Universidade de Brasília – UnB

Faculdade de Tecnologia – FT

Engenharia de Redes de Comunicação

AVALIAÇÃO DE CAPACIDADE DE INFRAESTRUTURAS DE REDES DE COMUNICAÇÃO DE DADOS

Autor: Arthur Antonoff dos Santos
Évilin Vieira Dantas

Orientador: Flávio Elias Gomes de Deus

Brasília, DF

2019

Arthur Antonoff dos Santos
Évilin Vieira Dantas

AVALIAÇÃO DE CAPACIDADE DE INFRAESTRUTURAS DE REDES DE COMUNICAÇÃO DE DADOS

Monografia submetida ao curso de graduação em Engenharia de Redes de Comunicação da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Redes de Comunicação.

Universidade de Brasília – UnB

Faculdade de Tecnologia – FT

Orientador: Flávio Elias Gomes de Deus

Brasília, DF

2019

Arthur Antonoff dos Santos
Évilin Vieira Dantas
AVALIAÇÃO DE CAPACIDADE DE INFRAESTRUTURAS DE REDES DE COMUNICAÇÃO DE DADOS/ Arthur Antonoff dos Santos
Évilin Vieira Dantas. – Brasília, DF, 2019-
76 p. : il. (algumas color.) ; 30 cm.

Orientador: Flávio Elias Gomes de Deus

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade de Tecnologia – FT , 2019.

1. Redes, Capacidade, Tecnologia da Informação, Custo, Benefício, Avaliação, Desempenho, Infraestrutura. 2. . I. Flávio Elias Gomes de Deus. II. Universidade de Brasília. III. Faculdade de Tecnologia. IV. AVALIAÇÃO DE CAPACIDADE DE INFRAESTRUTURAS DE REDES DE COMUNICAÇÃO DE DADOS

CDU 02:141:005.6

Arthur Antonoff dos Santos
Évilin Vieira Dantas

AVALIAÇÃO DE CAPACIDADE DE INFRAESTRUTURAS DE REDES DE COMUNICAÇÃO DE DADOS

Monografia submetida ao curso de graduação em Engenharia de Redes de Comunicação da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Redes de Comunicação.

Trabalho aprovado. Brasília, DF, 11 de julho de 2019 – Data da aprovação do trabalho:

Flávio Elias Gomes de Deus
Orientador

Prof. Msc. Rodrygo Torres Cordova
Convidado 1

Brasília, DF
2019

Esse trabalho é dedicado aos engenheiros de Redes de Comunicação. Esperamos que seja proveitoso em sua profissão.

Agradecimentos

Primeiramente, meu agradecimento é todo a Ele, que me permitiu viver tudo o que vivi e chegar até aqui. À minha mãe, agradeço por toda a paciência que tem comigo. Sem ela eu realmente não teria conseguido. A meus pais que sempre me deram total condição e apoio durante toda essa longa jornada. Aos meus irmãos que aguentam meu estresse diariamente e ainda assim me amam. Aos meus amigos que sempre me deram forças, e estiveram presente comigo em todas as situações ruins. Em especial, Lara Daiana, minha prima amada, e as amigas que conheci por conta da Universidade, mas levarei pra vida toda, Débora Luiza, Tatiana Veloso e Heloísa Vitória. Às amigas da infância, que amo com todo o coração, Kamila, Mari, Thay e Jade, vocês são inspirações para mim. E também às amigas do colégio, Rafa e Andressa, que sempre se preocupam comigo. Tenho que agradecer imensamente a minha dupla desse projeto, Arthur Antonoff, que foi minha dupla em diversos outros trabalhos e sempre me dizia pra acreditar, porque sempre daria certo. Obrigada por nunca me deixar desistir. Vários outros amigos que fiz no curso de Engenharia e também sabem da minha gratidão, mas existem que eu não posso deixar de mencionar. Obrigada, Bessinha, Renan, Silveira, Caio, Matheus, Alvin, Yuri Moy e Binho.

(Évilin Vieira Dantas)

Agradeço a Deus por me permitir concluir mais esta etapa em minha vida, mas, sem esquecer minha família e amigos que sempre me apoiaram de todas as formas possíveis e as vezes impossíveis como minha Mãe, Catia Alves Antonoff, sempre conseguiu fazer. Em especial coloco três gerações que me sempre inspiraram: Meu avô, Tzvetan Antonoff, imigrante búlgaro da época da Segunda Guerra, o primeiro engenheiro presente em minha vida e o primeiro adversário vencido no Xadrez sem ajuda (eu acho). Meu pai, Marco Túlio dos Santos, que me ensinou a amá-lo por toda sua vida e por toda a minha, deixando a saudade do grande pai e amigo que sempre foi. E ao meu filho, Ícaro Coelho Antonoff, que me obriga a ser minha melhor versão de mim a cada dia, me oferecendo a oportunidade crescer, lutar e amar por aquilo que acredito e quero em minha vida.

(Arthur Antonoff dos Santos)

"Recebei a instrução e não o dinheiro.
Preferi a ciência ao fino ouro, pois a Sabedoria vale mais que as pérolas e joia alguma a pode igualar. "
(Bíblia Sagrada, Provérbios 8, 10-11)

Resumo

Este trabalho propõe uma nova metodologia para se realizar avaliações de capacidade de infraestruturas de redes de comunicação apoiada em teoremas estatísticos. Para este fim, inicialmente é sugerido levantar relações custos x benefícios ao se investir em infraestrutura na área da Tecnologia da Informação. Porém, devido à complexidade de se mensurar os gastos ao longo da vida útil desses ativos, e assim, mensurar o valor investido no total, buscou-se determinar outras medidas, sendo essas da estatística. Tais medidas, aliadas a modelagens analíticas e simulações computacionais, determinam e justificam necessidades de upgrades ou manutenções futuras na infraestrutura. Por fim, as considerações finais estipulam, com exemplos numéricos e gráficos, a utilidade e o benefício desta metodologia aplicada. Sugestões de trabalhos futuros encerram esse trabalho.

Palavras-chaves: Tecnologia da Informação, Custo, Benefício, Avaliação, Desempenho, Capacidade, Infraestrutura, Redes.

Abstract

This study proposes a new methodology for conducting evaluations of the capacity of communication network infrastructures based on statistical theorems. To this end, it is initially suggested to raise cost-benefit ratios when investing in information technology infrastructure. However, due to the complexity of measuring the expenditures over the useful life of these assets, and thus, to measure the amount invested in the total, appeared the need to determine other measure, these being the statistics. Such measure, combined with analytical modeling and computational simulations, determine and justify the need for future upgrades or maintenance of the infrastructure. Finally, the final considerations stipulate, with numerical and graphical examples, the usefulness and the benefit of this applied methodology. Suggestions for future work close this study.

Key-words: Information Technology, Cost, Benefit, Evaluation, Performance, Capacity, Infrastructure, Network.

Lista de ilustrações

Lista de tabelas

Sumário

1 INTRODUÇÃO

A globalização trouxe consigo grande necessidade de estabelecer comunicação de formas cada vez mais eficientes e junto a ela o crescimento tecnológico vem proporcionando a criação de novos dispositivos capazes de se conectarem à rede de dados. Essa diversidade de dispositivos gera um desafio à segurança da informação em todos seus pilares, pois, o dado deve respeitar a seguinte tríade: confidencialidade, integridade e disponibilidade, para os diferentes dispositivos e plataformas. Tal cenário é observado em vários contextos sociais, seja esse um ambiente doméstico, acadêmico ou corporativo. Dando um enfoque especial ao meio corporativo, enxerga-se a tecnologia como um mecanismo que impulsiona o desenvolvimento de uma empresa, acelerando e sofisticando o processo produtivo e diminuindo seus gastos, tornando-a, assim, mais eficaz. Tal característica garante a necessidade do vasto uso de tecnologia na área.

A conectividade vem se tornando crucial entre os novos eletrônicos, mas, a implementação dessas tecnologias possui custos, que vão além do gasto com a aquisição de equipamentos. Em se tratando de redes de computadores, os meios de acesso à internet são limitados à banda disponível e a infraestrutura existente. Uma vez que essa limitação existe, é necessário que haja um maior investimento financeiro para realizar a expansão dessa infraestrutura, à medida que a quantidade de dispositivos e usuários crescem. Portanto, nesse contexto, é necessária uma avaliação de custo x benefício dos investimentos realizados.

A capacidade do ambiente reflete diretamente na execução dos serviços e tem impacto relevante na balança de qualidade de serviço e qualidade de experiência. Logo, a avaliação de capacidade é indispensável em ambientes corporativos, pois, os custos de infraestrutura e a qualidade do serviço prestado são determinantes para o rendimento da empresa.

Tendo esse cenário em vista e o estudo de casos reais em uma grande empresa pública, esta pesquisa tem como objetivo propor novas metodologias de avaliação de capacidade de rede apoiadas em teoremas de probabilidade e, por conseguinte, trazer dados significantes que possam embasar as necessidades de ampliação da infraestrutura a partir da relação de custo x benefício.

Um desses teoremas utilizados é o Teorema Central do Limite da probabilidade. Este, somado a coletas de dados reais por meio do protocolo SNMP (Simple Network Management Protocol), auxiliam o processo de avaliação de capacidade, a qual apresenta um intervalo de confiança de probabilidade e, possibilita que as conclusões desta análise respondam à relação custo x benefício de forma pragmática.

O artigo está organizado como segue. O capítulo 2 trata de trabalhos relacionados. Nele é abordado os fundamentos teóricos em que se baseiam os pilares da pesquisa. O capítulo 3 discorre sobre a metodologia aplicada para se alcançar o objetivo mencionado. O capítulo 4 aborda as propostas sugeridas e o capítulo que segue traz as análises e resultados obtidos. Encerra-se esse trabalho com as considerações finais e sugestões de projetos futuros.

2 REFERENCIAL TEÓRICO

Em uma pesquisa realizada por Gartner, Zwicker e Rödder (2009) foi testado um modelo baseado na função de produção de Cobb-Douglas, que apresentou indícios de que o aumento de investimento na área de TI foi acompanhado de um acréscimo positivo nas receitas. O trabalho mostrou, ainda, indícios de que houve eficiência marginal nos investimentos em tecnologia da informação.

Ao se estabelecer que as empresas desenvolvem a TI por meio dos Investimentos em Tecnologia da Informação (ITI), pode-se situar essa análise de efeitos indiretos no âmbito econômico. Nesse caso, o foco dessa análise baseia-se na mensuração do retorno dos ITI, operados por cada empresa, sendo esses investimentos aqui definidos como gastos em hardware e seus aplicativos, software e seus recursos, sistemas de telecomunicações, gestão de dados. Em se tratando de uma grande empresa pública que visa lucros, as adaptações e as atualizações da tecnologia de produção e de serviços são fatores de extrema relevância.

Para se obter essa mensuração do retorno dos ITI, utiliza-se de artifícios métricos da área de finanças e contábeis. Um dos muitos indicadores de desempenho existentes para avaliar o chamado custo-benefício é o ROI (*Return On Investment*), também conhecido como taxa de retorno sobre investimento. Consiste em uma métrica utilizada para mensurar o rendimento obtido com uma dada quantia de recursos. O ROI é dado pela razão entre o lucro líquido alcançado e o investimento efetuado dentro de um dado período. Para calcular o ROI pode-se subtrair o ganho obtido a partir do investimento pela quantia gasta com o investimento e dividindo o resultado novamente pela quantia gasta com o investimento.

Julga-se ser bastante simples a realização desse cálculo. Entretanto, em se tratando de projetos de TI, seja em investimentos de aquisição de software ou hardware, a definição de expectativa ou cálculo posterior e quanto se obteve de ROI podem ser bastante complicados. Isso acontece, pois os custos envolvidos, ou TCO (*Total Cost of Ownership*), podem ser muito variados e difíceis de serem mensurados. Portanto, ainda que o ROI seja uma medida bastante relevante, não deve ser a única a se considerar numa avaliação.

Quando se compra um equipamento espera-se que ele seja utilizado para produzir valor, seja um computador ou máquina, sua aquisição está condicionada à geração de lucros. Entretanto, bens como esses sofrem um efeito conhecido como depreciação e estão sujeitos a outros detalhes que afetam seu rendimento, como manutenções ou mesmo a obsolescência. Tendo em vista esse cenário, a reutilização desses equipamentos em outra área da empresa, por exemplo, é uma prática muito bem vista, pois gera um menor "desperdício" desses bens.

Outros artifícios métricos que são bastante utilizados para gestão de ativos são o CAPEX e o OPEX. CAPEX é a sigla da expressão inglesa *Capital Expenditure* (em português, despesas de capital ou investimento em bens de capital) que designa o montante de dinheiro despendido na aquisição (ou introdução de melhorias) de bens de capital de uma determinada empresa. O CAPEX é, portanto, o montante de investimentos realizados em equipamentos e instalações de forma a manter a produção de um produto ou serviço ou manter em funcionamento um negócio ou um determinado sistema (VERNIMMEN et al., 2014).

OPEX é uma sigla derivada da expressão *Operational Expenditure*, que significa o capital utilizado para manter ou melhorar os bens físicos de uma empresa, tais como equipamentos, propriedades e imóveis. As despesas operacionais (muitas vezes abreviado a OPEX) são os preços contínuos para dirigir um produto, o negócio, ou o sistema. O seu contrário, despesas de capital (CAPEX), refere-se ao preço de desenvolvimento ou fornecimento de partes não consumíveis do produto ou sistema (VERNIMMEN et al., 2014).

A gestão de ativos atende a demandas de planejamento dos custos de ciclo de vida dos ativos físicos. Estes custos, segundo (SINISUKA; NUGRAHA, 2013), incluem os determinísticos (como aquisição e descarte) e os probabilísticos (como os de mão de obra de reparo e demanda de sobressalentes). Lloyd et al. (2010) afirma que apenas 20% do custo de vida é gasto durante projeto e aquisição, mas estas fases definem 80% do custo de ciclo de vida. Ou seja, a manutenção responde pela maior parte dos custos de vida, mas por não atuar ativamente na especificação, projeto, aquisição, construção e instalação dos ativos, não pode atuar de maneira totalmente eficaz no controle destes custos.

Um fato importante é que os ativos em empresas como a em análise operam 24 horas por dia, e as avarias interrompem a operação de vários sistemas operacionais, causando grandes prejuízos indesejados.

2.1 Fundamentos Teóricos e Práticos

Uma análise é realizada a partir de dados obtidos através de coleta ou simulação do ambiente avaliado.

Rapidamente, a simulação traz, a baixos custos, dados passíveis de análise, agilizando a avaliação e possibilitando prever diferentes necessidades a partir de modelos de cenários esperados para a infraestrutura simulada. Porém, este ambiente simulado é controlado e, portanto, pode não refletir a realidade do ambiente. Ferramentas como o NS-3 (*Network Simulator 3*) e *Cisco Packet Tracer* são exemplos de simuladores de rede que permitem a criação dos cenários e realização da avaliação de forma controlada.

A coleta dos dados diretamente dos dispositivos é um procedimento com custo elevado e lento. Por se tratar de dados reais, esta coleta permite que seja realizada uma análise consistente da estrutura de rede avaliada. Porém, para que se tenha dados suficientes para uma análise será necessário o armazenamento durante um prazo mínimo conforme a necessidade de avaliação. Este prazo pode ser de minutos, para avaliações de saúde do ambiente, ou anos, para projeções anuais de crescimento, por exemplo. O SNMP é um protocolo que permite o acesso aos dispositivos de rede de forma simples, realizando as coletas de dados necessários, que servirão para a avaliação do ambiente.

Com os dados coletados e armazenados passa-se para a metodologia de avaliação da infraestrutura. Fundamentado nos resultados teóricos obtidos com os cálculos definidos na metodologia, tem-se informações suficientes para realizar a análise de capacidade da infraestrutura.

Esta pesquisa tem como objetivo propor a utilização conjunta da coleta de dados do protocolo SNMP e o Teorema Central do Limite da probabilidade, para avaliação da capacidade, apresentando no final um intervalo de confiança de probabilidade e, então, ser capaz de agregar valor às métricas de perdas, descartes, retransmissões, indisponibilidade, entre outras, possibilitando que as conclusões da análise respondam a relação custo x benefício de forma pragmática.

2.1.1 SNMP – *Simple Network Management Protocol*

O SNMP é um protocolo que permite o acesso aos dispositivos de rede, realizando as coletas de dados necessários, que servirão para a avaliação do ambiente. Este protocolo permite o gerenciamento de dispositivos de rede remotamente através de operações simples. Atualmente o protocolo SNMP está em sua terceira versão. Para que o dispositivo seja gerenciado é exigido a utilização do protocolo IP (*Internet Protocol*). Este protocolo utiliza um modelo com duas entidades: o gerente e o agente.

O gerente que é um servidor com a função de NMS (*Network Management System*) é, portanto, responsável pelo recebimento e solicitação de informações. O agente é um software instalado no dispositivo que provê as informações ao gerente. A busca de informações realizada pelo gerente é denominada *polling* e o agente retorna uma resposta com os dados solicitados pelo gerente. Além da resposta de um *polling*, o agente realiza uma ação que é denominada *trap*, que é a informação enviada ao gerente da rede para comunicar qualquer alteração de status.

O protocolo SNMP utiliza o protocolo UDP (*User Data Protocol*) como o protocolo da camada de transporte entre a comunicação das entidades. O UDP possui menos *overhead* que o protocolo TCP (*Transmission Control Protocol*) e não realizada a confirmação de entrega dos dados, assim, a escolha de utilização do UDP gera melhor

performance, apesar de não oferecer a garantia na comunicação.

Todas as informações acessadas pelo protocolo SNMP estão contidas na MIB (*Management Information Base*). As MIBs são compostas por módulos, podendo ser padronizados ou patenteados, que representam um conjunto de objetos. Assim, se faz necessário um levantamento de quais MIBs estão disponíveis e as informações que podem ser acessadas pelo gerente da rede. Um dos principais grupos de gerenciamento é o MIB-II e é padrão em qualquer MIB, portanto, carrega informações consideradas essenciais para o gerenciamento.

As ferramentas que o protocolo dispõe permite a monitoração da rede possibilitando até a configuração dos equipamentos gerenciados. Assim, o estudo e detalhamento do protocolo SNMP e das MIBs é de extrema importância para a coleta dos dados de forma eficiente para qualquer avaliação de capacidade de rede.

Contudo, para a realização deste estudo, não será construído de um script para realizar a monitoração via SNMP, pois, conforme já citado, o processo de construção de uma base com volume de dados grande o suficiente para a realização das análises é lento. Portanto, uma base de um ano de coleta, já disponibilizada em um banco de dados Oracle, será a fonte dos dados deste trabalho.

2.1.2 SPSS Modeler

O SPSS Modeler é um software de mineração de dados que, de forma prática e visual, cria modelos de predição. Ele foi projetado com o objetivo de obtenção de melhores resultados para as empresas centralizando todo o processo de mineração e gerando informações que podem subsidiar as tomadas de decisão.

