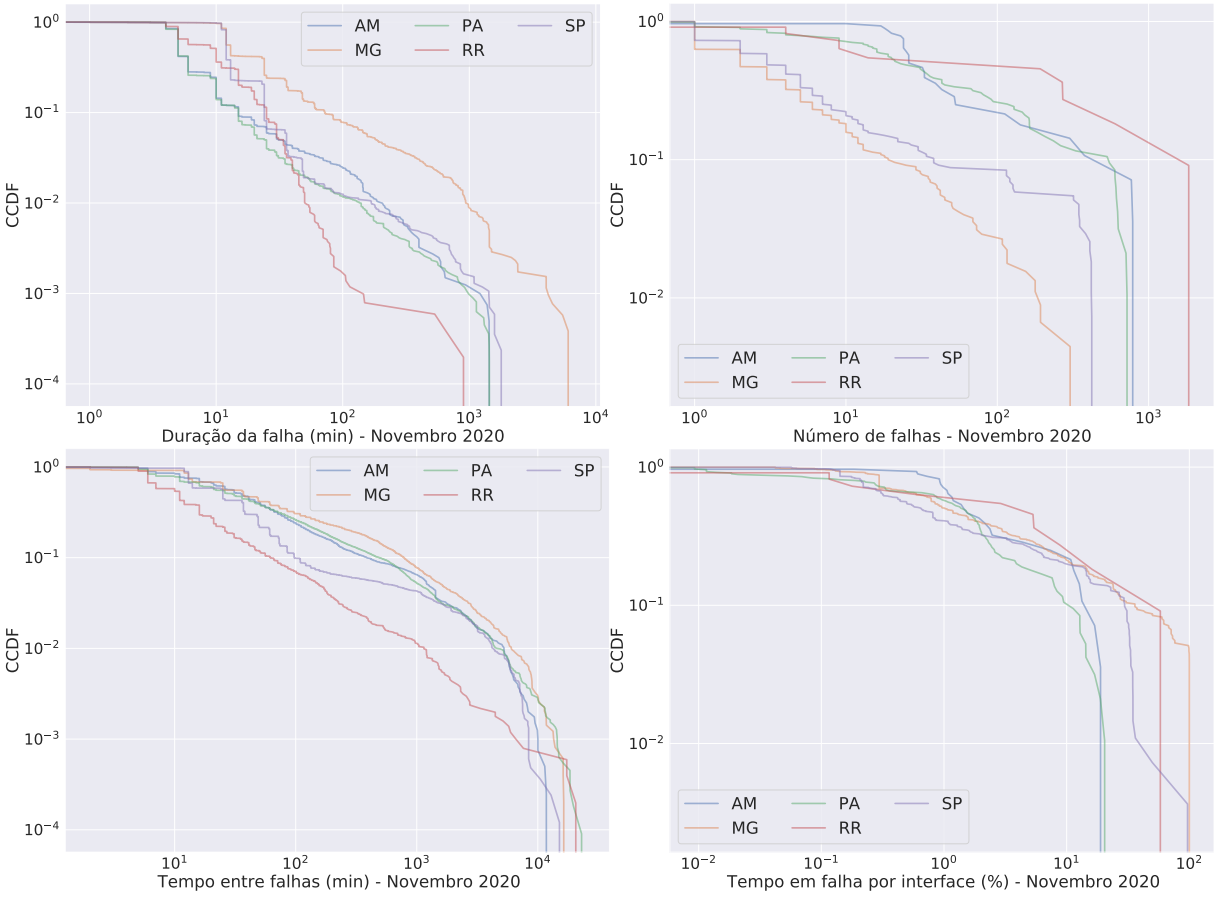

Figura 23 – Distribuições empíricas do Pará.

conectividade provido não é homogêneo: o PoP de cada estado apresenta desafios próprios.

A análise realizada na Seção 4.4, em conjunto com esta seção, mostra a complexidade da Rede Ipê e sua heterogeneidade; o comportamento das falhas é sensível a fatores espaciais e temporais. Uma análise completa precisa levar em consideração suas características e especificidades geográficas em diferentes níveis (nacional, regional e estadual) e também sua variação ao longo do tempo. Essa complexidade torna difícil o problema de predição apresentado na seção seguinte, de modo que as observações levantadas nesta seção precisam ser consideradas para a construção de modelos adequados.

6 Previsão de Falhas

O objetivo deste Capítulo é apresentar a metodologia utilizada e os resultados relacionados ao problema de predição de falhas na Rede Ipê. A Seção 6.1 define o problema de predição, a Seção 6.2 apresenta a metodologia utilizada para a avaliação de modelos, bem como a divisão e o pré-processamento dos dados a serem utilizados num problema de previsão e de aprendizado de máquina supervisionado.

6.1 Definição do problema de previsão

Para definir o problema de previsão, considera-se a definição de falha apresentada na Seção 5.1, onde, para uma sequência $T = [t_1, t_2, ...t_w, t_{w+1}, ...t_n]$ de medições, com intervalo fixo entre medições, entre duas entidades A e B possuindo um serviço de conectividade, incorre numa sequência de estados $S = [s_{t_1}, s_{t_2}, ..., s_{t_w}, s_{t_{w+1}}, ...s_{t_n}]$. O problema de predição encontra-se ilustrado na Figura 24 e consiste na seguinte pergunta: *Considerando uma janela das últimas W medições, é possível detectar a ocorrência de ao menos uma falha nas próximas K medições?*

A escolha dos parâmetros W e K depende de diversos fatores, dentre os quais se destacam: natureza do serviço, granularidade das medições, variáveis presentes no vetor s e grau de tolerância. Os dados utilizados neste trabalho apresentam o intervalo de medição de um minuto, e a escolha foi de 60 minutos para W e 15 minutos para K. Dessa forma, a tarefa de predição consiste em: num tempo t, usando uma janela de uma hora de medições anteriores, detectar se ocorrerá uma falha nos próximos 15 minutos em um serviço de conectividade.

Quanto maior o valor de W, maior a quantidade de informação usada para a predição, mas maior será o custo computacional e de memória; busca-se, então, um valor intermediário que não custe o mínimo de recursos sem prejudicar o processo de previsão. Para a escolha de W, foi feito um processo de seleção de hiperparâmetros, descrito na Seção 6.2.

O valor de K influencia diretamente a dificuldade do problema. Valores altos tornam o problema mais fácil, mas a aplicabilidade do modelo diminui proporcionalmente, podendo até tornar-se nula caso o valor escolhido seja muito alto. A escolha de 15 minutos foi feita de forma a estabelecer um compromisso entre as duas demandas em conflito, ponderando a razoabilidade e a utilidade do modelo.