Conforme declarado no documento de guia do Usuário Modeler, a ferramenta oferece métodos de modelagem a partir de estatística, inteligência artificial e Machine Learning. Para cada amostragem existe uma predição adequada e com o SPSS é possível identificar os métodos mais indicados para problemas específicos.

Com um ambiente visual simples, é possível explorar as possibilidades e os modelos disponíveis sem necessidade de ter um conhecimento estatístico prévio, assim, o usuário pode se concentrar em ajustar os relacionamentos das variáveis mineradas e analisar os resultados obtidos.

A conexão com o banco de dados permite ao SPSS a exploração de grandes massas de dados, a manipulação, modelagem e mineração através de uma sequência de operações e a gravação dos resultados em tabelas do mesmo banco.

Como o objetivo desta pesquisa é a aplicação do Teorema Central do Limite em dados de predição para a observação de intervalos de confiança, não se aprofundará nos

conceitos e construção das predições do SPSS Modeler, ou de qualquer outra ferramenta preditiva, e assim, focara-se em analisar os resultados obtidos.

2.1.3 Linguagem Python de Programação

Segundo o site oficial, o python[1] é uma linguagem programação interpretada, orientada a objetos e de alto nível, com semântica dinâmica. Sua estrutura de dados e sintaxe simples colaboram para o fácil aprendizado e reduz o custo de manutenção. O suporte a módulos e pacotes, que proporciona ao python o aproveito de códigos de maneira prática, incentiva a modularidade e o reuso de código. O próprio site apresenta comparações[2] com outras linguagens, apresentando os pontos fortes e fracos e indica, por exemplo, que um mesmo programa escrito em Python pode apresentar estrutura de 3 a 5 vezes menor do que o Java e de 5 a 10 vezes menor do que o C++, o que destaca sua simplicidade de estrutura e sintaxe.

A produtividade do python é enfatizada devido à sua característica de depuração rápida do programa. A linguagem não é compilada, acelerando a fase de teste do código. O interpretador do código executará uma exceção ao encontrar um erro conhecido ou um rastreamento da pilha é impresso quando o erro não é identificado.

A partir destas características e o fato de estar disponível em formato binário e sem nenhum custo para todas as principais plataformas, a linguagem python ganhou espaço no desenvolvimento ágil de aplicativos por sua eficácia.

Existem duas versões do Python usualmente utilizadas no mercado: o Python2 e o Python3, em que a versão mais recente é o Python 3.7.1, sendo esta a versão utilizada no código deste estudo.

2.1.4 Probabilidade e Estatística

Segundo Barbetta, Reis e Bornia (2004), estatística é a ciência que se ocupa de organizar, descrever, analisar e interpretar dados para que seja possível a tomada de decisões e/ou a validação científica de uma conclusão. A Estatística Descritiva é, em geral, utilizada na etapa inicial da análise, quando se tem contato com os dados pela primeira vez (MAGALHÃES; LIMA, 2010).

Inferência Estatística é o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados, das informações e conclusões obtidas a partir de subconjuntos de valores, usualmente de dimensão muito menor (MAGALHÃES; LIMA, 2010). Ou seja, o objetivo da Inferência Estatística é tirar conclusões sobre a população com base na

[1] https://www.python.org/doc/essays/blurb/
[2] https://www.python.org/doc/essays/comparisons/

informação fornecida por uma amostra. Quando se tem acesso a todos os elementos que deseja se estudar, não é necessário o uso das técnicas de inferência estatística.

Uma vez tendo coletado os dados, seja através de censo ou por amostragem, é preciso resumi-los e organizá-los de maneira a permitir uma primeira análise, e posterior uso das informações (BARBETTA; REIS; BORNIA, 2004). Ainda, segundo o autor, uma das formas de resumir o conjunto de dados é por meio das medidas de síntese ou estatísticas. As principais estatísticas são a média, o desvio padrão, a variância e a proporção.

- Média: trata-se de uma estatística que caracteriza o "centro de massa" do conjunto de dados (valor esperado).
- Variância: trata-se de uma estatística que mede a dispersão em torno da média do conjunto (em torno do valor esperado), possuindo uma unidade que é o quadrado da unidade da média.
- Desvio padrão: é a raiz quadrada positiva da variância, tendo portanto uma unidade que é igual à unidade da média, sendo muitas vezes preferida para efeito de mensuração da dispersão.
- Proporção: consiste em calcular a razão entre o número de ocorrências do valor de interesse de uma variável qualitativa e o número total de ocorrências registradas no conjunto.

O conhecimento das distribuições amostrais das principais estatísticas é necessário para fazer inferências sobre os parâmetros do modelo probabilístico da população. Uma vez que se têm as informações das distribuições amostrais, é possível obter-se gráficos bastante úteis para análises e projeções. Um modelo desses gráficos é o histograma. Para a obtenção deste, basta aglomerar os dados da amostra de acordo com a frequência que os mesmos aparecem.

A distribuição das médias amostrais da variável pode ser aproximada por uma distribuição normal. Quanto maior o tamanho da amostra mais o histograma aproximar-se-á de uma distribuição normal, independentemente do formato da distribuição da variável na população (BARBETTA; REIS; BORNIA, 2004).

Ao observar uma distribuição e realizar uma estimativa sobre ela, procurando inferir as probabilidades das ocorrências a partir de algum modelo, entra-se na disciplina de probabilidade. Segundo BUSSAB, tais modelos são chamados de modelos probabilísticos (BUSSAB; MORETTIN, 2017).

A probabilidade condicional refere-se ao cálculo de uma estimativa onde já se é conhecida uma situação prévia que pode interferir ou não no resultado. Sua fórmula é definida como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

Onde a interseção entre A e B pode ser definida pela equação 2.2

$$P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(A|B) \tag{2.2}$$

E por fim, pode-se encontrar uma fórmula simples para calcular a probabilidade condicional:

$$P(A|B) = \frac{P(A) \times P(A|B)}{P(B)} \tag{2.3}$$

2.1.5 Teorema Central do Limite (TCL)

De forma empírica, Teorema Central do Limite de-se que quanto mais distante da média, menos provável a ocorrência de um evento, e por consequência, quanto mais próximo da média, mais vezes esperamos que seja obtido o valor. Portanto, quanto mais eventos forem experimentados, é esperado um grande volume de resultados próximos da média e poucos resultados distantes da média. Assim, observa-se a formação da curva de Gauss.

A partir desta avaliação empírica nota-se que o Teorema Central do Limite afirma que a soma de N variáveis aleatórias e independentes, com qualquer distribuição e variâncias iguais, é equivalente a uma variável com distribuição próxima à normal (distribuição de Gauss) quando se tem um valor de N elevado. Para variáveis aleatórias não identicamente distribuídas o teorema também é válido, porém, é necessário que sejam satisfeitas algumas condições, que não serão necessárias para o escopo desta pesquisa, pois obtém-se os dados a partir de um único ambiente que segue a mesma distribuição.

Bussab e Morettin (2017) apresenta o modelo Normal originado com Gauss realizando estudos sobre os erros de observações astronômicas, por volta de 1810, gerando assim o nome de distribuição gaussiana. Por meio desse modelo de distribuição, se torna possível inferir sobre dados esperados através dos resultados de média e desvio padrão obtidos de coletas suficientemente grandes do ambiente a ser avaliado. Ou seja, para uma amostra aleatória $(X_1, X_2, ..., X_2)$ de tamanho n, obtém-se valores de média $= \mu$ e variância $= \sigma^2$, e, à medida que o número n de amostras cresce, a distribuição da média se aproxima da distribuição normal.

Moore (2017) apresenta três razões para a grande importância da distribuição Normal na estatística: A primeira razão é devido a sua característica de descrever bem os dados de coleta reais com características aproximadas. A segunda razão é a sua boa aproximação para resultados aleatórios. A terceira razão é que na inferência estatística, a

distribuição Normal também tem boa aproximação de outras distribuições aproximadamente simétricas.

Moore (2017) ainda define a regra 68-95-99,7, em que ele afirma que apesar das diferentes curvas normais, todas elas apresentam uma propriedade em comum que é esta regra, que, na distribuição Normal com média μ e desvio padrão σ tem-se:

- Regra 68: Aproximadamente 68% das observações estão a menos de σ de distância da média.
- Regra 95: Aproximadamente 95% das observações estão a menos de duas vezes σ de distância da média.
- Regra 99,7: Aproximadamente 99,7% das observações estão a menos de três vezes σ de distância da média.

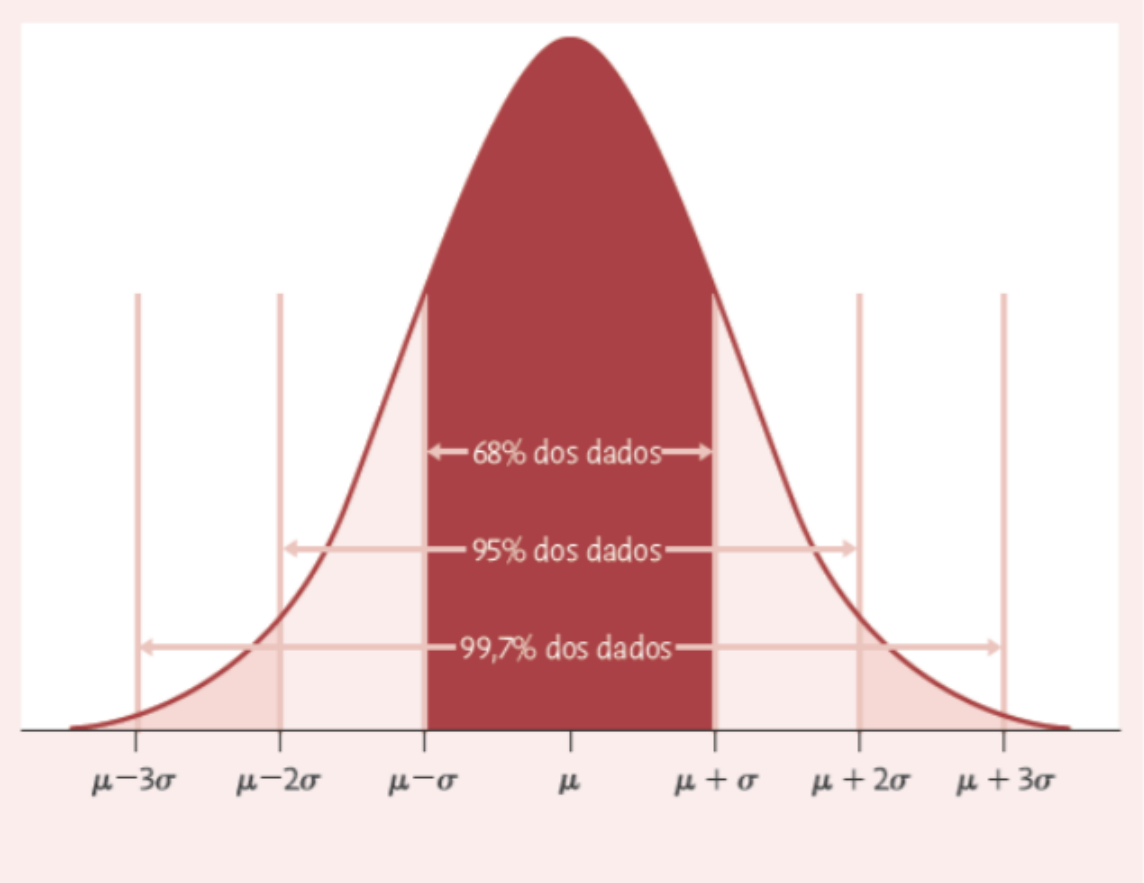

Figura 1 – A regra 68-95-99,7 para distribuições Normais. Fonte: Moore (2017)

Segundo o autor, a regra 68-95-99,7 descreve distribuições que são exatamente Normais. Então, para cada regra tem-se por consequência o percentual de observações que estão fora da regra. Ou seja, se proximamente 68% das observações estão a menos de

(sigma) da média, então 32% das observações estão a mais de (sigma) da média conforme o gráfico abaixo.

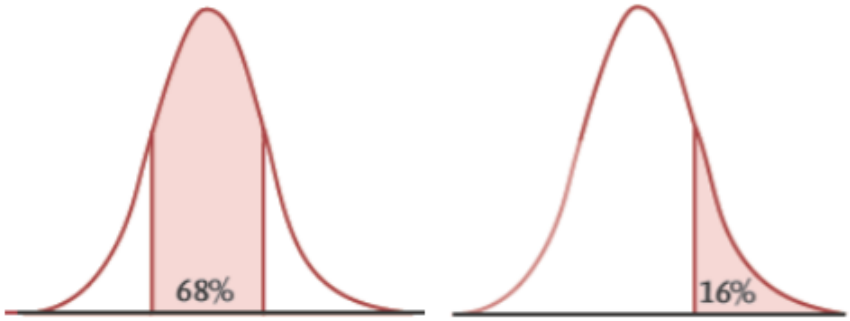

Figura 2 – A regra 68 da distribuição Normal e seus complementos. Fonte: Moore (2017)

Bussab e Morettin (2017) define que a variável aleatória igual a 1 tem distribuição normal para os parâmetros μ e σ^2 se sua densidade é dada pela equação 2.4:

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \tag{2.4}$$

E, portanto, pode-se provar que a integral definida com intervalo de ∞ até ∞ tem resultado igual a 1, ou seja, abrange 100% das variáveis.

Utilizando a função de densidade de probabilidade (PDF – *Probability Density Function*) pode-se encontrar a probabilidade de uma variável estar em um certo intervalo semiaberto $(X_1, X_2]$. Assim, o Teorema Central do Limite determina que:

$$Pr(Yn \leq Y) = \int_{-\infty}^{x_1} \frac{1}{\sqrt{2\pi}} e^{\frac{-(Y-\mu)^2}{2\sigma^2}} d\left[\frac{Y-\mu}{\sigma}\right] \tag{2.5}$$

E consequentemente, a soma normalizada tende a uma variável aleatória gaussiana:

$$Yn \rightarrow N(0, 1) \tag{2.6}$$

Porém, precisa-se considerar uma soma não normalizada de variáveis aleatórias, assim, obtendo uma variável aleatória Zn, com, $Zn = \sum_{i=1}^{n} x_i = \sigma\sqrt{n}(yn - n\mu)$ e para valores grandes de amostras (n), também pode-se aproximá-la por uma Gaussiana:

$$Zn \approx N(n\mu, n\sigma^2) \tag{2.7}$$

Logo, pode-se resumir e definir que a probabilidade de um evento estar dentro de um intervalo definido é dado por:

$$PrPr(Z < X) - N(n\mu, n\sigma^2) = 1 - Q(n\mu, n\sigma^2) \tag{2.8}$$

Onde:

$$n\mu, n\sigma^2 = \frac{X - \mu}{\sigma} \tag{2.9}$$

Então, com a ajuda do Teorema Central do Limite pode-se estabelecer um intervalo de confiança baseado em probabilidade para valores de média e desvio padrão obtidos a partir de coletas realizadas na rede. Pode-se alcançar também, para o mesmo intervalo de confiança, os valores de máximo e mínimo esperados utilizando a PDF (Função de Densidade de Probabilidade) para uma distribuição de variável aleatória conhecida.

A tabela com os valores da distribuição Normal está anexada ao final do relatório como Anexo A.

3 METODOLOGIA

Parte da estratégia está baseada na implementação de dois *trade-offs* de ciclo de vida, suportando a tomada de decisão para gestão de ativos, que consistem em levar em conta a soma destes custos ao tomar decisões de especificação, aquisição, operação, manutenção, reforma e descarte de ativos.

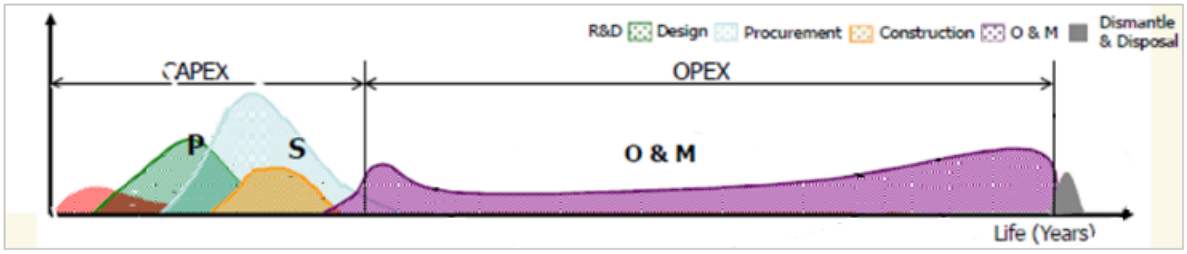

Figura 3 – Custos ao longo da vida de um ativo. Fonte: Azevedo (2007)

O *trade-off* CAPEX/OPEX é explicado por autores como Blanchard, Fabrycky e Fabrycky (1990), Lloyd et al. (2010), Kelly (2006), Azevedo (2007), que afirmam que cerca de 80% dos custos de ciclo de vida de um ativo são operacionais. Porém, as decisões tomadas durante o projeto definem 80% do custo ao longo de toda a vida. Ou seja, a busca por reduzir custos de manutenção tomando ações em ativos em idade operacional não é suficiente para otimização do custo. Para isso, é necessário manter uma política adequada de investimentos de capital, mantendo a base de ativos com a capacidade otimizada.

O *trade-off* CURTO/LONGO PRAZO se baseia na ideia de que existe um momento ideal de substituição de cada ativo, chamado vida econômica. Isto é definido por Woodward (1997) como o período em que a obsolescência econômica demanda a substituição por uma alternativa de menor custo. Os custos de manutenção e operação crescem com o tempo, devido à degradação dos ativos; por isso é preciso monitorar esta evolução e considerar os custos em longo prazo e não cair no raciocínio errôneo de que "o ativo já está pago, então este gasto vale à pena". Como já mencionado anteriormente, é de boa prática a reutilização de ativos que podem ser obsoletos para algum setor, mas de grande utilidade para outro.

Além do custo de manutenção dos equipamentos existe o custo de prestação de serviço por parte das operadoras de telecomunicações. Estes contratos elevam o custo operacional agregando um alto custo de manutenção que também tendem a crescer com o tempo. Este crescimento, diferente do crescimento devido à degradação dos ativos, tem outros fatores: crescimento do uso de internet, crescimento no tamanho dos dados de navegação devido às melhorias de qualidade e tamanho do tráfego, atualização de preços

por parte das prestadoras de serviço, entre outros.

Como fonte de dados para a avaliação deste custo operacional, serão utilizadas as coletas de *throughput* realizadas via SNMP dos circuitos da empresa pública avaliada. As amostras de dados utilizadas serão mascaradas e estarão sem quaisquer informações relativas aos equipamentos de rede, devido às condições de sigilo que devem ser mantidas. Desta forma, utilizou-se dados brutos reais de um ambiente de data center com classificação Tier4 para realizar a predição de consumo com o SPSS Modeler, realizar a avaliação da capacidade da infraestrutura de rede de comunicação no ponto de vista de banda disponível e, por fim, o dimensionamento e planejamento de contratação dos serviços de telecomunicação com a utilização do Teorema Central do Limite para fornecer uma nova metodologia baseada em probabilidade.

3.1 Codificação: Scripts em python para automação dos cálculos

Para a aplicação do Teorema Central do Limite e obtenção das métricas citadas utiliza-se a linguagem Python para criar um código de programação em que são realizados todos os cálculos e armazenamento dos valores necessários para a validação das métricas. O Python possui inúmeras bibliotecas e pacotes e auxiliam a codificação e implementação de forma simples, além de possuir as bibliotecas necessárias para conexão direta com o banco de dados utilizado, possibilitando a criação de rotinas de automação dos processos de análise.

Através do protocolo SNMP, uma rotina, já existente, realiza a coleta dos dados do roteador através das funções de GET a cada cinco minutos, totalizando 288 coletas. Através do SNMP coletam-se quaisquer informações da MIB do roteador, porém, para esta seção, o cerne se baseia no consumo de banda ou vazão de dados. Este dado, coletado em bits por segundo (bps) é armazenado no banco de dados Oracle, permanecendo disponível para consulta de quaisquer ferramentas que possuam acesso à base Oracle.

Os códigos criados e implementados estão nos anexos B e C, respectivamente, e seguem os seguintes fluxogramas:

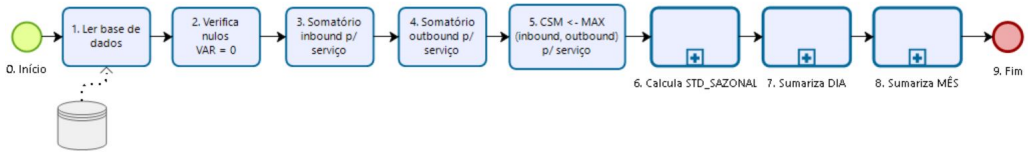

Figura 4 – Fluxograma do código de agregação de dados.

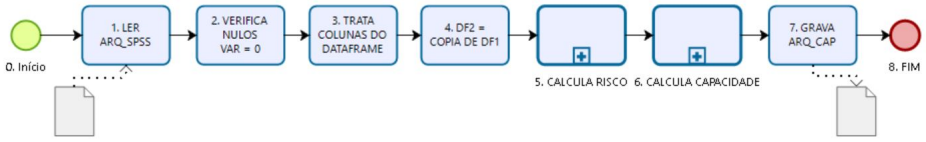

Figura 5 – Fluxograma do código de risco condicional

3.1.1 Bibliotecas Python

Os códigos apresentados nos fluxogramas, que realizam todo o procedimento de sumarização e aplicação do Teorema Central do Limite, não foram construídos com a biblioteca padrão Python, ou seja, foram importadas bibliotecas para auxiliar a sua criação. As bibliotecas python, assim como em outras linguagens, são arquivos que contém funções que podem ser incorporadas a outros programas. "*Quando as rotinas em uma biblioteca são linkeditadas com o restante do seu programa, apenas as funções que seu programa realmente usa são carregadas e linkeditadas.*" (SCHILDT; MAYER, 1997). Ou seja, assim como na linguagem C, usando bibliotecas, pode-se importar somente as funções que se tem necessidade para a sua construção.

Dentre as bibliotecas importadas, serão listadas as mais importantes, ou seja, que tiveram atuação direta na construção do código para a implementação do teorema do limite central, e houve a necessidade de importação das mesmas.

3.1.1.1 cx_Oracle

A biblioteca cx_Oracle possibilita o acesso a bases de dados oracle. Ela está atualmente na versão 7.1.3. Além da instalação desta biblioteca, para garantir a conexão com a base de dados oracle foi necessária a instalação de um cliente oracle (*Oracle Instant Client*), pois, o python e o banco de dados não estão na mesma máquina (ORACLE, 2019).

Assim, pode-se executar o passo "1. LER BASE DADOS" apresentado no fluxograma 1. Utilizando a biblioteca cx_Oracle do Python conecta-se com o banco de dados Oracle e realiza-se a operação de seleção para obter os dados armazenados (ORACLE, 2019).

3.1.1.2 Pandas

"A biblioteca Pandas é um pacote Python que fornece estruturas de dados rápidas, flexíveis e expressivas, projetadas para tornar o trabalho com dados estruturados (tabulares, multidimensionais, potencialmente heterogêneos) e séries temporais fáceis e intuitivos." (trecho retirado de (PANDAS, 2019)).

Devido à sua poderosa estrutura a biblioteca pandas é amplamente utilizada em sistemas de análise de dados, aprendizado de máquina, *big data*, séries temporais, etc. Na construção dos códigos apresentados nos fluxogramas, o Pandas está presente em todo o projeto. Com esta biblioteca recebe-se os dados da seleção realizada com a biblioteca Oracle em um *dataframe* e, com a estrutura de um *dataframe* pandas, executa-se inúmeras funções: tratamento de campos nulos, gravação de arquivos em formato xlsx (formato de tabelas Excel), filtros de conteúdo por colunas existentes, criação de novas colunas calculadas, entre outras (PANDAS, 2019).

3.1.1.3 NumPy

"NumPy é o pacote fundamental para computação científica com Python." (trecho retirado de (NUMPY, 2019)).

A biblioteca NumPy é uma biblioteca que possui recursos que facilitam o tratamento e a execução de cálculos que, às vezes, podem ser relativamente complexos de serem programados em qualquer linguagem. Nos códigos utilizados, três funções da biblioteca foram fundamentais: mean, std, NaN. A biblioteca NumPy também possui a função *percentile*, mas, este cálculo foi programado manualmente devido à sua simplicidade e à complexidade encontrada em configurar corretamente a interpolação e outros argumentos da função numpy.percentile. Portanto, a função da biblioteca NumPy foi, essencialmente, simplificar (NUMPY, 2019).

A função *numpy.mean* poderia facilmente ser programada manualmente, porém, como a biblioteca já estava importada, utilizá-la foi consequência da necessidade das outras funções. A função *numpy.std* é a função mais utilizada no código de sumarização e seu resultado, o desvio padrão de uma amostra, é de fundamental importância que seja precisamente calculado, para que se obtenham os cálculos corretos das probabilidades de risco e da aplicação da regra 68-9599,7 de Moore (2017) para obtenção dos máximos e mínimos de uma amostra normal. A função *numpy.NaN* aparece na característica de tratamento de dados da biblioteca. Esta função é autoexplicativa: *"not a number"* (traduzido do inglês: não é um número), ela verifica se a variável é não numérica. Utilizada, sucessivamente, em todos os processos de leitura de dados e antes da gravação dos arquivos sumarizados, esta função simples, protege o programa de receber valores que não são calculáveis e mantém uma base de dados confiável para os cálculos do Teorema Central do Limite.

3.1.1.4 Outras

Outras bibliotecas foram utilizadas nos scripts, porém, não tiveram função para a implementação do Teorema Central do Limite, mas constaram no código para auxiliar de

alguma forma a sua construção. A tabela abaixo apresenta cada biblioteca e sua respectiva função no código:

Tabela 1 – Bibliotecas Python e suas funções nos scripts

BIBLIOTECA	FUNÇÃO NO CÓDIGO
os	Manipulação do prompt de execução do python.
random	Criação de uma máscara randômica junto com a senha para mascarar dados.
scipy	Análise estatística.
warning	Impedimento de impressão de warnings na manipulação dos dataframes pandas.
matplotlib.pyplot	Geração dos gráficos para visualização dos resultados.
oracleclass	Importação do cx_Oracle e conexão com o banco de dados.

3.1.2 Agregação dos dados e métricas

Os dados de consumo de banda dos equipamentos coletados e armazenados fornecem as informações necessárias para gerenciar a capacidade da infraestrutura, assim como possibilitam a monitoração de serviços e aplicações. Segundo o autor Ogbonna (2017) em seu livro *A-Z of Capacity Management*: "Os dados desempenham um papel proeminente no gerenciamento, e sua importância não pode ser enfatizada demais". A qualidade das informações fornecidas pelos dados depende muito da granularidade da coleta inicial, granularidade da agregação e o método de agregação utilizado. Portanto, devem ser avaliados os três pontos apresentados pelo autor para a obtenção de resultados satisfatórios.

3.1.2.1 Granularidade da coleta inicial

O primeiro ponto abordado pelo autor, para garantir a qualidade das informações, é a granularidade da coleta inicial dos dados. As informações de consumo com esta granularidade são utilizadas para calcular o desvio padrão sazonal do consumo dentro de um mesmo mês, comparando a diferença de tráfego do serviço em um mesmo momento, que não seria eficiente após a sumarização da coleta. Este subprocesso é definido pelo fluxograma na figura 6 abaixo:

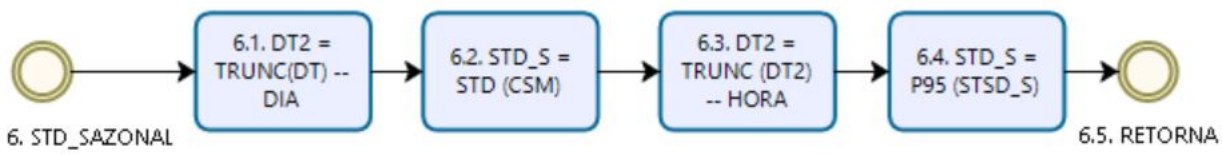

Figura 6 – Fluxograma do subprocesso de cálculo de desvio padrão.

No código criado, as informações de consumo são selecionadas de um banco de dados oracle, conforme citado na seção 2.1.1 Esta operação de seleção de dados é realizada com as cláusulas de condição restringidas às informações já contidas na base acerca do tipo de serviço de rede, que podem conter uma ou mais interfaces que atendem esta aplicação selecionada.

A granularidade destes dados é uma média de cinco minutos e no caso de serviços que são vinculados a mais de uma interface de rede, o tráfego do serviço corresponde ao valor da soma do tráfego individual no mesmo período. Esta soma é um processo de agregação que será apresentado e explicado no item de Granularidade da agregação.

Com a granularidade definida para cinco minutos, obtém-se então durante um dia o número de 288 amostras de dados e um total de 8640 amostras durante um mês (considerando um mês com 30 dias). Quanto menor é a granularidade definida, maior é a qualidade dos dados, porém, os impactos na infraestrutura coletada também serão proporcionalmente aumentados com a redução da média de coleta. Observando este impacto, Dominic Ogbanna relata em seu livro acerca da granularidade da coleta: "Portanto, deve-se tomar extremo cuidado para garantir que as atividades de coleta, agregação e armazenamento de dados sejam feitas corretamente, sem afetar o desempenho da infraestrutura de destino que fornece os dados." (OGBONNA, 2017). Assim, a média de coleta adotada é adequada para o tamanho do parque observado, pois, além de se adequar às recomendações mais comuns de intervalo médio de coleta citados pelo mesmo autor, que são as médias de 15 segundos, 30 segundos, 60 segundos e 5 minutos, com esta média de coleta garantiu-se uma performance razoável de processamento e armazenamento sem perder a qualidade da informação

3.1.2.2 Granularidade da agregação

A agregação dos dados é a forma utilizada para agrupar um volume de dados e representá-los como um único valor. Conforme o autor Dominic Ogbonna: "Agregação de dados é uma técnica de resumir inúmeros pontos de dados coletados em um drive de capacidade específico." (OGBONNA, 2017). Assim, foi realizada a agregação de valores das coletas para se obter dados mais adequados para a projeção de consumo.

A primeira agregação realizada, citada na seção de Granularidade da coleta inicial, é a soma da banda utilizada das diferentes interfaces de um mesmo serviço. A necessidade de se agrupar o consumo de diferentes interfaces é importante para a determinação de um consumo do serviço que está sendo avaliado, pois, como este é suportado por mais de um caminho, a junção de todas as possibilidades fornece a infraestrutura que deve estar à disposição para manter as funcionalidades da aplicação.

Para agregar o consumo de banda por serviço o script soma-se os consumos de banda do mesmo período de coleta, ou seja, que possuem valores iguais de data (*times-*

*tamp**) em que foi realizada a coleta dos dados via SNMP e armazenada no banco de dados. No fluxograma 1, esta soma é realizada nos passos 3 e 4, onde soma-se separadamente os tráfegos de download (*inbound*) e upload (*outbound*) da base de dados.

A tabela 2 apresenta dados fictícios de consumo de rede para a ilustração do modelo de soma para a agregação.

Tabela 2 – Tabela com dados para exemplificar modelo de soma

SERVIÇO	INTERFACE	DATA	CONSUMO
Rede 1	1	15/05/2019 20:15:00	1000
Rede 2	1	15/05/2019 20:15:00	2000
Rede 1	2	15/05/2019 20:15:00	3000
Rede 1	3	15/05/2019 20:15:00	4000
Rede 2	2	15/05/2019 20:15:00	5000
Rede 1	4	15/05/2019 20:15:00	6000
Rede 3	1	15/05/2019 20:15:00	7000

Para a Rede 1, no momento da coleta às 15/05/2019 20:15:00 o consumo é de:

Rede 1 = 1000 + 3000 + 4000 + 6000 = 14000.

Para a Rede 2, no momento da coleta às 15/05/2019 20:15:00 o consumo é de:

Rede 2 = 2000 + 4000 = 7000.

Para a Rede 3, no momento da coleta às 15/05/2019 20:15:00 consumo é de:

Rede 3 = 7000.

A granularidade com períodos de tempo maiores que as coletas é explicada pelo autor Dominic Ogbonna: "O valor agregado de uma métrica é o valor único que pode ser usado na capacidade de relatório e outras funções relacionadas à capacidade, como modelagem. Por exemplo, se uma ferramenta de monitoramento coleta uma métrica específica a cada 5 segundos, então, em um dia ela teria coletado 21600 pontos de dados. No entanto, com a agregação de dados, você pode ter um valor agregado para cada dia ou hora." (OGBONNA, 2017). Com esta citação exemplifica-se a agregação por períodos de coleta e, nas figuras abaixo, o fluxograma dos subprocessos de sumarização e processamento das datas:

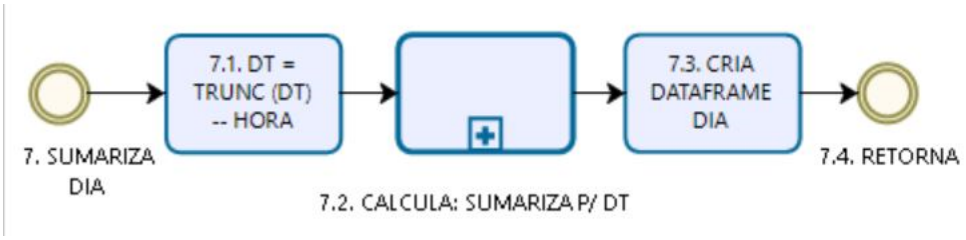

Figura 7 – Fluxograma do subprocesso de sumarização diária

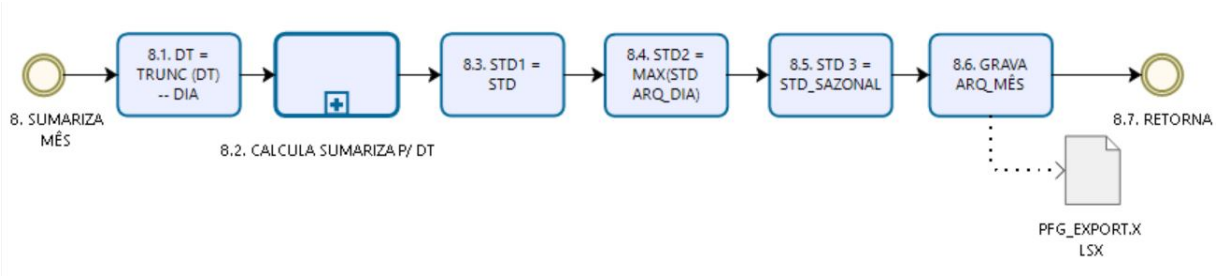

Figura 8 – Fluxograma do subprocesso de sumarização mensal

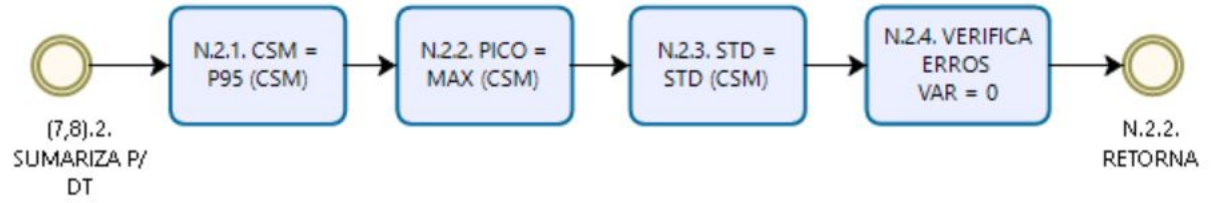

Figura 9 – Fluxograma do subprocesso de cálculo da sumarização

Ainda, de acordo com o Ogbonna (2017), os tipos mais comuns de agregação são: segundo, minuto, hora, dia, semana, mês, quadrimestre e ano. Cada uma destas métricas, de acordo com o autor, possui seus casos de uso mais indicados. A tabela 3 apresenta os casos de uso das métricas utilizadas neste tabalho descritas pelo autor no livro *A-Z of capacity management*:

Tabela 3 – Métricas utilizadas neste trabalho. Descrição retirada de Ogbonna (2017)

MÉTRICA	**DESCRIÇÃO/CASO DE USO***
Minuto	Medindo atividades do usuário ou carga de trabalho de um aplicativo.
Dia	Para análise semanal e mensal do relatório; Bom para modelar e prever o uso de recursos.
Mês	Para relatórios históricos e comparativos; Bom para relatórios de suporte do plano de capacidade.

3.1.2.3 Método de agregação

A partir das métricas já escolhidas, em relação à sua granularidade, é necessário realizar matematicamente a sua sumarização. A definição do método de agregação é um processo importante, pois, este cálculo servirá como base para as comparações e análises de todo o consumo da infraestrutura.

"O método de agregação de dados que você escolher determinará as saídas obtidas seu processo de gerenciamento de capacidade e as decisões resultantes que você faz. Consequentemente, o método de agregação específico adotado para um processo deve ser alinhado às características do conjunto de dados." (OGBONNA, 2017).

Os métodos mais comuns de agregação dos dados são: média, mediana, máximo e percentil. Todos as métodos, conforme a necessidade da aplicação e do conjunto de dados coletados, possuem características que podem apresentar mais ou menos benefícios. Neste trabalho, utilizam-se três métodos de agregação a fim de obter o melhor resultado para modelagem: média, máximo e percentil.

3.1.2.3.1 Média

Segundo Moore (2017): "A média de um conjunto de observações é sua média aritmética. Se considerarmos as observações como pesos distribuídos ao longo de uma varinha, a média é o ponto no qual a varinha se equilibraria."