A proposta de solução deste trabalho para o problema descrito envolve a utilização de modelos de aprendizado de máquina. Especificamente, tem-se um problema

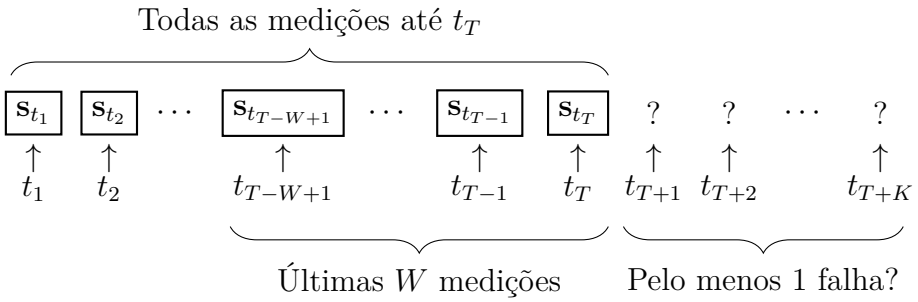

Figura 24 – Problema de previsão: dada uma janela das últimas W observações do estado do serviço e dado que no tempo t_T o serviço não está em estado de falha, o serviço se encontrará em estado de falha em pelo menos um dos K momentos futuros?

de aprendizado supervisionado do tipo classificação binária, onde o rótulo 1 representa uma falha e o rótulo 0 a ausência de uma falha. A arquitetura do modelo utiliza redes neurais, e especificamente redes neurais recorrentes do tipo LSTM. A descrição do modelo encontra-se na Seção 6.2.2

6.2 Metodologia

6.2.1 Adequação e divisão dos dados

A proposta de solução escolhida envolve o uso de **aprendizado de máquina supervisionado de classificação binária**. Os dados coletados devem então ser transformados em um formato adequado para esse tipo de modelo, assumindo a forma de um conjunto de n pares formado por uma entrada com um vetor de características $X \in \mathbb{R}^d$ e uma saída representada por um rótulo $y \in \{0, 1\}$: $D = \{(X_1, y_1), ..., (X_n, y_n)\}$, onde n representa o número de objetos de estudo e d é a dimensão do vetor de entrada.

O conjunto de dados coletado para a realização do trabalho é descrito na Seção 4.4. O total corresponde ao período de novembro de 2020 até novembro de 2021, ou seja, 13 meses consecutivos de dados. As características selecionadas para o modelo de aprendizado são as quatro variáveis numéricas presentes no conjunto: *packet loss, RTT, download e upload*. O estado do serviço no tempo t é representado por um vetor s_t possuindo os valores das características escolhidas em seu momento de medição:

$$\boldsymbol{s_t} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} (\text{packet loss})_t \\ (\text{RTT})_t \\ (\text{download})_t \\ (\text{upload})_t \end{bmatrix}$$

A janela escolhida para o trabalho é de 60 medições, portanto uma entrada X_t é composta pela última medição num instante t, s_t e as 59 medições anteriores:

$\{s_t, s_{t-1}, ..., s_{t-59}\}$. A dimensão d da entrada apresenta a forma (tamanho da janela × número de características) e X_t pode ser representada como uma matriz:

$$\boldsymbol{X_t} = \begin{bmatrix} v_{t-59,1} & v_{t-59,2} & v_{t-59,3} & v_{t-59,4} \\ v_{t-58,1} & v_{t-58,2} & v_{t-58,3} & v_{t-58,4} \\ \vdots & \vdots & \vdots & \vdots \\ v_{t,1} & v_{t,2} & v_{t,3} & v_{t,4} \end{bmatrix} = \begin{bmatrix} \text{———} & \boldsymbol{s_{t-59}^T} & \text{———} \\ \text{———} & \boldsymbol{s_{t-58}^T} & \text{———} \\ & \vdots & \\ \text{———} & \boldsymbol{s_t^T} & \text{———} \end{bmatrix}$$

O rótulo associado a uma entrada X_i, considerando o valor K escolhido de 15, é:

$$y_t = \begin{cases} 1, \text{ Se ocorre ao menos uma falha em } \{s_{t+1}, ..., s_{t+15}\} \\ 0, \text{ Caso contrário} \end{cases}$$

É importante lembrar que **falha** foi definida como o momento em que ocorre a transição de um estado de operação normal para um estado anormal, e que não faz muito sentido, do ponto de vista operacional, realizar previsões durante um estado de falha. Dessa forma, pares (X_i, y_i) onde a medição ocorre **durante** um instante t em estado de falha são removidos do novo conjunto de dados D.

A justificativa pode ser ilustrada com um exemplo simplificado do problema de predição: seja uma sequência de estados de conectividade onde 1 representa uma falha e 0 uma não falha, $Seq = [s_1, s_2, ..., s_{100}]$, em que:

$$s_i = \begin{cases} 1, \text{ se } 21 \leq i \leq 30 \text{ ou } 61 \leq i \leq 70 \\ 0, \text{ caso contrário} \end{cases}$$

Seja $M(\cdot)$ um modelo que prevê para um instante t o estado s_{t+1} com base nos valores históricos anteriores e segue a seguinte regra:

$$M(s_i) = \begin{cases} 0, \text{ se } i = 1 \\ s_{i-1}, \text{ caso contrário} \end{cases}$$

É possível verificar que, nesse caso, o modelo erra apenas no momento de transição entre os dois possíveis estados em s_{21}, s_{30}, s_{61} e s_{70} durante os momentos de transição entre estados distintos. As métricas de acurácia, precisão e revocamento são, respectivamente: 92%, 80% e 80%. O modelo apresenta resultados bons, mas, na prática, não acertou nenhuma falha antes que ocorressem; houve acerto apenas nos casos em que o sistema já estava em falha. Ao remover esses casos do conjunto de dados (não realizando predição em estados de falha), a utilidade do modelo não é afetada e a definição de falha é respeitada.

A escolha das características é feita a partir da observação dos dados da Figura 25, que apresenta a distribuição das variáveis referentes aos três minutos finais de cada entrada