Complementando o entendimento de Moore (2017), o autor Ogbonna (2017) cita algumas vantagens e desvantagens do uso de média em avaliações de capacidade, dentre elas, ele cita a média como boa medida para indicação de ocupação do recurso e distribuição uniforme para detecção de tendências e situações atípicas, porém, a média esconde os máximos e mínimos de utilização, conforme descrição de Moore, e, se utilizada como medida de planejamento, pode elevar o risco de esgotamento de recurso por não atendimento durante eventos de alta utilização.

3.1.2.3.2 Máximo

O ponto de máximo consumo, definido no código construído como "pico de consumo", é definido no livro de Ogbonna (2017) como ATH (todo tempo alto, do inglês, *all-time-high*). Esta métrica é o valor mais alto de uma determinada medida e pode ser utilizada como ponto de referência para decisões relacionadas à capacidade da infraestrutura, porém, devido a sua característica volátil não é um valor adequado para planejamentos e modelos preditivos.

Os picos de utilização servirão como ponto de referência para a determinação de uma área de risco que será apresentada no capítulo 5, oferecendo um valor real comparativo para adequação da necessidade de banda, visto que, mesmo tratando-se de um evento atípico, não há motivos para justificar o esgotamento dos recursos nestes eventos.

3.1.2.3.3 Percentil

O percentil é uma medida que divide uma amostra em cem partes organizados de forma crescente. Pode-se calcular o valor do percentil conforme a descrição no livro

Practical Statistics for Data Scientists: "O valor tal que P por cento dos valores assume esse valor ou menos e (100-P) por cento assume esse valor ou mais." (BRUCE; BRUCE, 2017).

Moore (2017) destaca em seu livro que se deve procurar por valores atípicos, buscando corrigi-los ou justificar sua remoção antes de realizar os procedimentos z ou outra inferência com base estatística, assim, com o intuito de obter dados mais próximos da distribuição normal e com críticos mais precisas de avaliação.

"Para as métricas de recurso do sistema, o método de agregação baseado em percentil é mais indicado para planejamento de capacidade, pois pode ser utilizado para eliminar picos pontuais e focando em picos sustentados." (OGBONNA, 2017). Com esta assertiva, o autor indica a utilização do percentil para a agregação das métricas que serão utilizadas no planejamento de capacidade.

De acordo com (OGBONNA, 2017), os valores de percentil mais comumente utilizados são os percentis 95, 98 e 99. O percentil 95 é bem aceito e utilizado no mercado como indicador de avaliação de rede. Diversos *papers* e empresas utilizam tal medida para definir os picos de utilização retirando da amostra os valores falsos positivos, ou seja, os picos pontuais.

Um artigo de plano de capacidade de rede que ilustra a utilização do percentil 95 pelo mercado é o "*Movinar Network Capacity Planning*" de 2016, escrito em parceria pelas empresas Moviri, Entuity e BMC. No ano em que foi escrito, dentre as empresas participantes do artigo, a mais recente no mercado era a Moviri, com 16 anos, enquanto a BMC já alcançava seus 46 anos de fundação.

Outros exemplos de empresas que utilizam o percentil 95: Produto IP on Demand[1] e a Semaphore[2] . Exemplos de papers que também abordam o assunto: Century Link: Percentil de Pagamentos[3] , Para métodos de pagamento[4] . E uma tese de Mestrado na Universidade do Minho: Monitorização da Qualidade de Serviço da Rede [5].

Assim todos os cálculos de percentil executados no script construído para sumarizar as coletas brutas de dados são realizados com o valor de percentil 95. Além de se obter as médias de consumo com o valor percentual, outras medidas que precisavam ser agregadas também foram sumarizadas pelo percentil. Pode-se observar a aplicação do percentil nos dois fluxogramas indicados pela Sigla "P95".

[1] http://www.w8telecom.com.br/pagina-inicial
[2] https://www.semaphore.com/95th-percentile-bandwidth-metering-explained-and-analyzed/
[3] https://www.nanog.org/meetings/nanog53/presentations/Monday/AElcanFin.pdf
[4] https://www.tik.ee.ethz.ch/file/2083f48f3be1ba7b12a9af1e6db71ca2/p95pam.pdf.1.pdf
[5] https://repositorium.sdum.uminho.pt/bitstream/1822/27508/1/Tese_Pedro%20Queir%C3%B3s_2013.pdf

3.1.3 Cálculos executados

A fim de atender todas as necessidades e processos da proposta, o script construído conta também com resultados gerados a partir de fórmulas criadas manualmente. Nesta seção, estes métodos serão expostos e explicados para que sejam esclarecidos suas necessidades.

3.1.3.1 Desvio Padrão utilizando regra 68-95-99,7 de Moore

Quando realiza-se a agregação dos dados, independente do método escolhido para a agregação, está se realizando um resumo de toda a informação coletada. Contudo, mesmo sendo necessário para o planejamento de capacidade, esse processo oculta a variação existente no ambiente, portanto, é necessária uma forma de calcular a dispersão das informações antes da sumarização.

A variância, conforme definido em 2 é a medida que calcula esta dispersão dos dados, mas, por ser da unidade quadrática seu valor não é diretamente comparável à amostra, por isso, a utilização do desvio padrão é mais comum como definido em 1.1.1.3. Além disso, a fórmula geral do Teorema Central do Limite possui como uma de suas variáveis o desvio padrão. Assim, definir o valor de desvio padrão tem grande importância para o cálculo dos limites de confiança, pois, é a partir dele que, matematicamente, as curvas de máximos e mínimos se aproximam ou afastam da média de consumo.

A forma definida para o estimação do desvio padrão consiste em calcular a variação do consumo de banda dentro do mesmo período de tempo (timestamp) dentro de um mesmo mês (da mesma forma que na agregação do consumo) e calcular o valor de percentil 95, removendo os falsos positivos assim como no cálculo da média.

A partir da definição dos desvios padrão, pode-se então utilizar a regra de Moore para definir os valores que estão dentro ou fora de uma distribuição com comportamento Normal.

3.1.3.2 Teorema Central do Limite

O Teorema Central do Limite é a base fundamental para as métodologias que serão propostas neste trabalho e por isso sua definição é muito importante para a fundamentação do estudo. De forma simplificada, utiliza-se a equação 2.5 apresentada no capítulo 2, em dois momentos distintos do código, chamando os subprocessos Calcula Risco e Calcula Capacidade para determinar o valor de duas variáveis: O Risco e a Capacidade.

No primeiro momento, calcula-se o risco baseado no valor da capacidade que é fixo. Este risco, portanto, varia mês a mês conforme a coleta é realizada. Em um segundo momento do *script*, o risco é fixado, e, então, é calculado um valor de capacidade. Esta matemática reversa é possível pois a média e o desvio padrão da coleta são conhecidos.

Os subprocessos para o cálculo do risco e da capacidade são apresentados nas figuras abaixo:

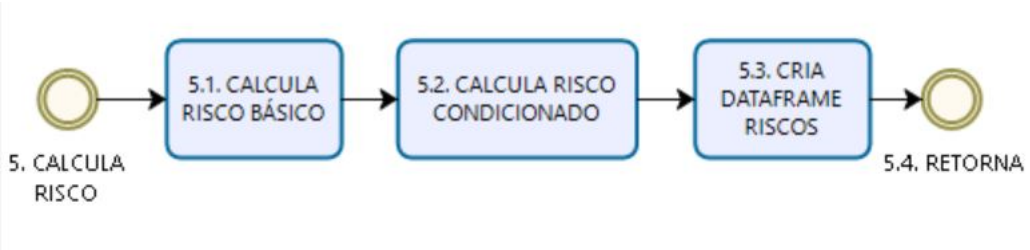

Figura 10 – Fluxograma do subprocesso de cálculo de risco

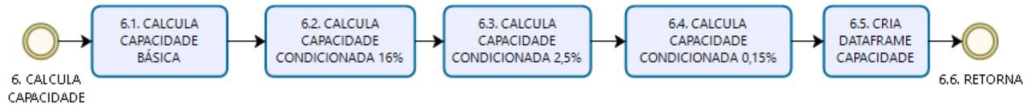

Figura 11 – Fluxograma do subprocesso de cálculo de capacidade

3.1.3.3 Probabilidade Condicional

A probabilidade condicional, definida em 2.1.4, foi utilizada para o cálculo dos risco condicionado que será apresentado no próximo capítulo. O risco condicional está incluído no código na mesma posição do cálculo do Teorema Central do Limite, pois, a probabilidade condicional é relativa ao Teorema Central do Limite e, portanto, estão na mesma função do *script*.

4 PROPOSTAS

Conforme o objetivo desta pesquisa, a definição de metodologias para a avaliação de rede é uma necessidade clara para a avaliação do custo x benefício de um investimento em expansão de capacidade da infraestrutura. Com a metodologia definida, esta deve ser utilizada para a determinação de parâmetros que justificarão as necessidades de *upgrade*, *downgrade* ou manutenção da capacidade instalada ou provisionada.

De forma geral, a capacidade de qualquer ambiente definido baseado em métodos de agregação (médias, máximos, medianas, etc). Porém, nesta pesquisa, com o auxílio do Teorema Central do Limite, calculam-se probabilidades de variáveis aleatórias normais, condicionadas a situações previstas, atingirem valores acima dos limites superiores, ou seja, estima-se o risco de uma amostra estar além da capacidade esperada.

4.1 Desvio Padrão relativo

O cálculo do desvio padrão apresentado na seção 2.1.4 tem grande relevância para esta avaliação conforme suas definições e necessidade para o cálculo do Teorema Central do Limite. Sua fórmula padrão é calculada pela diferença das amostras em relação a sua média. Assim, é possível escrever a fórmula completa do teorema como:

$$Pr(Z < X) = N((X - x)/\sigma) = N\left((X - x)/\sqrt{\sum_{i=0}^{n}(x_i - x)^2/(n-1)}\right) \tag{4.1}$$

Contudo, conforme a metodologia definida, utiliza-se como média o valor do percentil 95, e portanto, para equilibrar o resultado ao utilizar o Teorema Central do Limite, pode-se reescrever a mesma fórmula substituindo o x pelo percentil 95:

$$N\left((X - p95)/\sqrt{\sum_{i=0}^{n}(x_i - p95)^2/(n-1)}\right) \tag{4.2}$$

Com essa alteração, propõe-se adequar o desvio padrão utilizado no Teorema Central do Limite à mesma grandeza de média, substituindo em toda a fórmula do teorema e não em uma só parte. Com o aumento do valor calculado do desvio, o impacto observado nesta alteração indica uma dispersão maior dos dados, devido à distância observada entre os valores menores da amostra e o percentil. Assim, ao se observar a curva normal, nota-se a menor concentração no centro da *gaussiana*, e por consequência, o aumento de observações conforme a regra de Moore indicada na seção 2.1.5.

4.2 Risco: Uma Probabilidade condicional

A probabilidade condicional, conforme estabelecido em 2.1.4, retrata o resultado de um evento baseado em outro já conhecido. Quando utilizamos o percentil trata-se os dados para que apresente informações mais relevantes para a avaliação do ambiente, porém, a situação de risco ocorre exatamente nos momentos descartados pelo percentil e, por isso, o cálculo de um valor de risco é condicionado às situações atípicas, onde há possibilidade de escassez de recursos.

Portanto, para propor um indicador de verificação do risco a partir das coletas de consumo e da capacidade da infraestrutura, pode-se inferir a utilização da seguinte fórmula para o risco de forma condicionada:

$$Pr(x > C|x > p95) \tag{4.3}$$

Observa-se que a proposta da equação 4.3 consiste em calcular a condição de que uma variável aleatória normal seja maior que a capacidade (C) dado que já é sabido que esta mesma variável aleatória é maior que a média calculada com o percentil 95, ou seja, se obtém, a probabilidade condicional de uma variável distribuída normalmente, dentre as descartadas pelo percentil, esgotar os recursos de capacidade. Assim, utilizando a definição de probabilidade condicional, é possível determinar:

$$Pr(x > C|x > p95) = \frac{Pr(x > C|x > p95) \times Pr(x > C)}{Pr(x > p95)} \tag{4.4}$$

Ao definir esta expressão da equação 4.4, pode-se aferir algumas informações a fim de resumir o cálculo que deverá ser utilizado na obtenção do risco.

A primeira informação da equação 4.4 apresenta a probabilidade condicional de uma variável aleatória ser maior que sua média dada a informação que ela é superior a sua capacidade. O resultado desta expressão é 1, pois, para qualquer caso que a variável aleatória seja superior a sua capacidade, necessariamente esta variável é maior que o valor do percentil 95 da amostra.

$$Pr(x > p95|x > C) = 1 \tag{4.5}$$

A segunda informação extraída da equação 4.4 é a probabilidade de uma variável aleatória ser superior a sua média. Esta probabilidade pode ser calculada por um evento simples, onde escolhe-se ao acaso o conjunto de amostras que sejam maiores que a média e calcula-se a proporção do conjunto total de amostras. Como esta amostra média é um percentil 95, nota-se que a probabilidade de se obter ao acaso um valor maior que o

percentil é exatamente o total de amostras não contempladas neste percentil, conforme a eq. 4.6 abaixo:

$$Pr(x > p95) = 1 - Pr(x < p95) = 0,05 \tag{4.6}$$

Portanto, utilizando as equações 4.5 e 4.6 pode-se resumir a probabilidade condicional proposta no estudo como a equação 4.7:

$$Pr(x > p95|x > C) = \frac{Pr(x > C)}{0,05} \tag{4.7}$$

Ou pela equação 4.8:

$$Pr(x > p95|x > C) = 20 \times Pr(x > C) \tag{4.8}$$

Esta é uma aproximação válida, porém, deve-se notar que à medida que a probabilidade da distribuição normal ser maior que a capacidade, apresentada na equação 4.9 abaixo, o resultado da equação 4.8 pode ser superior a 100%.

$$Pr(x > C) \tag{4.9}$$

Isso ocorre, pois, o número de amostras superiores à condição é fixo conforme a equação 4.6 enquanto o número de amostras acima da capacidade é flutuante, e, para os casos onde a equação 4.9 tem valor maior que a equação 4.6, significaria que a variável definida para a capacidade é menor que o valor do percentil. O fato de se calcular o valor utilizando dois modelos de distribuição justifica a probabilidade fora do intervalo de uma proporção real variando de 0 a 1.

Quando se trata de probabilidade e estatística, uma observação com 100% de probabilidade já resulta em todos os resultados possíveis, portanto, para a proposta, a obtenção de valores superiores a 1 são limitados superiormente em 100%, ou seja, para todas as amostras maiores que o percentil dado, todas serão maiores que a capacidade do ambiente, portanto a probabilidade é igual a 1.

4.3 Capacidade definida pelo risco

A capacidade determinada para um recurso qualquer define o limite de produção máximo alcançável, e por isso, sua determinação é tão importante. Ao longo deste trabalho, definiram-se metodologias para o apontamento e/ou criação de indicadores que sirvam de base para o ajuste da capacidade de um ambiente e, assim, foi apresentado o risco como uma probabilidade condicionada em 4.2.

Ao se definir um planejamento de capacidade, entende-se este como o processo pelo qual procura-se garantir que a infraestrutura oferecida atenda todas as necessidades do cliente e/ou aplicação. Deste modo, é comum que a capacidade seja calculada baseada no consumo médio observado, sendo, portanto, um valor proporcional à média estabelecida.

Quando é fixado o valor de capacidade para o ambiente, baseado em um valor médio de consumo, entende-se que esta capacidade é proporcional à média. Obtém-se, assim, um valor de proporção para estimação da capacidade, conforme ilustrado na equação 4.10 abaixo:

Para valores fictícios de Média = 8, Capacidade=10, Proporção = 125%

$$125\% = \frac{10}{8} \tag{4.10}$$

Neste pequeno exemplo, tem-se que a capacidade é 25% superior ao consumo médio. Porém, este valor proporcional é reflexo somente da média apurada para o período, e, portanto, não reage com o desvio padrão no mesmo período. Neste sentido, utilizando a proposta estabelecida em 4.2, obtêm-se resultados diferentes de risco para cada mês.

Deste modo, apresenta-se a proposta de obter o valor de capacidade baseado no valor de risco, a fim de se estabelecer ao longo do período planejado um risco máximo de esgotamento de recurso bem definido que pode ser calculado apoiado no Teorema Central do Limite em sintonia com a proposta de um risco condicional.

Assim, a partir da equação 4.2.6, é possível obter a capacidade a partir de um risco informado chegando à equação abaixo:

$$Pr(x > p95|x > C) = 20 \times Pr(x > C) \tag{4.11}$$

Logo:

$$Pr(x > C) = \frac{Pr(x > C|x > p95)}{20} \tag{4.12}$$

Lembrando que:

$$Pr(x > C) = 1 - N(z) \tag{4.13}$$

e

$$z = (C - \mu)/\sigma \tag{4.14}$$

Assim, a partir das equações 4.12 e 4.13 têm-se que:

$$N(z) = 1 - Pr(x > C|x > p95)/20 \tag{4.15}$$

Com o resultado da equação 4.13 e com a tabela Normal padrão (anexo A), encontra-se o valor de Z, e assim, entendendo como N^{-1} como a inversa da normal:

$$z = N^{-1}(1 - (Pr(x > C|x > p95))/20) \tag{4.16}$$

Utilizando a equação 4.14, obtém-se a equação:

$$C = \mu + \sigma * N^{-1}(1 - (Pr(x > C|x > p95))/20) \tag{4.17}$$

Portanto, quando define-se o risco para o período como um risco fixo Rf, obtém-se a equação que estabelece a capacidade baseada em um risco.

$$C = \mu + \sigma * N^{-1}(1 - (Rf/20) \tag{4.18}$$

Com esta fórmula estabelecida, entende-se que, para um período especificado, a capacidade do ambiente estará baseada em um cálculo bem definido e justificada pelo próprio histórico de consumo, tendo em vista o risco máximo que o ambiente deve estar preparado para suportar ou o nível de confiança e segurança que este deve apresentar, tendo em vista a distribuição normal de seu comportamento.

5 RESULTADOS E DISCUSSÕES

Neste capítulo, apresentaremos os resultados das propostas apresentadas no capítulo anterior. Ao longo da apresentação dos resultados, as devidas discussões serão realizadas a fim de se obter, ainda nesta seção, algumas conclusões acerca dos indicadores obtidos.. Para estes resultados, foram avaliados oito serviços, nomeados como rede 1 a rede 8, e suas respectivas interfaces que os compõem.

5.1 Agregação dos dados

Na seção 3.1.2.2 foi citado o fato de que precisa-se agrupar os consumos de rede pelo serviço que é oferecido, devido ao fato do roteamento ocorrer em diferentes interfaces e mudanças nas configurações das próprias interfaces. Serão apresentados os gráficos das redes 5 e 7 e suas respectivas interfaces a fim de exemplificar esta agregação. Após as análises individuais, será apresentado uma análise geral dos resultados em relação à agregação dos dados de consumo.

5.1.1 Gráficos de consumo da Rede 5 e análise dos dados

A rede 5 é composta pelas interfaces: interface 51, interface 52, interface 53, interface 54. Esta rede possui características de volumes constantes de dados com seus máximos de utilização muito próximo do percentil 95 em resposta da constante utilização pelas aplicações e clientes que utilizam esta rede.

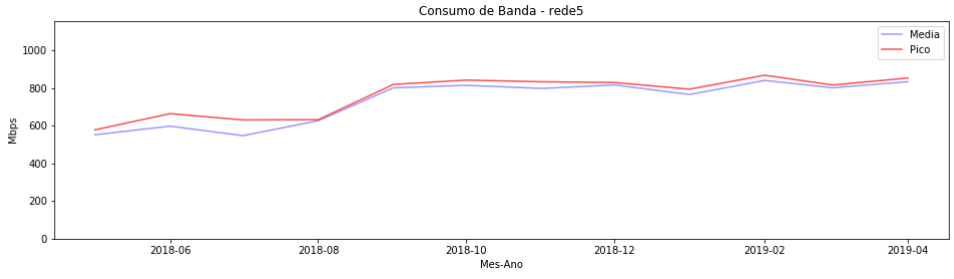

Figura 12 – Consumo de banda - Rede 5

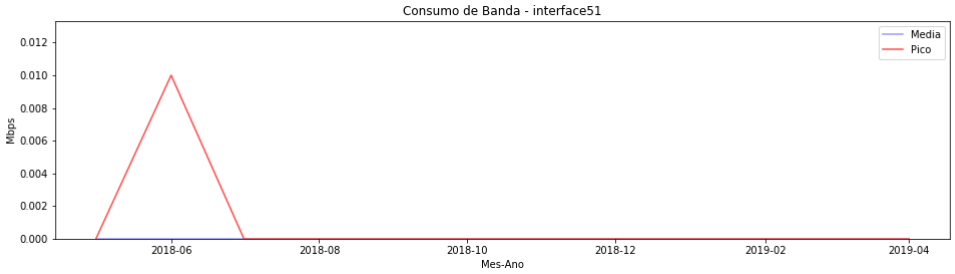

Figura 13 – Consumo de banda - Interface 51

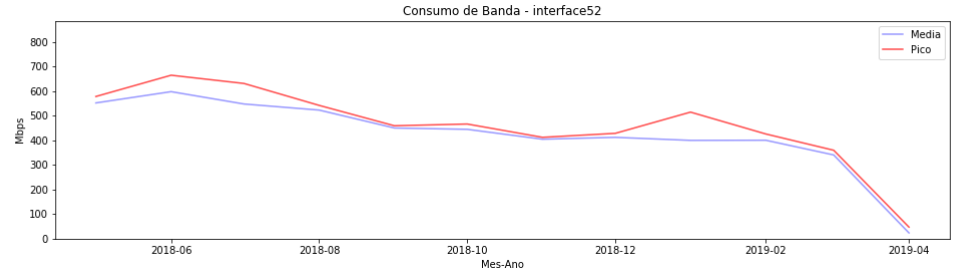

Figura 14 – Consumo de banda - Interface 52

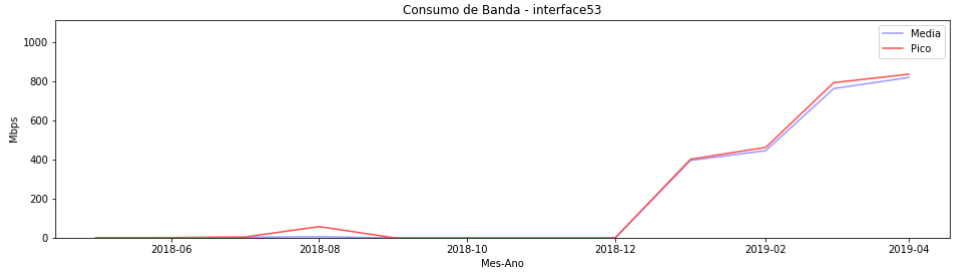

Figura 15 – Consumo de banda - Interface 53

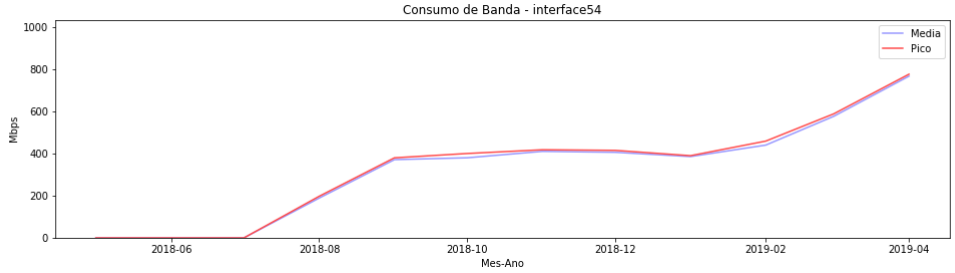

Figura 16 – Consumo de banda - Interface 54

A tabela 4 abaixo apresenta os dados dos gráficos no mês de março:

Tabela 4 – Dados dos gráficos no mês de março - Rede 5

DATA	INTERFACES	MÉDIA	PICO
01/03/2019	interface51	0.0	0.0
01/03/2019	interface52	340.88	359.95
01/03/2019	interface53	763.40	793.54
01/03/2019	interface54	577.30	589.02
01/03/2019	rede5	802.18	816.64

A tabela 5 abaixo apresentam os dados dos gráficos no mês de abril:

Tabela 5 – Dados dos gráficos no mês de Abril - Rede 5

DATA	INTERFACES	MÉDIA	PICO
01/04/2019	interface51	0.0	0.0
01/04/2019	interface52	24.21	47.79
01/04/2019	interface53	820.13	836.88
01/04/2019	interface54	767.99	777.18
01/04/2019	rede5	834.27	853.83

Observa-se nos gráficos que as variações de consumo ao longo dos períodos apresenta uma grande variação. Esta variação, porém, é completamente absorvida quando realiza-se a agregação do período mais bruto da coleta dos dados comprovando a necessidade da unificação do consumo de banda das interfaces.

Com os dados observados nas tabelas, é possível verificar a grande diferença entre o consumo real da aplicação e a soma dos volumes de tráfego das interfaces após as interfaces. No caso das tabelas apresentadas, é possível verificar que a soma das interfaces sumarizadas chega a ser superior a 1600 Mbps, o que corresponde aproximadamente a 200% o volume da rede 5.

5.1.2 Gráficos de consumo da Rede 7 e análise dos dados

A rede 7 é composta pelas interfaces: interface 71, interface 72, interface 73, interface 74, interface 75 e interface 76. Esta rede é caracterizada por ser utilizada por clientes internos, ou seja, os usuários dessa aplicação são conhecidos, portanto, seu consumo possui desvio padrão relativamente baixo.

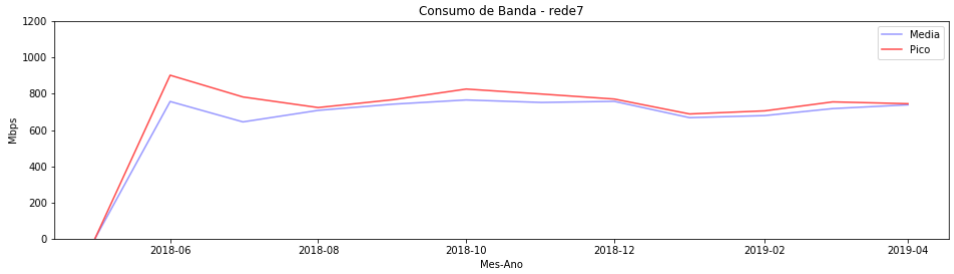

Figura 17 – Consumo de banda - Rede 7

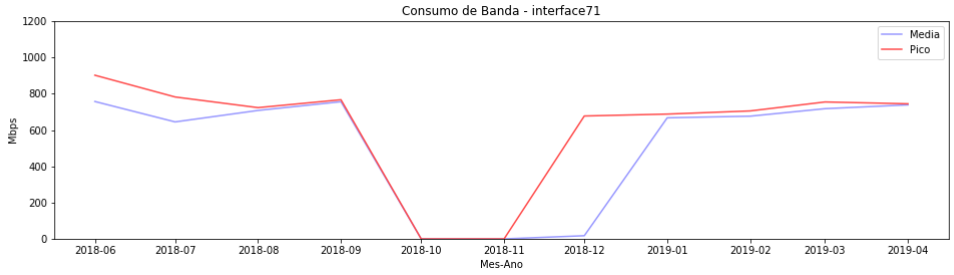

Figura 18 – Consumo de banda - Interface 71

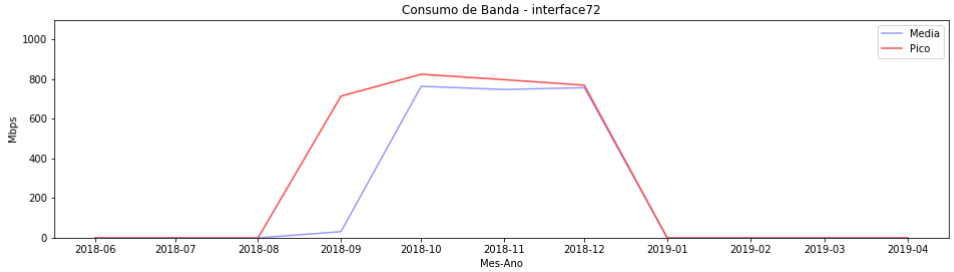

Figura 19 – Consumo de banda - Interface 72

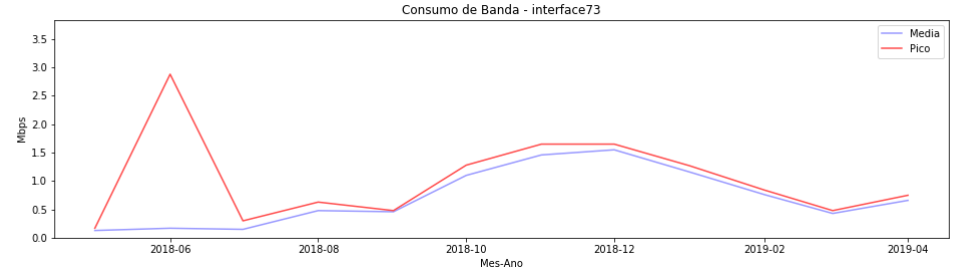

Figura 20 – Consumo de banda - Interface 73

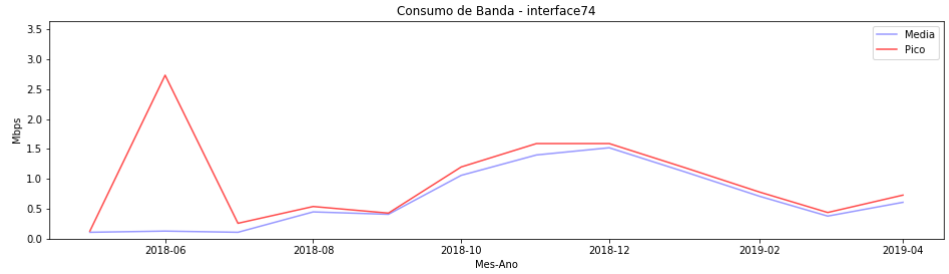

Figura 21 – Consumo de banda - Interface 74

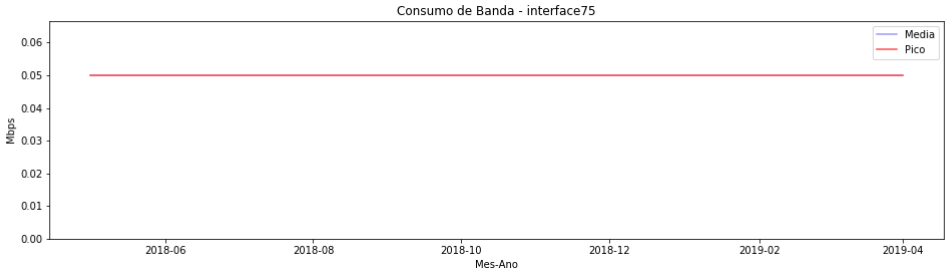

Figura 22 – Consumo de banda - Interface 75

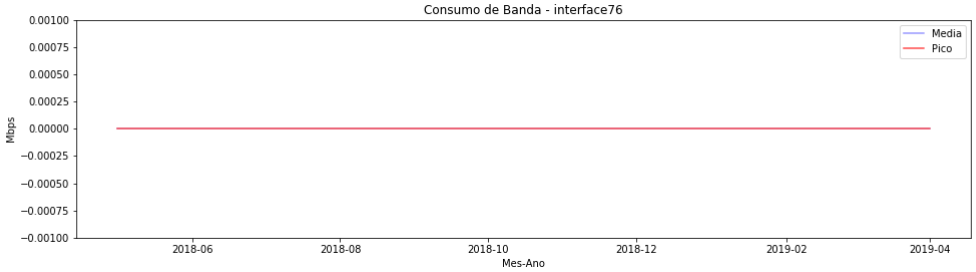

Figura 23 – Consumo de banda - Interface 76

A tabela 6 abaixo apresenta os dados dos gráficos no mês de março:

Tabela 6 – Dados dos gráficos no mês de Março - Rede 7

DATA	**INTERFACES**	**MÉDIA**	**PICO**
01/03/2019	interface71	717.54	754.55
01/03/2019	interface72	0.0	0.0
01/03/2019	interface73	0.43	0.48
01/03/2019	interface74	0.38	0.44
01/03/2019	interface75	0.05	0.05
01/03/2019	interface76	0.0	0.0
01/03/2019	rede7	718.13	755.12

A tabela 7 abaixo apresenta os dados dos gráficos no mês de abril:

Tabela 7 – Dados dos gráficos no mês de Abril - Rede 7

DATA	**INTERFACES**	**MÉDIA**	**PICO**
01/04/2019	interface71	738.53	744.52
01/04/2019	interface72	0.0	0.0
01/04/2019	interface73	0.66	0.75
01/04/2019	interface74	0.61	0.73
01/04/2019	interface75	0.05	0.05
01/04/2019	interface76	0.0	0.0
01/04/2019	rede7	738.95	745.12

Conforme observa-se nas tabela e nos gráficos, o tráfego é direcionado é configurado para a interface 71, ocasionando uma subutilização dos outros links que permanecem em regime de *stand by*, ou seja, permanecem sem utilização efetiva até que sejam ativados por alguma necessidade. Esta necessidade é observada pela ativação da interface 72 que absorveu o tráfego que deveria estar na interface comumente utilizada. Mais uma vez, nota-se visão da rede de forma absoluta e sem perdas de informação de consumo mesmo com o desvio do tráfego entre as interfaces, estabelecendo uma confiança maior na observação e caracterização da curva de consumo do serviço.

5.1.3 Análise geral de agregação de consumo por serviço

De modo geral, a visão por serviço é necessariamente mais complexa pois se trata de um conjunto de informações individuais, pois, vai depender de uma boa modelagem dos dados e da verificação constante da configuração da infraestrutura que é dinâmica.

Além da necessidade de uma base de configuração atualizada, foi demonstrada a necessidade da soma dos intervalos de consumo previamente à sumarização a fim de garantir a maior proximidade com a situação real da situação da rede. Em contrapartida, o volume de armazenamento destas informações sem sumarização e o processamento para se obter a visão por serviço é muito custosa. Visto que para uma coleta com intervalos de 5 minutos, tem-se em média 288 amostras por dia, para a execução deste estudo com 42 interfaces em 12 meses de coleta, estima-se o processamento e armazenamento superior a 4,3 milhões de linhas para apenas 8 redes.

Contudo, apesar das dificuldades de modelagem e processamento, nota-se positivamente que a agregação estabelece uma melhor condição de avaliação de modo geral. A visualização é mais clara do comportamento da rede para responsáveis pelo suporte, a projeção de consumo é mais uniforme ao longo do tempo proporcionando um melhor dimensionamento, redução de erros por superdimensionamento ocasionado pela agregação feita após a sumarização evitando custos excedentes desnecessários para atender o serviço e permite estabelecer melhor o custo benefício da aplicação em relação ao custo de infra.

5.2 Resultados do desvio Padrão Relativo

Conforme descrito em 4.1, o desvio padrão tem grande relevância na distribuição normal alterando o comportamento da curva, por isso, é importante a análise do desvio padrão proposto. Em paralelo com esta análise, precisa-se identificar a normalidade da distribuição dos dados coletados nas interfaces. Para isso, podem ser utilizados gráficos de distribuição de frequência (histogramas), que são muito utilizados em análises estatísticas, projetados juntamente com a curva de distribuição normal desta frequência.

A curva gaussiana é distribuída em torno do centro e dispersa conforme o valor do desvio padrão, assim, alterando estes parâmetros pode-se comparar a normalidade da distribuição das amostras e ao mesmo tempo verificar a proposta de se utilizar um desvio padrão relativo à média como percentil.

Para exemplificar esta análise, utilizou-se os gráficos das redes 5 e 7, que já foram caracterizadas na seção 5.1, e suas respectivas interfaces. Após as análises individuais, será apresentado uma análise geral dos resultados em relação ao desvio padrão relativo proposto.

5.2.1 Gráficos de Frequência da Rede 5 e análise dos dados

As características da rede 5 estão descritas na seção 5.1.1.