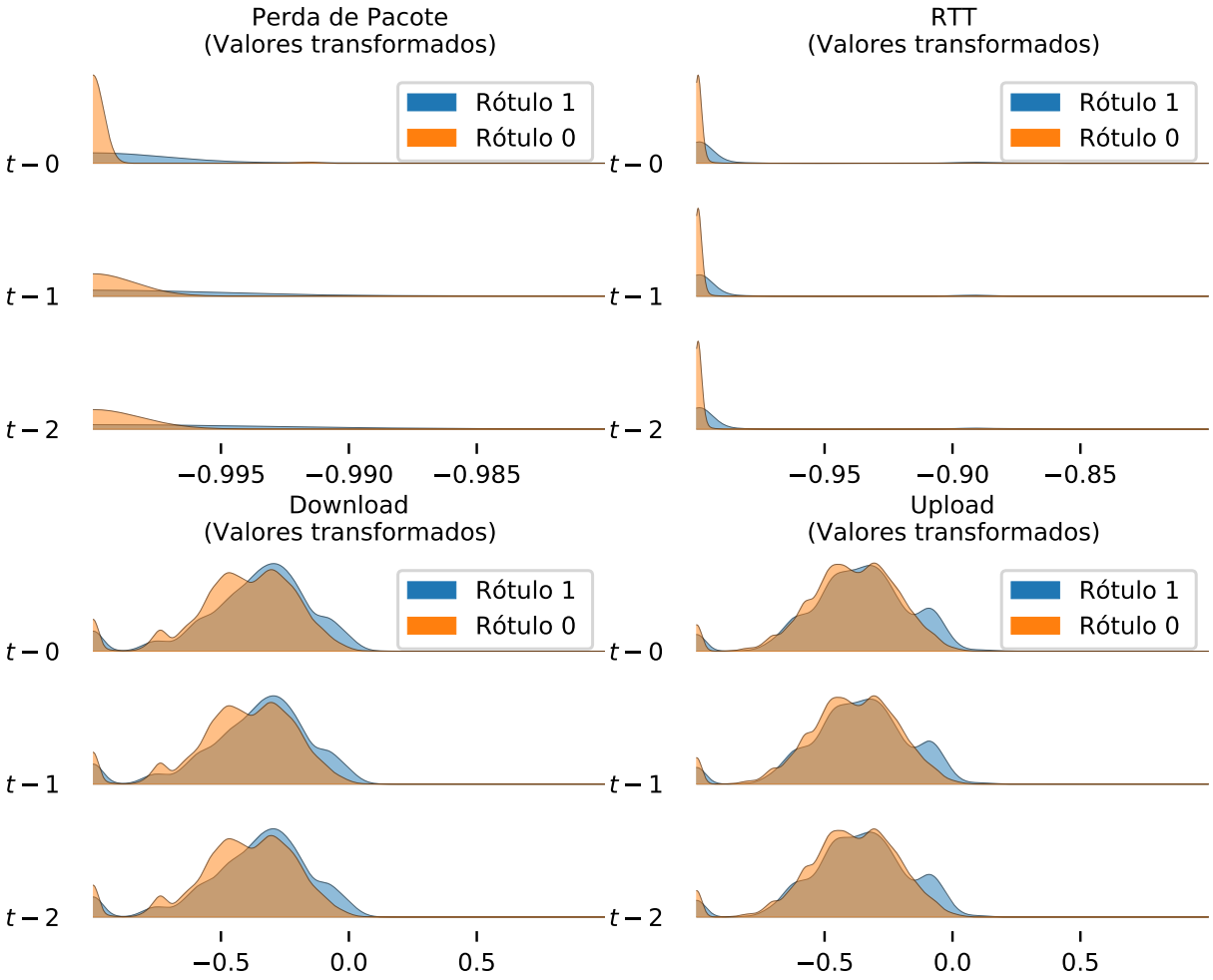

Figura 25 – Distribuições empíricas referentes aos últimos minutos de uma de entrada do problema de predição. Normalizados em [-1, 1].

X_i (as três últimas linhas de cada matriz, representando os momentos t, t-1 e t-2) para os rótulos 1, caso ocorra pelo menos uma falha nos momentos t+1, ..., t+15 e o momento t não estiver em estado de falha, e 0, caso contrário. As distribuições encontram-se normalizadas para apresentar valores no intervalo de [-1, 1]. A figura mostra que, apesar de pequena, existe uma diferença entre as distribuições dos rótulos 1 e 0. Momentos anteriores a um período contendo falhas são caracterizados por aumentos nas taxas de *packet loss*, RTT, download e upload. Esse fato remete à observação apresentada anteriormente sobre a Figura 10 de que existem falhas que ocorrem após uma degradação gradual em suas métricas de desempenho e qualidade de serviço.

Um problema de aprendizado supervisionado divide o conjunto de dados em três: os conjuntos de **treino**, **validação** e **teste**. A natureza temporal dos dados utilizados requer um cuidado especial, para evitar violações de causalidade, e limita a forma como os dados são tratados e como deve ser realizada a divisão.

Para o treinamento dos modelos referentes à Seção 6.3 utiliza-se apenas o mês de novembro de 2020, e este se encontra dividido da seguinte forma:

- Conjunto de Treino: 21 dias, de 01/11/2020 a 21/11/2020.
- Conjunto de Validação: 2 dias, 22/11/2020 e 23/11/2020.
- Conjunto de Testes: 7 dias, de 24/11/2020 a 30/11/10202.

O conjunto de dados possui granularidade alta, e as medições são feitas a cada minuto. Quando necessário, é realizado um procedimento de amostragem para tratar das limitações de recursos de *hardware*. Nesses casos, a amostragem é feita de forma estratificada (mantendo as proporções dos rótulos) e respeitando a causalidade. Mais detalhes do processo de amostragem são dados caso a caso durante seu uso. A escolha da divisão também considera a importância de que o conjunto de testes apresente todos os dias da semana e, dessa forma, seja mais representativo.

Para os modelos da Seção 6.3.5, o mês de novembro de 2020 apresenta a mesma divisão, e os demais meses, utilizados apenas para teste, passam também por um processo de amostragem aleatória estratificada.

O processo de pré-processamento das *features*, como anteriormente apresentado na Figura 25, coloca os valores das variáveis numa escala [-1, 1] através de um *scaler* MinMax[1] e, nos casos de download e upload, em variáveis com variância muito elevada, é também aplicado o logaritmo dos valores, para estabilizar a variância antes de passar o valor para a escala. Para não ocorrer violações na causalidade durante o processo, os *scalers* são criados sempre com base no conjunto de treino e, em seguida, aplicados nos demais, para evitar *data leakage*[2].

6.2.2 Modelos de Rede Neural Artificial

Os modelos propostos para resolver o problema de predição através do aprendizado de máquina são os de Redes Neurais Artificiais. Um modelo base utilizando a arquitetura de *Multi-Layer Perceptron* foi criado. Por causa da natureza temporal dos dados, foram também construídos modelos do tipo *Long Short-Term Memory*, que formam a proposta de solução deste trabalho.

[1] *Scaler* MinMax: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>

[2] *Data Leakage* refere-se à ocorrência de um 'vazamento' indevido de informações. Nesse caso, informações do futuro causando influência no passado. Um exemplo ocorreria caso o *scaler* usasse valores referentes ao conjunto de teste e, então, ao realizar a transformação dos dados de treino, informações do futuro se propagariam para o passado.