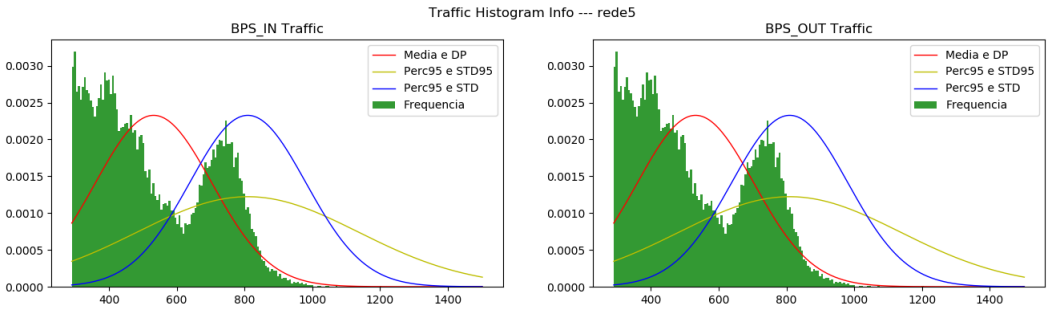

Figura 24 – Histograma de tráfego - Rede 5

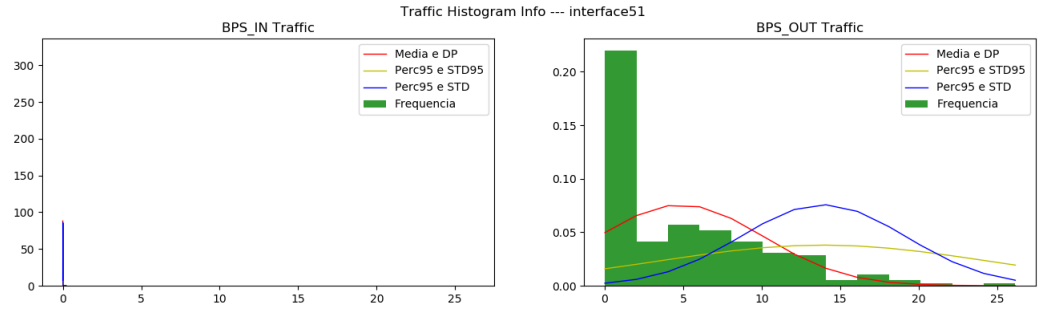

Figura 25 – Histograma de tráfego - Interface 51

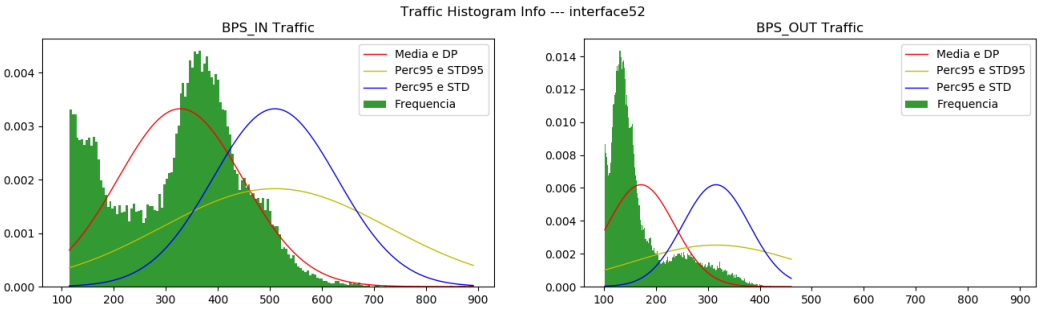

Figura 26 – Histograma de tráfego - Interface 52

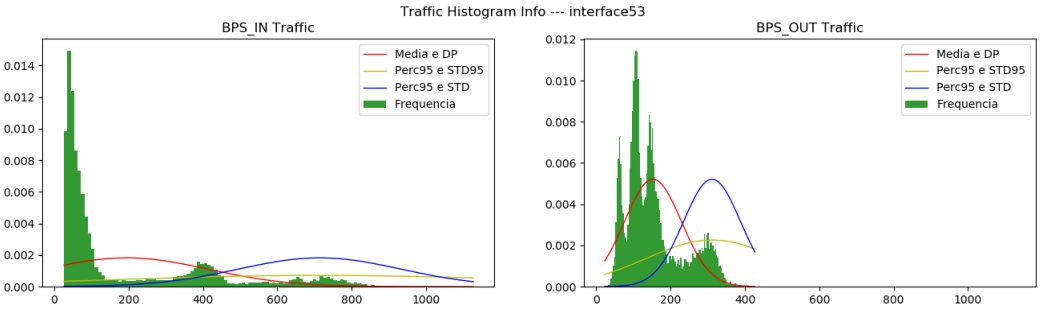

Figura 27 – Histograma de tráfego - Interface 53

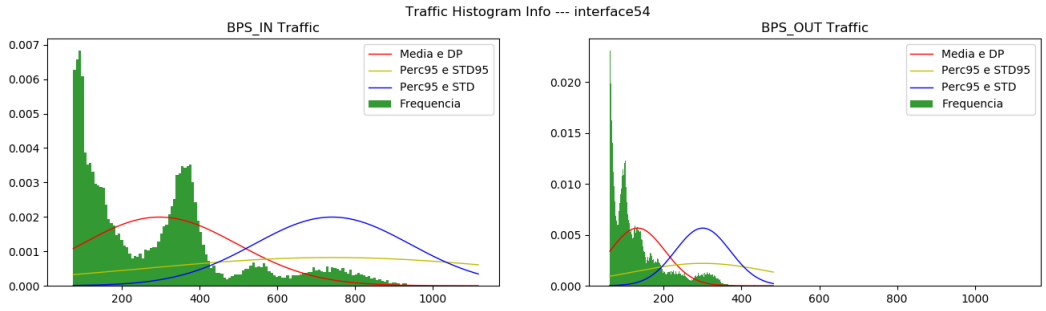

Figura 28 – Histograma de tráfego - Interface 54

Ao observar as distribuições de frequência da rede 5 e suas interfaces, notam-se, com exceção da interface 51, dois picos de frequência de consumo que podem ser aproximados de uma curva gaussiana. Para a distribuição da rede este comportamento aproximando da curva normal fica mais nítido em seu segundo pico.

Analisando as curvas normais em torno da distribuição nota-se que o comportamento das curvas vermelha e azul se aproximam dos picos de frequência, porém, a proposta apresentada de se utilizar o desvio padrão relativo ao percentil apresentou uma curva que destoa significativamente da frequência resultante nos histogramas. As observações não estão centralizadas, o que sugere que pode-se aperfeiçoar o cálculo do percentil.

5.2.2 Gráficos de Frequência da Rede 7 e análise dos dados

As características da rede 7 estão descritas na seção 5.1.2.

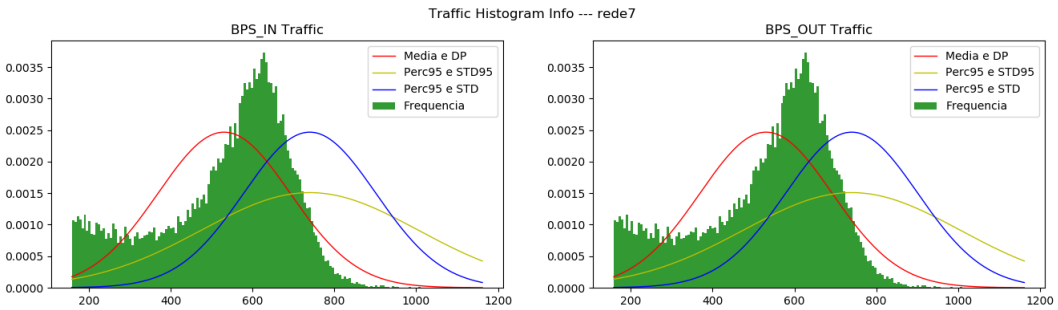

Figura 29 – Histograma de tráfego - Rede 7

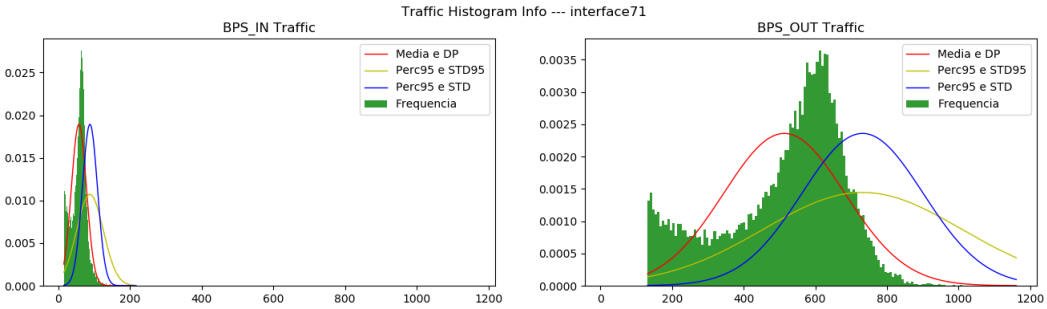

Figura 30 – Histograma de tráfego - Interface 71

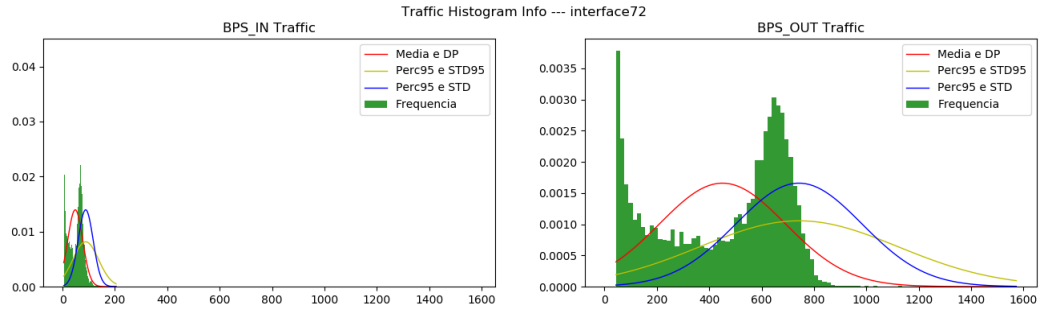

Figura 31 – Histograma de tráfego - Interface 72

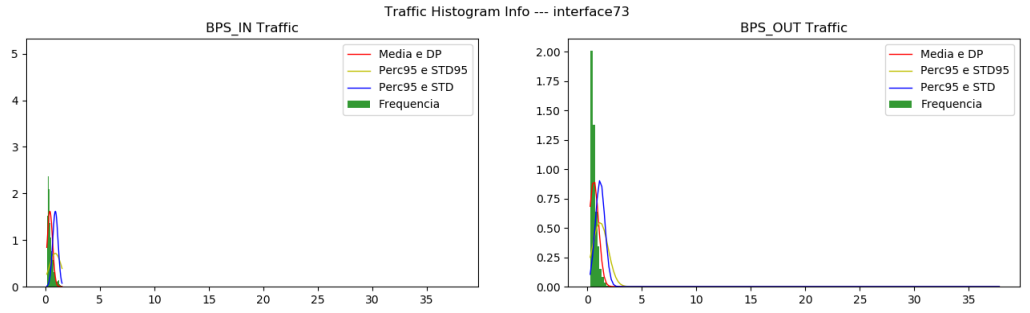

Figura 32 – Histograma de tráfego - Interface 73

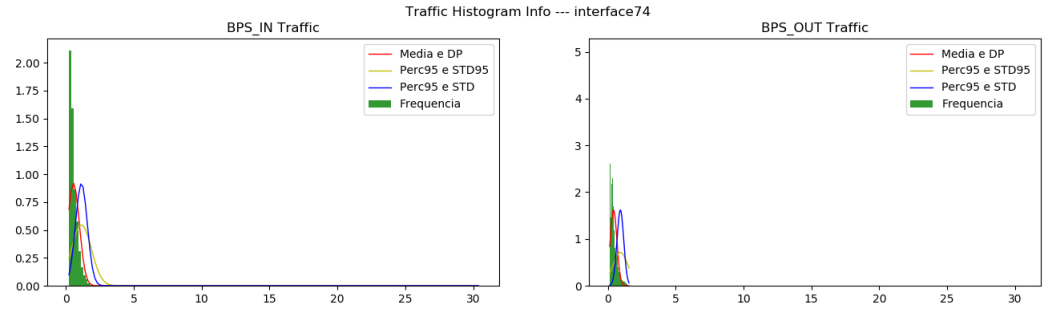

Figura 33 – Histograma de tráfego - Interface 74

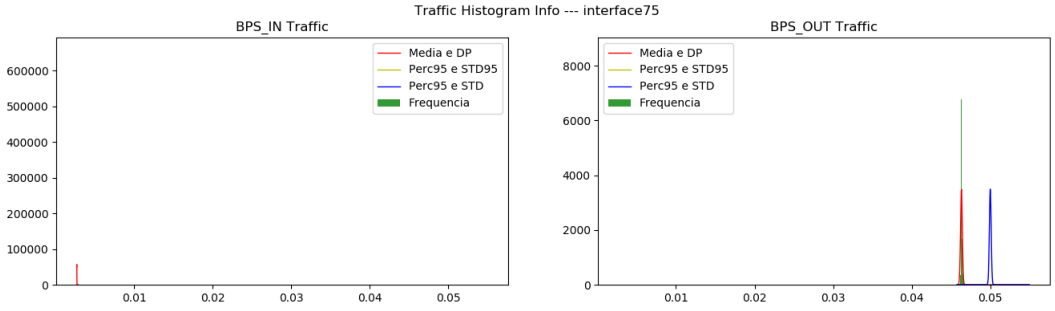

Figura 34 – Histograma de tráfego - Interface 75

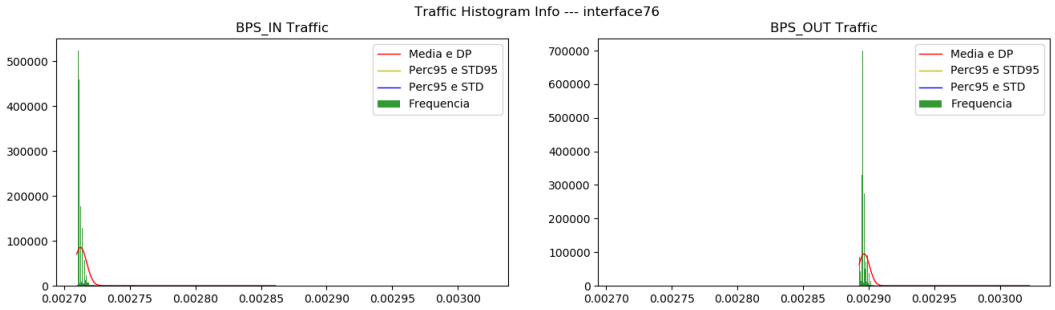

Figura 35 – Histograma de tráfego - Interface 76

Para a rede 7, a distribuição normal das amostras é muito mais visível, neste sentido, à medida que o volume de dados aumenta, esta curva tende a se tornar cada vez mais próxima à gaussiana. Contudo, mesmo com o comportamento mais tendente a uma curva normal, nota-se que o desvio padrão relativo não acompanha o comportamento da amostra, e, portanto, não é uma métrica adequada para utilização.

Assim como na rede 5, nota-se que o centro da curvas normais em vermelho e azul, que demonstram um comportamento próximo ao de uma distribuição de frequências, estão centralizadas em pontos diferentes do centro do histograma.

5.2.3 Análise geral do desvio Padrão Relativo

Ao utilizar o desvio padrão relativo proposto em 4.1 nota-se o crescimento esperado do desvio padrão, e, assim, a distribuição normal tende a se estender os limites inferiores e superiores da amostra, porém, imediatamente nota-se que este comportamento mais distribuído não acompanha a distribuição de frequências de amostras observados através do histograma. Ao contrário da proposta, nota-se que a redução do desvio padrão apresenta um comportamento mais provável para a curva normal das amostras.

Ao buscar os resultados para a proposta 4.1, foi possível observar a curva sob a ótica do seu centro, ou seja, a média utilizada para aproximar a gaussiana da distribuição de frequências. Apresentado nas redes 5 e 7, a distância entre o centro das curvas e o centro do histograma é observado também nas outras redes.

Assim, mesmo apoiado na teoria estatística de que qualquer distribuição tende à distribuição Normal quando o número de amostras tende ao infinito, observa-se que é necessário um ajuste ou correção no cálculo do desvio padrão e da distribuição aplicada, a fim de melhorar a precisão dos resultados obtidos com o Teorema Central do Limite, evitando o mal dimensionamento do ambiente.

5.3 Resultados do cálculo de Risco como probabilidade condicional

Ao se propor o risco como uma probabilidade condicional, em 4.2, determina-se uma área de consumo de banda que pode ser caracterizada como área de risco para o ambiente. Desta forma, para que possa ser observado o resultado desta proposta esta seção foi dividida em três partes, para que possam em conjunto justificar utilizar o risco como uma probabilidade condicional e apresentar os resultados desta proposta.

Visto que, em 5.1 foi demonstrado a melhoria na observação dos resultados, a partir desta seção serão observados somente as redes como serviço e não as interfaces individualizadas.

Para esta seção, foram utilizadas as redes 3, 4 e 8 para representar os resultados. As redes 3 e 4 são caracterizadas pelo atendimento de clientes externos, ou seja, existe uma maior aleatoriedade em seu consumo visto que não se conhece os padrões que estes vão consumir os recursos oferecidos. Por este motivo nota-se uma maior diferença entre o consumo médio e os picos de utilização que ocasionam um aumento do desvio padrão observado. Para a rede 8, a sua principal característica é agregar vários serviços que não possuem uma rede específica, assim, o tráfego pode ser mais volumoso que algumas aplicações, porém, não são serviços críticos e por isso, menos utilizados de forma geral, apresentam uma variação de consumo menor.

5.3.1 Resultados Área segura

Ao definir uma área como segura significa que toda amostras coletadas contidas neste intervalo não apresentam possibilidades de causar esgotamento de recurso. Esta área de segurança é estabelecida com o percentil escolhido para sumarização dos dados coletados.

Visto que toda a metodologia para avaliação da capacidade está baseada no consumo médio estabelecido, qualquer valor de consumo que seja inferior ou igual a esta grandeza, necessariamente está contido num espaço amostral já definido como base para se definir a capacidade suportada e, portanto, está contida em uma área segura de disponibilidade de recursos.

Assim, a área segura é definida como todo o intervalo contido na média escolhida, o percentil 95, e apresentado nos gráficos abaixo a sua representação.

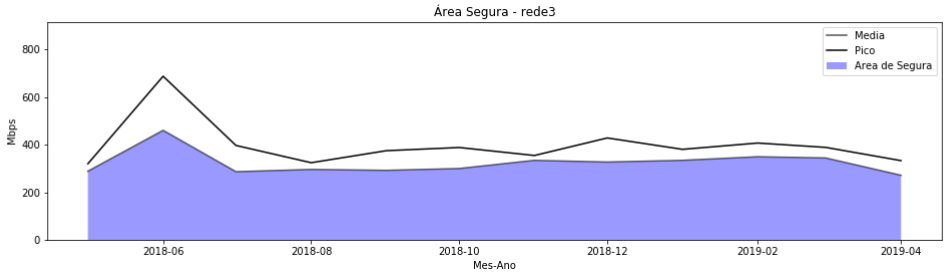

Figura 36 – Área segura - Rede 3

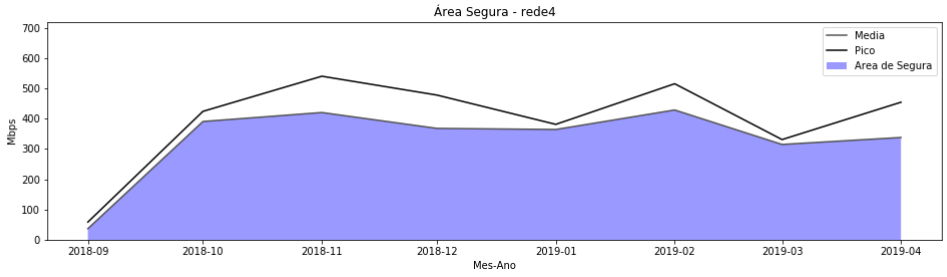

Figura 37 – Área segura - Rede 4

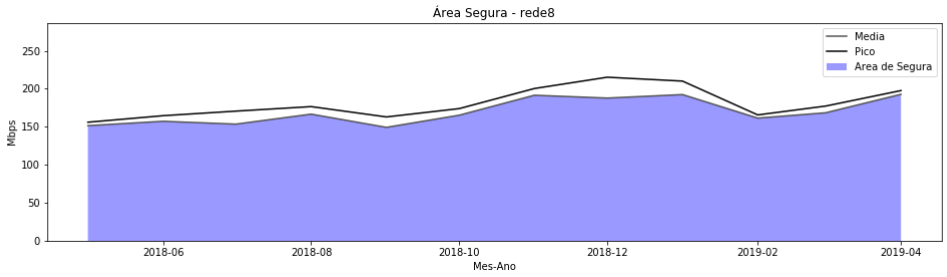

Figura 38 – Área segura - Rede 8

5.3.2 Resultados Área de risco

Ao contrário da observação em 5.3.1, a área segura é o espaço amostral que não foi estabelecido como média e que pode de alguma forma apresentar riscos durante as operações. Visto que o percentil 95 elimina de sua amostra eventos atípicos para melhor definir uma curva de tendência de consumo, estes picos extraídos da amostra são os valores cujo podem apresentar problemas para a aplicação e riscos reais à capacidade da infraestrutura preparada.