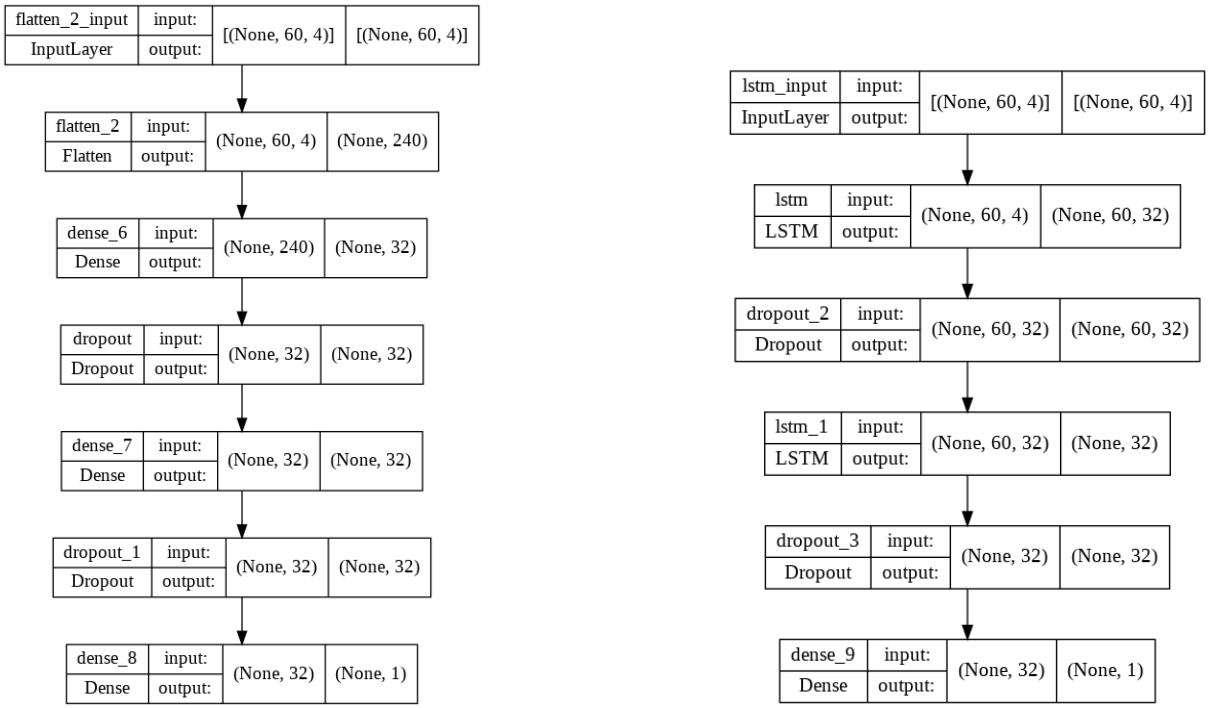

Figura 26 – Arquitetura dos modelos de rede neural MLP e LSTM, respectivamente.

Escolha de hiperparâmetros

A especificação de cada arquitetura e escolha de hiperparâmetros passa primeiro por uma etapa de *tuning* com o auxílio da biblioteca Keras Tuner[3], presente na interface Keras do Tensorflow. O processo de *tuning* usa a Otimização Bayesiana através do *tuner* BayesianOptimization[4]. Assim, não é preciso realizar a busca em todo o espaço definido para os possíveis valores dos hiperparâmetros. A seleção desse espaço segue as boas práticas encontradas na literatura de aprendizado de máquina e redes neurais e os *insights* construídos com a parte do conjunto de dados descrita na Seção 4.2, utilizada para a construção das ferramentas do trabalho.

Arquitetura de Rede Neural

A arquitetura escolhida para o modelo MLP é formada pelas seguintes camadas: uma camada *flatten*, que transforma o vetor de entrada para uma dimensão, seguida de duas camadas densas (contendo 32 unidades cada), com duas camadas de regularização usando *dropout* (com probabilidade de descarte de unidade de entrada igual a 20%), uma após cada camada densa. A última camada é do tipo densa (contendo uma unidade).

Para o modelo LSTM é utilizada uma arquitetura semelhante: duas camadas LSTM (contendo 32 unidades cada), cada uma seguida de uma camada de regularização usando *dropout* (também com probabilidade de 20%) e, ao fim, uma camada densa (contendo uma

[3]Keras Tuning: <https://www.tensorflow.org/tutorials/keras/keras_tuner>
[4]BayesianOptimization tuner: <https://keras.io/api/keras_tuner/tuners/bayesian/>

unidade). A Figura 26 apresenta as camadas e as dimensões dos dados em cada parte do modelo MLP à esquerda e do modelo LSTM à direita.

O algoritmo de otimização utilizado é o NADAM (*Nesterov Adaptive Moment Estimation*) (DOZAT, 2016), modificação do algoritmo ADAM (KINGMA; BA, 2017) com a inclusão do Momento de Nesterov, com taxa de aprendizado de 10^{-4}, e a função de perda *Focal Cross Entropy*, como definida em (Lin et al., 2017), uma modificação da entropia cruzada escolhida por apresentar resultados significantemente superiores na precisão dos modelos durante a busca por hiperparâmetros.

6.2.3 Métricas de Avaliação

As métricas de avaliação escolhidas para analisar os modelos são: **acurácia**, **precisão** e **revocação**. O conjunto de dados, como será descrito na Seção 6.3, apresenta alto desbalanceamento, favorecendo o rótulo 0. Por consequência, a acurácia por si só não teria a capacidade de avaliar o modelo como bom ou ruim. Uma forma de visualizar esse fato é com o seguinte exemplo:

Para um conjunto de dados $D = \{(X_1, y_1), ..., (X_n, y_n)\}$, onde o rótulo r_1 é muito maior que a ocorrência do rótulo r_2, modelo $M(\cdot)$ com o seguinte comportamento:

$$M(X_i) = r_1, \forall i$$

apresentaria um resultado alto de acurácia, equivalente à proporção de rótulos r_1 em relação ao total de rótulos. Sendo assim, é importante que os resultados sejam acompanhados da precisão e da acurácia do modelo.

Considerando as métricas escolhidas para o trabalho, a importância que cada uma apresenta pode variar com o contexto em que são aplicadas. Do ponto de vista operacional, em alguns casos, pode ser importante ter um valor baixo para o número de falhas que passam desapercebidas pelo modelo (**alta revocação**). No caso em que existe um alto custo para os falsos positivos, é importante que a maioria das falhas previstas pelo modelo sejam de fato falhas (**alta precisão**). Acurácia, como anteriormente discutido, é apresentada acompanhando os resultados, mas não é o valor principal a ser considerado por sua limitação em conjuntos de dados desbalanceados.

Na Seção 4.4, é possível verificar que algumas falhas ocorrem repentinamente. Tais ocorrências podem ser difíceis ou até impossíveis de se prever, afetando o valor máximo de revocação possível. Dessa forma, a revocação é uma métrica importante, mas não a principal. Portanto, para este trabalho, o foco é escolher modelos pela sua **precisão**, ou seja, quando um modelo afirma que uma falha ocorrerá nos próximos 15 minutos, espera-se que, de fato, essa falha ocorrerá.

Tabela 5 – Resultados obtidos após o balanceamento dos rótulos considerando todo o Brasil.

Proporção classe positiva	Acurácia	Precisão	Revocação
~10%	0.92	0.56	0.55
~25%	0.88	0.80	0.61
~33%	0.84	0.87	0.60

O objetivo das métricas não é apenas definir quais modelos apresentaram desempenho melhor que os demais, mas, efetivamente, responder às questões de pesquisa levantadas e concluir se existe ou não algum sinal nos dados que permite a realização de predições no contexto definido de falhas em serviços de conectividade. O refinamento de modelos até um desempenho suficiente para sua utilização em contextos operacionais é um dos possíveis desmembramentos do trabalho em projetos futuros.

6.3 Modelos de Predição

O processo de caracterização descrito na Seção 5 revela a heterogeneidade do comportamento das falhas em território nacional e a necessidade de seu estudo em diferentes níveis. De forma análoga, a construção dos modelos também abordará descrições mais gerais, em que se considera todo o território nacional, e descrições mais restritas, em que se considera apenas um estado. O modelo proposto pelo trabalho é o de redes neurais do tipo LSTM, para explorar as características temporais dos dados.

6.3.1 Modelos Balanceados

A grande maioria (cerca de 99%) das instâncias do conjunto de dados possui rótulos negativos (não ocorrência de falhas). Uma primeira tentativa de solução, criada para verificar a capacidade do modelo proposto, utiliza uma técnica de *undersampling*[5] para realizar um balanceamento artificial. São criados três novos conjuntos de dados apresentando diferentes proporções entre os rótulos positivos (ocorrência de falhas) e os negativos. São utilizados dados de todo o território nacional e a divisão descrita para o mês de novembro de 2020 na seção anterior.

A Tabela 5 apresenta os resultados obtidos após o balanceamento. É possível observar que os modelos apresentam melhora significativa, principalmente em relação à precisão, à medida que o grau de balanceamento aumenta. Sendo assim, existe evidência de que o modelo utilizado consegue reconhecer o sinal presente nos dados e separar os dois

[5]Na técnica de amostragem de *undersampling*, as instâncias que apresentam o rótulo minoritário são mantidas, e uma amostra aleatória é realizada nas instâncias que apresentam o rótulo majoritário, para modificar a proporção entre os rótulos.

Tabela 6 – Resultados obtidos para os modelos considerando todo o Brasil e Regiões Sudeste e Norte, considerando todas as interfaces ou apenas as 5% piores

	Conjunto de interfaces	**Acurácia**	**Precisão**	**Revocação**
Brasil	Todas	0.97	0.19	0.43
	5% piores	0.87	0.38	0.50
Sudeste	Todas	0.96	0.22	0.72
	5% piores	0.86	0.44	0.65
Norte	Todas	0.94	0.25	0.61
	5% piores	0.83	0.51	0.32

rótulos de forma satisfatória. Entretanto, do ponto de vista operacional, esses resultados não são úteis. O cenário real é desbalanceado, e não é possível a aplicação do *undersampling* num cenário em que o rótulo das instâncias não testadas ainda não é conhecido.

O problema de classificação em bases de dados altamente desbalanceadas é desafiador na literatura de aprendizado de máquina. Na Seção 6.1, em que é definido o problema de predição, o uso da janela futura com tamanho K visa a aumentar a proporção de instâncias com rótulo positivo sem prejudicar sua aplicabilidade em soluções reais (quando o valor de K não é grande). Novas formas de se tratar esse desbalanceamento são possíveis direções para trabalhos futuros.

6.3.2 Modelos Regionais

A heterogeneidade apresentada no processo de caracterização conduz à conclusão de que modelos devem se adequar a características geográficas. Foram criados, portanto, modelos considerando todo o Brasil e as regiões Sudeste e Norte. Além da divisão geográfica, em cada caso foi feita também uma seleção para um modelo usando dados de apenas 5% das interfaces representando os piores resultados (maior número de falhas). Os modelos também utilizam apenas os dados referentes ao mês de novembro de 2020.

A Tabela 6 apresenta os resultados obtidos. É possível notar que, ao sair do caso mais geral (Brasil) para qualquer uma das regiões (Sudeste ou Norte), a precisão apresenta um aumento, indicando que a homogeneidade dos conjuntos de dados contribui para os resultados do modelo. Outra observação é que os casos em que se consideram apenas as piores interfaces apresentam um crescimento ainda maior na precisão do modelo, sugerindo maiores similaridades entre interfaces que falham muito, que podem ir além de apenas o número de falhas apresentados durante a análise. As piores interfaces também concentram grande parte das falhas presentes no conjunto de dados, demonstrando uma proporção mais favorável entre as classes positivas e negativas.

No processo de caracterização por regiões apresentado na Seção 5.2.2, é possível verificar que a região Norte apresenta mais falhas durante o período em análise, tanto em

Tabela 7 – Resultados obtidos para os modelos estaduais.

Estado	Acurácia	Precisão	Revocação
Amazonas	0.92	0.35	0.48
Minas Gerais	0.95	0.20	0.79
Pará	0.95	0.30	0.46
Roraima	0.96	0.83	0.69
São Paulo	0.96	0.41	0.83

totalidade (Figura 15) quanto em proporção (Figura 14), que as demais regiões. Pode-se observar também a Tabela 3, que apresenta estados mais homogêneos, com PoPs a atendendo um número menor de locais. Essas características refletem-se nos resultados, com o modelo apresentando resultados superiores nessa região.

6.3.3 Modelos Estaduais

Modelos estaduais são criados de forma que possam capturar características locais mais distintas do que quando se considera toda a região a que o estado pertence. Os estados escolhidos para a criação do modelo são os mesmos caracterizados na Seção 5.2.3. Dessa forma, uma análise mais minuciosa pode ser feita em relação aos resultados.

A Tabela 7 apresenta os resultados estaduais. A qualidade dos modelos varia significativamente. O pior caso é o de Minas Gerais, que apresenta comportamento semelhante ao modelo brasileiro, com 20% de precisão, sugerindo que o estado mineiro é mais heterogêneo que os demais. Isso pode ser percebido também na Tabela 3, em que se verifica que o PoP do estado atende a uma quantidade alta de locais. Uma investigação mais aprofundada revelou também que o PoP mineiro atende a instituições de **fora da região geográfica de Minas Gerais**, como a Universidade Federal do Amazonas, a Universidade Federal de Mato Grosso do Sul e o Instituto Federal de Educação, Ciência e Tecnologia do Maranhão. O estado mineiro também foi caracterizado anteriormente com a presença de falhas **mais longas** e **menos frequentes.**

O estado de Roraima apresentou os melhores resultados, com 83% de precisão. Roraima foi o estado que se destacou pela quantidade de falhas, apresentando grande número de falhas no mês de novembro, como mostra a Figura 17. Durante sua caracterização, verificou-se, também, que o estado apresenta falhas **curtas**, **numerosas** e **próximas**. O PoP do estado também atende a um número baixo de locais, como mostra a Tabela 3, sendo bastante **homogêneo**.