Qualquer valor atípico acima da média estabelecida, portanto, é um valor que pode causar risco, porém, baseado na própria coleta, nota-se que o pico de consumo é o maior valor obtido entre os dados registrados no mês, e assim, é o limite superior para o risco ocorrido no intervalo.

Então, pode-se restringir a área de risco como o intervalo entre o percentil 95 e o pico de consumo, representados pelos gráficos abaixo:

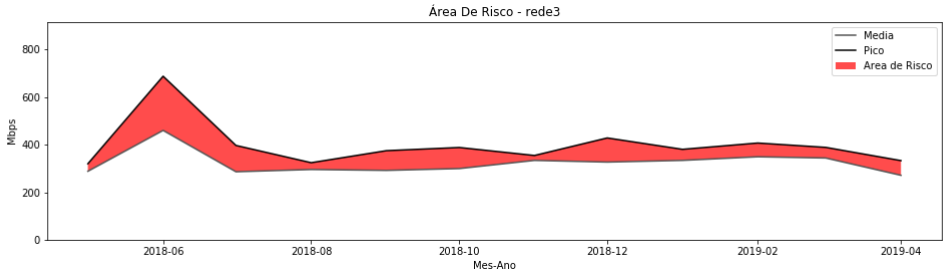

Figura 39 – Área de risco - Rede 3

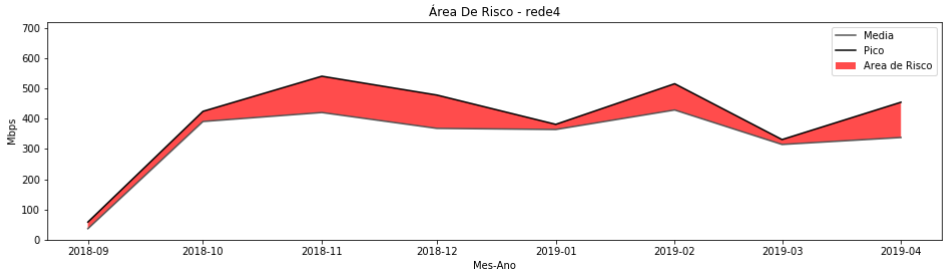

Figura 40 – Área de risco - Rede 4

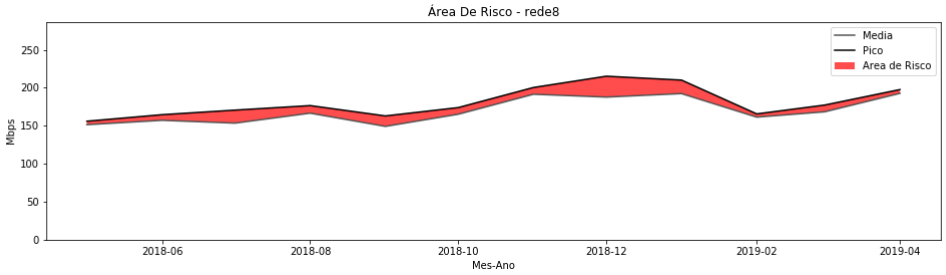

Figura 41 – Área de risco - Rede 8

5.3.3 Resultados de Risco

Por fim, justificadas as condições de segurança e risco, calcula-se o risco real apresentado para uma determinada capacidade existente da infraestrutura. A fim de simular o risco, o valor da capacidade de banda do serviço foi definido manualmente para que houvesse uma sobra de banda de 11,11%, ou seja, o percentil 95 corresponde à 90% da capacidade.

Com esta margem que excede o prospecção de consumo estipulada aplica-se o Teorema Central do Limite, obtendo os resultados de probabilidade de valores distribuídos normalmente que são maiores que a capacidade com e sem a condição proposta.

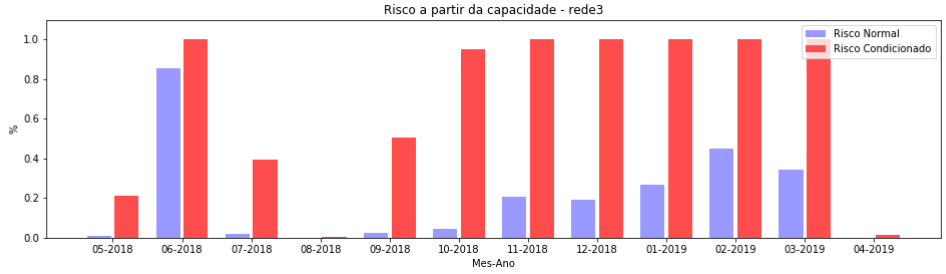

Figura 42 – Risco a partir da capacidade - Rede 3

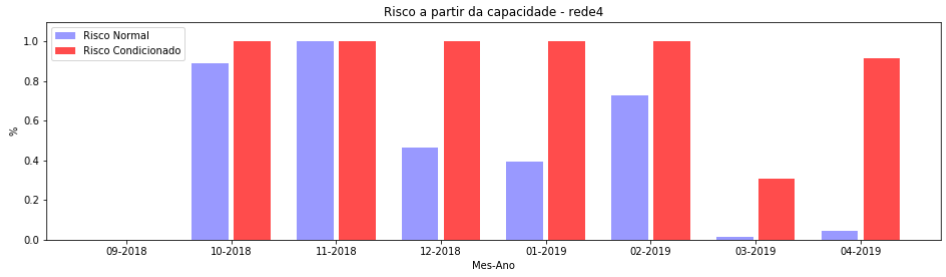

Figura 43 – Risco a partir da capacidade - Rede 4

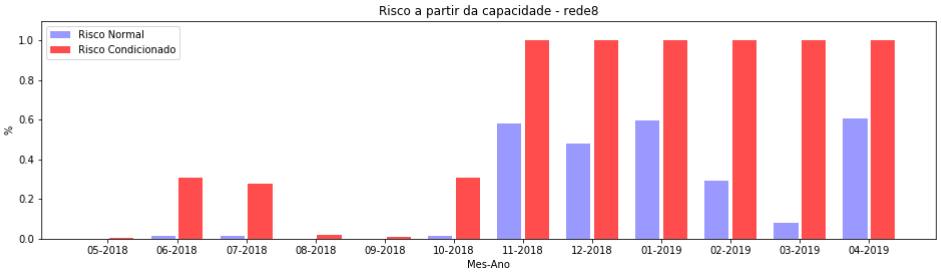

Figura 44 – Risco a partir da capacidade - Rede 8

A tabela 8 abaixo apresenta os dados dos gráficos no mês de março:

Tabela 8 – Gráficos no mês de Março

DATA	REDE	RISCO	RISCO CONDICIONADO
01/03/2019	Rede 3	0.3416	100.00
01/03/2019	Rede 4	0.0154	0.308
01/03/2019	Rede 8	0.08001	100.00

A tabela 9 abaixo apresenta os dados dos gráficos no mês de abril:

Tabela 9 – Gráficos no mês de Abril

DATA	REDE	RISCO	RISCO CONDICIONADO
01/04/2019	Rede 3	0.00088	0.0176
01/04/2019	Rede 4	0.04585	0.917
01/04/2019	Rede 8	0.60457	100.00

5.3.4 Análise Geral dos Resultados do Risco como probabilidade condicional

De fato, a preocupação com os dados dentro do parâmetro do *percentil* é mínima devido a resposta da capacidade já ser superior aos valores médios escolhidos e, quando realiza-se o cálculo do Teorema Central do Limite para este espaço amostral, obtém-se valores iguais ou próximos à zero. Ao impor a condição de risco condicional à essas mesmas medidas, a resposta deste cálculo é sempre zero. Alterando a condição para valores aleatórios abaixo da média pode-se adaptar a equação 4.4 como:

$$Pr(x > C|x < p95) = \frac{Pr(x < p95|x > C) \times Pr(x > C)}{Pr(x < p95)} \tag{5.1}$$

Onde, para uma situação de coleta de dados, todo valor de x maior que a capacidade, este sempre será maior que a média:

$$Pr(x < p95|x > C) = 0 \tag{5.2}$$

Utilizando a mesma analogia ao observar a área de risco, para a amostra coletada o maior valor possível é o pico, então, a probabilidade de um valor maior que o pico é nula. Assim, o limite superior da área de risco de uma amostra é sempre o pico de consumo observado.

A observação gráfico dos riscos colabora para a percepção da proposta 4.3, onde é visível o problema apresentado em fixar a capacidade baseado somente no consumo médio do período, pois, devido a esta designação não há uma mensura prévia do risco que se altera a cada mês sem estimações.

5.4 Resultados do cálculo de Capacidade baseado em risco

Introduzido na proposta estabelecida em 4.3, a estimação da capacidade da infraestrutura baseada em médias, embute ao seu ambiente a propriedade de suportar diversos riscos distintos ao longo dos períodos avaliados, pois, fixar a capacidade resulta em respostas diferentes devido aos diferentes desvios padrão das amostras.

Em contrapartida a essa flutuação de risco, fixá-lo permite obter faixas de capacidade com indicadores definidos, gerando mais controle do seu ambiente. Para simular a fixação desses indicadores de risco foram escolhidas três faixas baseadas no desvio padrão de Moore. De acordo com a regra de Moore, estabelecida em 2.1.5, ao somar o desvio padrão à média obtém-se um percentual de amostras normais contidas num intervalo. Contudo, definida a área de segurança na seção 5.3, pode-se definir o risco como o limite superior fora do intervalo da regra, ou seja, os valores superiores à média mais n vezes o desvio padrão, exatamente conforme a Figura 2 na seção 2.1.5. Assim, para a regras 68-95-99,7, tem-se três riscos definidor: Risco 16%, 2,5% e 0,15%. Com os riscos definidos, utiliza-se a equação eq. 4.3.9 para obter o valor de capacidade para cada mês.

Para esta seção, foram utilizadas as redes 1, 2 e 6 para representar os resultados. A rede 1 tem como principal característica a agregação das outras redes, por esse motivo apresenta um grande volume de dados e os maiores desvios padrão bruto, porém, proporcionalmente calculado, esta variância é muito próxima e dependente das variações dos outros serviços. A rede 2, é caracterizada por se tratar de uma rede interna, assim como a rede 7, portanto, padrões conhecidos e clientes bem determinados geram um desvio padrão muito pequeno. A rede 6 possui um padrão diferente das demais, apesar de ser uma rede para clientes internos, esta é uma rede caracterizada por utilização internacional, ou seja, conecta com outros países. A principal característica da rede 6 é que possui um consumo médio muito baixo, com situações de consumo pico entre 5 a 6 vezes maior que o consumo habitual, fugindo assim das tendências de curvas normais.

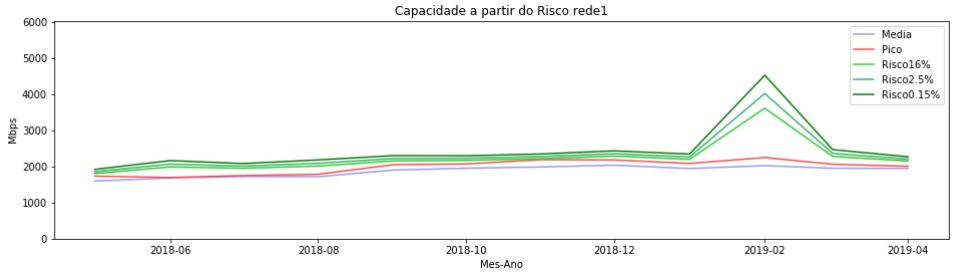

Figura 45 – Capacidade a partir do risco - Rede 1

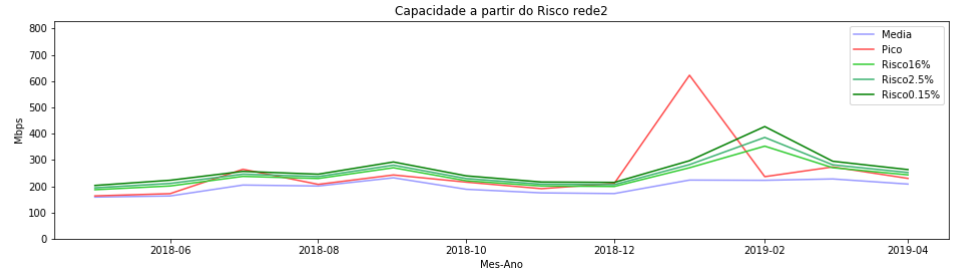

Figura 46 – Capacidade a partir do risco - Rede 2

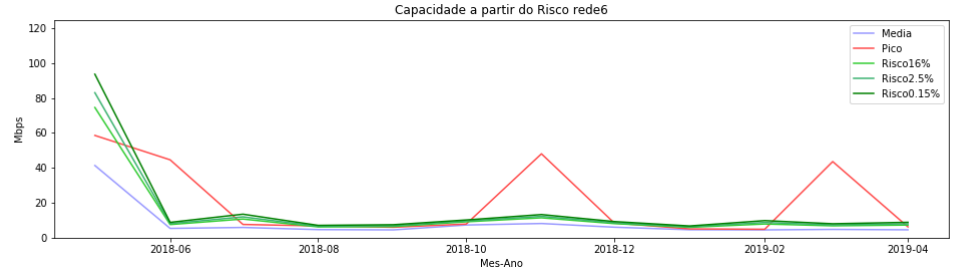

Figura 47 – Capacidade a partir do risco - Rede 6

A tabela 10 abaixo apresenta os dados dos gráficos no mês de março:

Tabela 10 – Capacidade a partir do Risco - mês de março

DATA	**REDE**	**RISCO 16%**	**RISCO 2,5%**	**RISCO 0,15%**
01/03/2019	Rede 1	2.284.155.227	2.367.766.303	2.472.238.731
01/03/2019	Rede2	270.688.251	281.594.311	295.221.483
01/03/2019	Rede 6	6.897.578	7.419.840	8.072.409

A tabela 11 abaixo apresenta os dados dos gráficos no mês de abril:

Tabela 11 – Capacidade a partir do Risco - mês de abril

DATA	REDE	RISCO 16%	RISCO 2,5%	RISCO 0,15%
01/04/2019	Rede 1	2.158.178.829	2.210.060.951	2.274.887.904
01/04/2019	Rede2	243.403.365	252.318.684	263.458.418
01/04/2019	Rede 6	7.342.075	8.036.376	8.903.908

5.4.1 Análise Geral dos Resultados do cálculo de capacidade baseado em risco

De modo geral, observa-se para a visão a nível de serviço uma variância pequena, e por esse motivo, as visões gráficas paras as três faixas de capacidade definidas estão muito próximas apesar da grande diferença do risco escolhido. Nota-se, neste sentido, que todos os momentos em que foi observado uma elevação do desvio padrão a diferença entre as faixas de capacidade aumentaram proporcionalmente para que o risco observado seja o mesmo ao longo do período.

Esta adaptação do ambiente limitado ao indicador de risco proporciona uma estimação que agrega mais valor apresentando o nível de segurança no qual o ambiente está configurado para suportar. Além disso, é possível observar que os cálculos realizados atribuíram à capacidade do ambiente valores que suportariam quase toda a área de risco.

Em contrapartida, nota-se que a forma de cálculo definida para o desvio padrão impactou diretamente no controle da estimação da capacidade, pois, a margem de segurança da capacidade é o produto entre o risco estabelecido e o desvio padrão da amostra. Assim, conforme registrado no resultado da análise do desvio padrão relativo, mesmo o desvio padrão da amostra pela média, afetou o cálculo do Teorema Central do Limite negativamente, sugerindo mais uma vez que a necessidade de correção ou ajuste da forma como foi calculado o desvio padrão.

Outro ponto a ser considerado é fruto da análise da rede 6, pois, apesar do Teorema Central do Limite se apresentar como um indicador bem estimado, esta medida não resolve os problemas de capacidade por si só, ou seja, ela agrega o valor necessário para uma avaliação. Devido ao comportamento de algumas amostras da rede 6, estas, independente do valor definido do risco, não são atendidas por uma definição de capacidade baseada em curvas gaussianas por se tratarem de amostras não distribuídas normalmente.

Então, o indicador é definido pelo risco a fim de oferecer suporte às tomadas de decisão para a construção ou manutenção da infraestrutura em relação à sua capacidade.

6 CONSIDERAÇÕES FINAIS

Considerando a importância que a capacidade da infraestrutura incide sobre a garantia de atendimento em todo o processo de serviços prestados, refletindo diretamente no custo efetivo final do ambiente, faz-se necessário a monitoração de recursos e o amadurecimento do processo de avaliação para que as previsões orçamentárias sejam definidas de maneiras mais precisas a fim de atingir um padrão aceitável de segurança e custo benefício.

Este trabalho vem, portanto, somar e apresentar uma nova visão do ambiente que pode auxiliar e amadurecer o estudo da avaliação de capacidade em relação ao seu custo operacional de contratos de acesso à internet. Desta forma, agrega-se valor à visão do analista responsável pelo gerenciamento da rede e, além de servir como indicativo para respostas de alarme, fundamenta a tomada de decisão durante a contratação de um circuito de rede.

De forma fundamentada, o estudo indica a melhoria ao realizar a agregação das informações que se referem a um mesmo serviço, observando a normalidade dos dados de coleta com volume elevado de informações e os benefícios de se observar e tomar decisões baseadas em condições pré-estabelecidas.

Apesar da dificuldade apresentada para se criar um modelo que se adeque e seja visível um casamento entre a coleta real e os gráficos calculados, devido à sua complexidade e o curto tempo hábil para o estudo, este trabalho é relevante para direcionamentos e prospecções futuras, onde pode-se apontar a correção dos indicadores calculados de percentil e desvio padrão como principal foco de direcionamento.

Com base no artigo "*Statistical Analysis and Modeling of the Internet Traffic*"do Janevski, Temkov e Tudzarov (2003) , que realiza o estudo de tráfego agregado de uma rede real, assim como neste trabalho, mas avaliando a auto similaridade do tráfego de internet usando as distribuições de Pareto e Exponencial, uma possível continuação deste estudo seria uma combinação de outras distribuições para corrigir e garantir a assertividade na avaliação da capacidade da rede. Neste sentido, a utilização do parâmetro de Hurst da distribuição de Pareto pode auxiliar na escolha da distribuição à ser combinada antes da aplicação do Teorema Central do Limite.