Os demais estados apresentaram comportamento intermediário: São Paulo tem o melhor resultado, com 41%, Amazonas aparece em seguida, com 35%, e o Pará apresenta 30%. Os resultados sugerem que existem situações em que a falha é de mais fácil previsão e situações em que as previsões são mais difíceis ou impossíveis, e que realizar recortes

Tabela 8 – Resultados obtidos para o comportamento espacial dos modelos em relação à precisão. Os modelos são treinados no estado referente à linha e a predição é realizada no conjunto de dados referente à coluna.

Precisão	**AM**	**MG**	**PA**	**RR**	**SP**
AM	**0.35**	0.07	0.13	0.21	0.13
MG	0.35	**0.20**	0.34	0.73	0.39
PA	0.35	0.14	**0.30**	0.44	0.36
RR	0.20	0.06	0.16	**0.83**	0.40
SP	0.34	0.13	0.16	0.79	**0.41**

geográficos, por si só, não é suficiente para garantir bons resultados. O estudo de fatores que influenciam os resultados de previsão e a classificação de categorias de falhas é mais uma tarefa que pode ser explorada em trabalhos futuros.

6.3.4 Comportamento de um modelo em relação ao espaço

Uma forma de explorar o comportamento dos modelos e entender a previsão de falhas é observar como um modelo treinado com os dados de um local se comporta em outra localidade. Modelos mais gerais apresentam resultados inferiores, mas uma pergunta interessante é se um modelo treinado, que apresenta resultados bons em sua localidade, consegue generalizar seus resultados em outros lugares. Para isso, modelos treinados em um estado são usados para prever falhas em dados pertencentes aos demais.

A Tabela 8 apresenta os resultados obtidos. Os modelos são treinados com os dados referentes ao estado em cada linha e a predição é realizada no estado referente a cada coluna; a diagonal é, então, o modelo treinado e testado em sua própria localidade. O resultado chama a atenção em alguns pontos. Para o estado do Amazonas, todos os modelos apresentaram resultados semelhantes, o que pode implicar a presença de uma categoria de falha mais geral presente na área em questão.

Para o modelo do estado de Minas Gerais, são encontrados resultados superiores aos de sua própria região e próximos aos resultados dos modelos em suas localidades originais, o que leva a crer que a heterogeneidade do estado, apesar de dificultar as previsões locais, pode ajudar em alguns casos com a generalização espacial do modelo.

Todos os modelos apresentaram bons resultados para os dados de Roraima, o que sugere que as falhas da região são, de fato, mais fáceis de serem previstas. Por outro lado, o modelo de Roraima não generaliza bem para os demais locais, sofrendo uma espécie de *overfitting* a essa categoria de falha que ocorre em grandes quantidades no estado. Seu único resultado próximo do resultado local ocorre com os dados de São Paulo, o que sugere a presença desse tipo de falha no estado paulista.

Tabela 9 – Resultados obtidos para o comportamento temporal dos modelos em relação à precisão.

	Amazonas	**Minas Gerais**	**Pará**	**Roraima**	**São Paulo**
11/2020	0.35	0.20	0.30	0.83	0.41
12/2020	0.21	0.13	0.08	0.82	0.40
01/2021	0.20	0.10	0.18	0.80	0.36
02/2021	0.27	0.07	0.17	0.80	0.03
03/2021	0.23	0.10	0.12	0.73	0.04
04/2021	0.28	0.12	0.27	0.00	0.05
05/2021	0.31	0.09	0.29	0.04	0.04
06/2021	0.13	0.08	0.25	0.04	0.08
07/2021	0.10	0.08	0.26	0.03	0.05
08/2021	0.19	0.13	0.17	0.05	0.09
09/2021	0.31	0.14	0.23	0.06	0.12
10/2021	0.34	0.26	0.18	0.03	0.22
11/2021	0.22	0.14	0.17	0.04	0.44

6.3.5 Comportamento de um modelo em relação ao tempo

Outra forma de se estudar o comportamento dos modelos é verificando o seu desempenho ao longo do tempo. Pode-se observar se o desempenho decai e, se preciso, qual o intervalo de tempo necessário para treinar novamente o modelo com dados novos. Para isso, os modelos são treinados em suas localidades utilizando os dados de novembro de 2020 e seu desempenho é testado nos demais meses, até novembro de 2021.

Os resultados se encontram na Tabela 9, que, quando vista em conjunto com a Figura 17, evidencia a influência do tempo nos resultados. O modelo de São Paulo é treinado num pico de falha e seu resultado se mantém enquanto a ocorrência de falha é alta. O desempenho cai quando os meses mudam sua característica, apresentando a ocorrência de poucas falhas e, por fim, recuperando parte do desempenho novamente no final do ano de 2021, quando um número maior de falhas volta a ocorrer.

Em Roraima, existe apenas um pico de falhas no começo do conjunto de dados. O desempenho do modelo cai drasticamente quando o período de alto número de falhas acaba. É um modelo que apresenta bons resultados, mas eles são limitados à sua região e ao tipo de período em que foi treinado.

O modelo do Amazonas manteve resultados semelhantes, variando sua precisão entre 20 e 30%. Apenas nos meses de junho e julho, quando o número de falhas atinge seu mínimo, o desempenho cai até 13% e 10%, respectivamente.

Pará apresenta uma grande queda em sua precisão de novembro para dezembro de 2020. São dois períodos com número de falhas semelhantes, levando a crer em uma possível mudança nas categorias de falhas ocorridas nesses dois meses.

Tabela 10 – Resultados para o modelo MLP considerando os estados de AM, MG, PA, RR e SP.

Estado	**Acurácia**	**Precisão**	**Revocação**
AM	0.73	0.14	0.76
MG	0.82	0.06	0.86
PA	0.79	0.09	0.65
RR	0.94	0.72	0.70
SP	0.89	0.19	0.91

6.3.6 Comparação entre MLP e LSTMs

A proposta deste trabalho é o uso de LSTMs para a solução do problema de predição. Para justificar esse uso, é feita uma comparação com redes neurais MLP. Dessa forma, compara-se o resultado obtido em cinco situações distintas, considerando-se os modelos estaduais. Os dados utilizados são os referentes ao mês de novembro de 2020.

Os resultados encontram-se na Tabela 10, em que se observa que, apesar de haver um resultado bom para Roraima, todos os resultados têm a precisão mais baixa que o modelo LSTM, especialmente nos casos de Minas Gerais e Pará. Dessa forma, justifica-se o uso dos modelos propostos e verifica-se a capacidade das redes neurais LSTM em trabalhar dados temporais.

7 Conclusões e Trabalhos Futuros

A Rede Ipê mostra-se fundamental para o desenvolvimento e a evolução da comunidade científica nacional. É uma rede de grandes dimensões que conecta universidades e centros de pesquisa em todo o país e apresenta também ligações com redes acadêmicas no exterior. Essa escala traz desafios para a sua operação; no entanto, como mostra o trabalho apresentado, no que tange às falhas em sistemas provedores de conectividade, é possível tratar essa complexidade com o empenho de técnicas adequadas de estatística e aprendizado de máquinas. Primeiramente, é realizada a coleta, armazenamento, tratamento e descrição dos dados e, em seguida os dados tratados são caracterizados através de suas distribuições de probabilidade, para gerar *insights* sobre o funcionamento da rede. Ao final, modelos de aprendizado são criados com base nas novas intuições adquirias.

Os Capítulos 4 e 5 apresentaram a caracterização tanto dos dados brutos coletados, quanto dos dados no contexto de falhas. O comportamento dos dados é explorado em diversos níveis geográficos e durante o período de um ano. Com base na análise realizada, é possível afirmar que a qualidade do serviço de conectividade apresentou grande variação nos PoPs do país. Regiões e estados apresentam características diferentes em seu comportamento com relação às falhas. A rede qualifica-se, então, por um comportamento heterogêneo. A diferença no comportamento das falhas ocorre também presente em relação ao tempo. O período estudado apresenta picos no número de falhas durante alguns meses em determinadas regiões, bem como momentos de baixa ocorrência.

O Capítulo 6 define o problema de predição e apresenta as propostas de modelos de aprendizado de máquina para seu tratamento. Os modelos são criados considerando-se três diferentes níveis geográficos: nacional, regional e estadual. Bons resultados são alcançados em alguns casos e o desempenho dos modelos apresenta melhora à medida que seu escopo é restringido, atingindo os melhores resultados nos casos dos modelos estaduais. O desempenho dos modelos estaduais também passa por duas avaliações: a capacidade do modelo treinado em um estado de generalizar e obter resultados semelhantes em outro estado e a capacidade do modelo manter sua performance por longos períodos. Para o estado de Roraima, modelos treinados em outras localidades exibem bons resultados, até superiores aos de seu local originário. Minas Gerais, por outro lado, é um estado onde os modelos de outras localidades mostram uma alta queda de performance. Quanto ao desempenho no tempo, quando o perfil do mês de treinamento se mantém, como no caso de modelos treinados num período caracterizado por um grande número de falhas, os modelos continuam alcançando resultados próximos, mas apresentando queda em seu desempenho quando ocorre a mudança para um mês de poucas falhas. Não é adequado, portanto, treinar um modelo e usá-lo por longos períodos sem atualizá-lo. O treinamento

contínuo com dados novos é um caminho a ser explorado para garantir a longevidade dos modelos.

O trabalho levantou três hipóteses: *(1) O uso de algoritmos de aprendizado de máquina é adequado para a previsão de falhas em enlaces de redes, (2) Falhas apresentam diferentes dificuldades de predição, podendo ser fáceis, difíceis ou até impossíveis e (3) O comportamento das falhas não é uniforme no território nacional.* Os resultados relacionados à previsão mostram que a abordagem proposta é promissora, atingindo, em alguns casos, precisão superior a 80%, mas, em outros casos, como nos modelos mais gerais, a metodologia e as técnicas ainda precisam ser aprimoradas. O maior desafio encontrado está relacionado ao alto desbalanceamento das classes. A ocorrência de falha é baixa em relação às não-falhas, dificultando o aprendizado e a generalização dos modelos. Portanto, a primeira hipótese é apenas parcialmente confirmada, e torna-se necessária a realização de trabalhos futuros para explorar o tema e sanar as dificuldades encontradas. A grande variação nos resultados sustenta as hipóteses 2 e 3: modelos que apresentam precisão baixa em suas próprias regiões obtiveram resultados superiores em outros estados, sugerindo a existência de falhas de fácil predição. O comportamento contrário também ocorre: modelos com bons resultados apresentam perda de performance em algumas localidades. O território nacional tem regiões em que falhas são raras durante todo o período estudado e regiões em que a presença de falhas apresenta-se alta durante a maior parte do tempo, como as regiões Nordeste e Norte, respectivamente. Alguns estados apresentaram momentos com falhas curtas e constantes, como Roraima, e outros apresentaram falhas mais longas e infrequentes, como Minas Gerais. Portanto, as falhas distinguem-se em sua dificuldade de predição e o território brasileiro apresenta uma distribuição heterogênea dos tipos de falha.

7.1 Trabalhos futuros

Este trabalho revela vários desafios e possibilidades de melhoria que podem ser objeto de estudos futuros. Entre as possibilidades, destacam-se:

- Implementação de algoritmos mais refinados. O campo de aprendizado de máquina apresenta grande variedade de modelos em desenvolvimento e refinamento. No campo das redes neurais recorrentes, as redes Transformers (VASWANI et al., 2017) aparecem em destaque nos últimos anos e, dessa forma, a exploração de seu desempenho é uma opção apropriada no contexto deste trabalho;

- Busca de novas fontes de dados para enriquecer o conjunto utilizado. O trabalho utilizou apenas quatro características para realizar a predição, e a adição de mais informação ao modelo pode contribuir com os resultados dos modelos;

- Criação de modelos de aprendizado on-line. Como se observou na Seção 6.3.5, o

desempenho dos modelos pode decair com o tempo, tornando necessário atualizar os modelos com dados atuais;

- Identificação e estudo dos tipos de falha existentes. Falhas fáceis e difíceis foram encontradas. Classificar as falhas e verificar sua distribuição no território nacional pode trazer melhorias ao modelo;
- Criação de *dashboards* interativos para a visualização dinâmica do comportamento das falhas em suas diversas dimensões. O alto volume de dados torna a visualização estática ineficiente e custosa, e criar todas as visualizações possíveis é inviável. Gráficos interativos em que o usuário especifica parâmetros para a visualização são mais eficientes;
- Criação de modelos para os estados e regiões não contemplados neste trabalho. A Rede Ipê abrange todo o país e, portanto, para uma análise completa, deve ser feita a extensão para os demais estados.

7.2 Trabalhos publicados

O trabalho foi desenvolvido em conjunto com professores e colaboradores do laboratório LabNERDS da UFES e gerou a seguinte publicação, e sua apresentação no evento SBRC 2021[1]:

ZANOTELLI, Vitor F.; COMARELA, Giovanni; VILLACA, Rodolfo S.; MARTINELLO, Magnos. Caracterização e Previsão de Falhas em Serviços de Conectividade: uma Aplicação à Rede Ipê. In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS (SBRC), 39. , 2021, Uberlândia. <https://doi.org/10.5753/sbrc.2021.16717>

Resumo: *A Rede Ipê é fundamental para a comunidade científica brasileira por interconectar universidades e centros de pesquisa de todo o país. Este artigo analisa algumas características da Rede Ipê e explora o uso de técnicas de aprendizado de máquina para predição de falhas em serviços de conectividade usando dados públicos disponibilizados pela ferramenta ViaIpê. O problema é abordado como uma tarefa de classificação binária utilizando redes neurais recorrentes. Os resultados mostram que a dependabilidade do serviço de conectividade varia significativamente nos diferentes PoPs da Rede Ipê. Além disso, apesar da heterogeneidade deste serviço, os modelos de predição mostram-se promissores, apresentando boa acurácia e boa precisão em alguns cenários.*

[1]SBRC 2021: <https://www.sbrc2021.facom.ufu.br/>

Referências

AVIžIENIS, A.; LAPRIE, J.-C.; RANDELL, B. Dependability and Its Threats: A Taxonomy. In: JACQUART, R. (Ed.). *Building the Information Society*. Boston, MA: Springer US, 2004. (IFIP International Federation for Information Processing), p. 91–120. ISBN 978-1-4020-8157-6. Citado na página 27.

Azzouni, A.; Pujolle, G. Neutm: A neural network-based framework for traffic matrix prediction in sdn. In: *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*. [S.l.: s.n.], 2018. p. 1–5. Citado na página 36.

BAGNALL, A. et al. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, v. 31, n. 3, p. 606–660, maio 2017. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-016-0483-9>. Citado na página 35.

BOSSHART, P. et al. P4: Programming protocol-independent packet processors. *SIGCOMM Comput. Commun. Rev.*, Association for Computing Machinery, New York, NY, USA, v. 44, n. 3, p. 87–95, jul. 2014. ISSN 0146-4833. Disponível em: <https://doi.org/10.1145/2656877.2656890>. Citado na página 35.

BOUTABA, R. et al. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, v. 9, n. 1, p. 16, dez. 2018. ISSN 1867-4828, 1869-0238. Disponível em: <https://jisajournal.springeropen.com/articles/10.1186/s13174-018-0087-2>. Citado 2 vezes nas páginas 23 e 35.

CARNEIRO, T. et al. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, v. 6, p. 61677–61685, 2018. ISSN 2169-3536. Conference Name: IEEE Access. Citado na página 49.

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. *Mathematics for Machine Learning*. [S.l.]: Cambridge University Press, 2020. Citado na página 28.

DOZAT, T. INCORPORATING NESTEROV MOMENTUM INTO ADAM. p. 4, 2016. Citado na página 75.

FAWAZ, H. I. et al. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, v. 33, n. 4, p. 917–963, jul. 2019. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-019-00619-1>. Citado na página 36.

GIANNAKOU, A.; DWIVEDI, D.; PEISERT, S. A machine learning approach for packet loss prediction in science flows. *Future Generation Computer Systems*, v. 102, p. 190–197, 2020. ISSN 0167-739X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167739X19305850>. Citado na página 37.

GILL, P.; JAIN, N.; NAGAPPAN, N. Understanding network failures in data centers: Measurement, analysis, and implications. *SIGCOMM Comput. Commun. Rev.*, Association for Computing Machinery, New York, NY, USA, v. 41, n. 4, p. 350–361, ago. 2011. ISSN

0146-4833. Disponível em: <https://doi.org/10.1145/2043164.2018477>. Citado na página 36.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado 2 vezes nas páginas 28 e 31.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Comput.*, MIT Press, Cambridge, MA, USA, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667. Disponível em: <https://doi.org/10.1162/neco.1997.9.8.1735>. Citado na página 33.

KINGMA, D. P.; BA, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, jan. 2017. ArXiv: 1412.6980. Disponível em: <http://arxiv.org/abs/1412.6980>. Citado na página 75.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, maio 2015. ISSN 1476-4687. Number: 7553 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/nature14539>. Citado na página 36.

LENAIL, A. Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, The Open Journal, v. 4, n. 33, p. 747, 2019. Disponível em: <https://doi.org/10.21105/joss.00747>. Citado na página 32.

Lens Shiang, E. P. et al. Gated recurrent unit network-based cellular trafile prediction. In: *2020 International Conference on Information Networking (ICOIN)*. [S.l.: s.n.], 2020. p. 471–476. Citado na página 36.

Liberato, A. et al. Rdna: Residue-defined networking architecture enabling ultra-reliable low-latency datacenters. *IEEE Transactions on Network and Service Management*, v. 15, n. 4, p. 1473–1487, 2018. Citado na página 35.

Lin, T. et al. Focal loss for dense object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. p. 2999–3007. Citado na página 75.

Markopoulou, A. et al. Characterization of failures in an operational ip backbone network. *IEEE/ACM Transactions on Networking*, v. 16, n. 4, p. 749–762, 2008. Citado na página 36.

MARTINELLO, M. *Availability Modeling and Evaluation of Web-based Services-A pragmatic approach*. Tese (Doutorado) — Institut National Polytechnique de Toulouse, INPT, 2005. Citado 3 vezes nas páginas 23, 27 e 51.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2. Citado na página 28.

OLIVEIRA, T. P.; BARBAR, J. S.; SOARES, A. S. Computer network traffic prediction: a comparison between traditional and deep learning neural networks. *International Journal of Big Data Intelligence*, v. 3, n. 1, p. 28, 2016. ISSN 2053-1389, 2053-1397. Disponível em: <http://www.inderscience.com/link.php?id=73903>. Citado na página 36.

VASWANI, A. et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964. Citado na página 84.

WANG, Z.; YAN, W.; OATES, T. Time series classification from scratch with deep neural networks: A strong baseline. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2017. p. 1578–1585. ISSN: 2161-4407. Citado na página 36.

WICKHAM, H. Tidy data. *The Journal of Statistical Software*, v. 59, 2014. Disponível em: <http://www.jstatsoft.org/v59/i10/>. Citado na página 44.

YANG, Q.; WU, X. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, v. 05, n. 04, p. 597–604, dez. 2006. ISSN 0219-6220. Publisher: World Scientific Publishing Co. Disponível em: <https://www.worldscientific.com/doi/abs/10.1142/S0219622006002258>. Citado na página 35.

ZHANG, K. et al. Automated IT system failure prediction: A deep learning approach. In: *2016 IEEE International Conference on Big Data (Big Data)*. [S.l.: s.n.], 2016. p. 1291–1300. Citado na página 37.

Zhong, J.; Guo, W.; Wang, Z. Study on network failure prediction based on alarm logs. In: *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*. [S.l.: s.n.], 2016. p. 1–7. Citado na página 37.

Zhou, Z.; Zhang, T. Applying machine learning to service assurance in network function virtualization environment. In: *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*. [S.l.: s.n.], 2018. p. 112–115. Citado na página 37.