Este estudo demonstrou ser útil para a avaliações de capacidade utilizando o conceitos estatísticos e probabilísticos, e, portanto, se corretamente adaptado, é indiferente em relação à disciplina de aplicação contanto que atenda às necessidades expostas neste trabalho para aplicação correta do Teorema Central do Limite.

Referências

AZEVEDO, C. de. *Se as máquinas falassem: uma conversa franca sobre a gestão de ativos industriais*. [S.l.]: São Paulo: Saraiva, 2007. Citado 2 vezes nas páginas 9 e 26.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística: para cursos de engenharia e informática*. [S.l.]: Atlas São Paulo, 2004. v. 3. Citado 2 vezes nas páginas 20 e 21.

BLANCHARD, B. S.; FABRYCKY, W. J.; FABRYCKY, W. J. *Systems engineering and analysis*. [S.l.]: Prentice Hall Englewood Cliffs, NJ, 1990. v. 4. Citado na página 26.

BRUCE, P.; BRUCE, A. *Practical statistics for data scientists: 50 essential concepts*. [S.l.]: "O'Reilly Media, Inc.", 2017. ISBN 1491952938. Citado na página 35.

BUSSAB, W. d. O.; MORETTIN, P. A. *Estatística básica*. 9ª. ed. [S.l.]: Editora Saraiva, 2017. ISBN 8502207172. Citado 3 vezes nas páginas 21, 22 e 24.

GARTNER, I. R.; ZWICKER, R.; RÖDDER, W. Investimentos em tecnologia da informação e impactos na produtividade empresarial: uma análise empírica à luz do paradoxo da produtividade. *Revista de Administração Contemporânea*, SciELO Brasil, v. 13, n. 3, p. 391–409, 2009. ISSN 1982-7849. Citado na página 16.

JANEVSKI, T.; TEMKOV, D.; TUDZAROV, A. Statistical analysis and modeling of the internet traffic. In: . [S.l.: s.n.], 2003. Citado na página 62.

KELLY, A. *Strategic maintenance planning*. [S.l.]: Elsevier, 2006. v. 1. ISBN 0080478999. Citado na página 26.

LLOYD, C. et al. Asset Management: Whole-life management of physical assets. In: *Proceedings of the Institution of Civil Engineers-Municipal Engineer*. [S.l.: s.n.], 2010. v. 163, n. 4, p. 221–224. Citado 2 vezes nas páginas 17 e 26.

MAGALHÃES, M. N.; LIMA, A. C. P. Noções de Probabilidade e Estatística. 6ª edição revista. *São Paulo, EDUSP*, 2010. Citado na página 20.

MOORE, D. S. A estatıstica básica e sua prática. *LTC 7rd ed., Rio de Janeiro, Brazil*, 2017. Citado 7 vezes nas páginas 9, 22, 23, 24, 29, 34 e 35.

NUMPY. *NumPy*. 2019. Disponível em: <https://www.numpy.org/>. Citado na página 29.

OGBONNA, D. *AZ of Capacity Management: Practical Guide for Implementing Enterprise IT Monitoring & Capacity Planning*. [S.l.]: Booklocker. com, Incorporated, 2017. ISBN 1634927575. Citado 7 vezes nas páginas 11, 30, 31, 32, 33, 34 e 35.

ORACLE. *Welcome to cx_Oracle's documentation! — cx_Oracle 7.2.0-dev documentation*. 2019. Disponível em: <https://cx-oracle.readthedocs.io/en/latest/>. Citado na página 28.

PANDAS. *pandas - PyPI*. 2019. Disponível em: <https://pypi.org/project/pandas/>. Citado 2 vezes nas páginas 28 e 29.

SCHILDT, H.; MAYER, R. C. *C completo e total*. [S.l.: s.n.], 1997. ISBN 8534605955. Citado na página 28.

SINISUKA, N. I.; NUGRAHA, H. Life cycle cost analysis on the operation of power generation. *Journal of Quality in Maintenance engineering*, Emerald Group Publishing Limited, v. 19, n. 1, p. 5–24, 2013. ISSN 1355-2511. Citado na página 17.

VERNIMMEN, P. et al. *Corporate finance: theory and practice*. [S.l.]: John Wiley & Sons, 2014. ISBN 1118849337. Citado na página 17.

WOODWARD, D. G. Life cycle costing—theory, information acquisition and application. *International journal of project management*, Elsevier, v. 15, n. 6, p. 335–344, 1997. ISSN 0263-7863. Citado na página 26.

Anexos

ANEXO A – TABELA DISTRIBUIÇÃO NORMAL PADRÃO Z - N(0,1)

Tabela da Distribuição Normal Padrão
P(Z<z)

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Figura 48 – Tabela da distribuição Normal Padrão

ANEXO B – CÓDIGO PYTHON DE AGREGAÇÃO DE DADOS

(pfg-agg.py)

```
import pandas as pd
import numpy as np
import time
import oracleclass as oc
obj = oc.oracle()
mascara = ***
#FUNCOES SIMPLES
def p95(l,p=0.95):
try:
return round(sorted(l)[int(len(l)*p)-1],2)
except Exception as e:
return 0
def std95(l):
try:
mu,n = p95(l),len(l)
return round(np.sqrt(sum(list(map(lambda x: float((x-mu)**2)/(n-1), l)))),2)
except Exception as e:
return 0
def coluna_std(df,column='DT_CSM'):
df['DT_STD']=df[column].apply(lambda x: x.replace(day=1, second=0))
return df
def coluna_dia(df,column='DT_CSM'):
df['DT_DIA'] = df[column].apply(lambda x: x.replace(hour=0, minute=0, second=0))
return df
def coluna_mes(df,column='DT_CSM'):
df['DT'] = df[column].apply(lambda x: x.replace(day=1, hour=0, minute=0, second=0))
return df
#LEITURA DO ARQUIVO DE DADOS BRUTOS
def le_entrada():
arq = pd.read_excel('arq.xlsx',sep=';')
df = arq.copy()
```

```
df.BPS_IN.fillna(0,inplace=True)
df.BPS_OUT.fillna(0,inplace=True)
return df
#DICIONARIO DE SERVIÇOS
def srvc_dict():
s1 = [***]
s2 = [
***
]
s3 = [i+1 for i in range(len(s1))]
s4 = [ [j+1 for j in range(len(i))] for i in s2]

srvc_dict = dict(zip(s1,s2))
sdm = dict(zip(s3,s4))
d = {}
for k,v in srvc_dict.items():
rede = list(sdm.keys())[list(srvc_dict.keys()).index(k)]
d.update({k:"rede{}".format(rede)})
dict_itce=dict(zip(v,list(range(1,len(v)+1))))
for k2,v2 in dict_itce.items():
d.update({k2:"interface{}{}".format(rede,dict_itce[k2])})
return srvc_dict,d
srvc_dict,sdm = srvc_dict()
#CALCULO DO DESVIO PADRÃO DE FORMA SAZONAL
def desvio_padrao_sazonal(df,srvcs,datas_std,datas_mes):
df2=pd.DataFrame([])
desvios = lambda l: (round(np.std(l),2),round(std95(l),2))
for s in srvcs:
df_filtered = df.copy().query("CD_SRVC in [{}]".format(s)) if type(s)==int else df.
std_csm = [desvios(list(df_filtered.CSM_MBPS.loc[df_filtered.DT_STD==dt])) for dt i
df_temp = pd.DataFrame(list(zip(datas_std,[s]*len(datas_std),std_csm[0],std_csm[1]))
df_temp = coluna_mes(df_temp,'DT_STD')
std_csm = [p95(list(df_temp.STD_CSM.loc[df_temp.DT==dt])) for dt in datas_mes]
std95_csm = [p95(list(df_temp.STD95_CSM.loc[df_temp.DT==dt])) for dt in datas_mes]
df2 = df2.append(pd.DataFrame(list(zip([s]*len(datas_mes),datas_mes,std_csm,std95_c
return df2

#AGREGAÇÃO DOS DADOS COM PERCENTIL 95 E PICO
def p95_agg(df,datas,srvcs,dia=True):
```

```
df_temp=pd.DataFrame([])
agg = lambda l: (p95(l),p95(l,1))
for s in srvcs:
df_filtered = df.copy().query("CD_SRVC in [{}]".format(s)) if type(s)==int else df.
if dia:
df_filtered = coluna_dia(df_filtered)
csm,pico = zip(*[agg(df_filtered.CSM_MBPS.loc[df_filtered.DT_DIA==dt]) for dt in da
col = 'DT_DIA'
else:
df_filtered = coluna_mes(df_filtered)
csm,pico = zip(*[agg(df_filtered.CSM_MBPS.loc[df_filtered.DT==dt]) for dt in datas]
col = 'DT'
df_temp = df_temp.append(pd.DataFrame(list(zip(datas,[s]*len(datas),csm,pico)), col
return df_temp

#SELECT ORACLE
def entrada_oracle(dt):
select = """
**********************************************************************
""".format(str(dt[0]),str(dt[1]))
obj.select = select
return obj.dataframe.copy()

#AQUI SE DEFINE O TAMANHO DO ARQUIVO COM A DATA SELECIONADA
def entrada_datas():
select = """SELECT distinct(DT_CSM) as DT_CSM FROM TAB_CSM_REDE_MES WHERE
--cd_srvc IN (SELECT CD_SRVC FROM TAB_SRVC_REDE WHERE NM_SRVC='INTERNET') and
dt_csm>=to_date('01-05-2018','dd-mm-yyyy') and
dt_csm<to_date('01-06-2019','dd-mm-yyyy')
"""
obj.select = select
return list(obj.dataframe.copy().DT_CSM)

def script_pfg_agg(a = False):
t = time.time()
df_export = pd.DataFrame([])
datas = entrada_datas()
for i in range(1,len(datas)):
print("{}ª Iteracao\nParcial1: {}min{}s".format(i,int(time.time() - t)//60,int(time
```

```
df = le_entrada() if not(a) else entrada_oracle((datas[i-1],datas[i]))
if len(df.columns)<4:
df_export = df_export.append(df,ignore_index=True)
continue
cd_itces = list(df.CD_SRVC.unique())
datas_5min = list(df.DT_CSM.unique())
print("Parcial2: Leitura = {}min{}s".format(int(time.time() - t)//60,int(time.time(
for k,v in srvc_dict.items():
df_filtered = df.copy().query("CD_SRVC in {}".format(v))
if len(df_filtered)==0:
continue
csm_mbps = [max(sum(df_filtered.BPS_IN.loc[(df_filtered.DT_CSM==dt)]),sum(df_filter
df = df.append(pd.DataFrame(list(zip([k]*len(csm_mbps), datas_5min,csm_mbps,csm_mbp
srvcs = df.CD_SRVC.unique()
print("Parcial3: Servicos = {}min{}s".format(int(time.time() - t)//60,int(time.time
l = list(map(lambda a,b: max(a,b)/mascara,list(df.BPS_IN),list(df.BPS_OUT)))
df['CSM_MBPS'] = l
print("Parcial4: Consumo = {}min{}s".format(int(time.time() - t)//60,int(time.time(
df = coluna_std(df)
df = coluna_dia(df)
df = coluna_mes(df)
datas_std=list(df.DT_STD.unique())
datas_dia=list(df.DT_DIA.unique())
datas_mes=list(df.DT.unique())
print("Parcial5: Datas = {}min{}s".format(int(time.time() - t)//60,int(time.time()
try:
df_desvios = desvio_padrao_sazonal(df,srvcs,datas_std,datas_mes)
except:
print("Ocorreram erros")
pass
print("Parcial6: Desvio Padrão = {}min{}s".format(int(time.time() - t)//60,int(time
dia = p95_agg(df,datas_dia,srvcs)
mes = p95_agg(dia,datas_mes,srvcs,False)
df = pd.merge(mes,df_desvios, how='left', ...
...left_on=['DT_CSM','CD_SRVC'], right_on=['DT_STD','CD_SRVC']).drop(['DT_STD'],axi
df_export = df_export.append(df,ignore_index=True)
print("Parcial7: Agg, Merge, Export = {}min{}s".format(int(time.time() - t)//60,...
...int(time.time() - t)%60))
#df.to_excel("arq5min.xlsx")
```

```
print("Tempo Final: {}min{}s".format(int(time.time() - t)//60,int(time.time() - t)%
return
if __name__=='__main__':
try:
#df = script_pfg_agg() # sem entradas (executa a leitura do arquivo.xlsx uma unica
#df = script_pfg_agg(entrada_datas(),False) # entrada de ...
...data sem chamada oracle ( executa n-1 vezes o tamanho do array datas)
df = script_pfg_agg(True) # entrada de data com chamada oracle ...
...( executa para todas as datas no array datas)
df['MASCARA'] = [sdm[cd] for cd in list(df.CD_SRVC)]
df.to_excel("pfg_export_teste_de_mascara.xlsx")
except Exception as e:
print(e)
input("\nAperte ENTER para encerrar. \n>>> ")
```

ANEXO C – CÓDIGO QUE CALCULA RISCO E CAPACIDADES

(pfg_risco_condicional.py)

```
import pandas as pd
import numpy as np
from scipy.stats import norm
import time

def p95(l,p=0.95):
try:
return round(sorted(l)[int(len(l)*p)-1],2)
except Exception as e:
return 0

def risco_basico(mu,dp,cap):
mu,dp,cap = (float(mu),float(dp),float(cap))
if not(dp):
return dp
return float(format(1 - norm.cdf( (cap-mu) / dp),'.5f'))
def risco_condicional_basico(mu,dp,cap,percentil=0.95):
mu,dp,cap = float(mu),float(dp),float(cap)
r = float(format(risco_basico(mu,dp,cap)/(1-percentil),'.5f'))
return 0 if r<0 else (1 if r>1 else r)
def risco_condicional_completo(mu,dp,cap,pico,percentil=0.95,n=288*30):
mu,dp,cap = float(mu),float(dp),float(cap)
r = float(format((risco_basico(mu,dp,cap)-1/n)/((1-1/n) - percentil),'.5f'))
return 0 if r<0 else (1 if r>1 else r)

def tratamento_df(df):
df = df.replace(np.NaN,0.0)
df = df.replace(np.zeros,0.0)
return df

def riscos(df):
```

```
df_r = pd.DataFrame([])
srvcs = list(df.CD_SRVC.unique())
datas = list(df.DT_CSM.unique())
for s in srvcs:
df_filtered = df.copy().query("CD_SRVC in [{}]".format(s)) if type(s)==int ...
...else df.copy().query("CD_SRVC in ['{}']".format(s))
l = [[],[],[],[],[],[]]
c = np.mean(df_filtered.CSM_MBPS)/0.9  #UM VALOR SETADO MANUALMENTE
for dt in datas:
m = df_filtered.CSM_MBPS.loc[df_filtered.DT_CSM==dt]
d = df_filtered.STD_CSM.loc[df_filtered.DT_CSM==dt]
d95 = df_filtered.STD_CSM.loc[df_filtered.DT_CSM==dt]
p = df_filtered.PICO_MBPS.loc[df_filtered.DT_CSM==dt]
if not(len(m)):
continue
l[0]+=[risco_basico(m,d,c)]
l[2]+=[risco_condicional_basico(m,d,c)]
l[4]+=[risco_condicional_completo(m,d,c,p)]
#l[1]+=[risco_basico(m,d95,c)]
#l[3]+=[risco_condicional_basico(m,d95,c)]
#l[5]+=[risco_condicional_completo(m,d95,c,p)]

df_filtered['CAP90']=[c]*len(l[0])
df_filtered['RSC'] = l[0]
df_filtered['RSC_C'] = l[2]
df_filtered['RSC_CC'] = l[4]
#df_filtered['rsk_basico_dp2'] = l[1]
#df_filtered['rsk_cond_bas_dp2'] = l[3]
#df_filtered['rsk_cond_comp_dp2'] = l[5]
df_r = df_r.append(df_filtered,ignore_index=True)
return df_r

def capacity(df):
df_r = pd.DataFrame([])
srvcs = list(df.CD_SRVC.unique())
datas = list(df.DT_CSM.unique())
for s in srvcs:
df_filtered = df.copy().query("CD_SRVC in [{}]".format(s)) if type(s)==int ...
...else df.copy().query("CD_SRVC in ['{}']".format(s))
```

```
l = [[],[],[],[],[],[]]
rb = np.mean(df_filtered.RSC)
rbc = np.mean(df_filtered.RSC_C)
rcc = np.mean(df_filtered.RSC_CC)
for dt in datas:
m = df_filtered.CSM_MBPS.loc[df_filtered.DT_CSM==dt]
if not(len(m)):
continue
else:
m = float(m)
d = float(df_filtered.STD_CSM.loc[df_filtered.DT_CSM==dt])
p = float(df_filtered.PICO_MBPS.loc[df_filtered.DT_CSM==dt])
l[0]+=[(float(format((1*d + m),'.3f')))]
l[2]+=[(float(format((2*d + m),'.3f')))]
l[4]+=[(float(format((3*d + m),'.3f')))]
l[1]+=[(risco_basico(p,d,(1*d + m)))]
l[3]+=[(risco_basico(p,d,(2*d + m)))]
l[5]+=[(risco_basico(p,d,(3*d + m)))]
df_filtered['CAP1DP'] = l[0]
df_filtered['RSC_PICO_CAP1DP'] = l[1]
df_filtered['CAP2DP'] = l[2]
df_filtered['RSC_PICO_CAP2DP'] = l[3]
df_filtered['CAP3DP'] = l[4]
df_filtered['RSC_PICO_CAP3DP'] = l[5]
df_r = df_r.append(df_filtered,ignore_index=True)
return df_r

def risc_capacity(df):
df_r = pd.DataFrame([])
srvcs = list(df.CD_SRVC.unique())
datas = list(df.DT_CSM.unique())
for s in srvcs:
df_filtered = df.copy().query("CD_SRVC in [{}]".format(s)) if type(s)==int else ...
...df.copy().query("CD_SRVC in ['{}']".format(s))
c1 = [float(norm.ppf(1-(0.16/20))*df_filtered.STD_CSM.loc[df_filtered.DT_CSM==dt]..
... + df_filtered.CSM_MBPS.loc[df_filtered.DT_CSM==dt]) for dt in df_filtered.DT_CS
c2 = [float(norm.ppf(1-(0.025/20))*df_filtered.STD_CSM.loc[df_filtered.DT_CSM==dt].
... + df_filtered.CSM_MBPS.loc[df_filtered.DT_CSM==dt]) for dt in df_filtered.DT_CS
c3 = [float(norm.ppf(1-(0.0015/20))*df_filtered.STD_CSM.loc[df_filtered.DT_CSM==dt]
```

```
... + df_filtered.CSM_MBPS.loc[df_filtered.DT_CSM==dt]) for dt in df_filtered.DT_CS
df_filtered['CAP_PARA_RSC16%']=c1
df_filtered['CAP_PARA_RSC2.5%']=c2
df_filtered['CAP_PARA_RSC0.15%']=c3
df_r = df_r.append(df_filtered,ignore_index=True)
return df_r

def script_pfg_risco():
t = time.time()
df_export = pd.DataFrame([])
df = pd.read_excel("pfg_export.xlsx")
#df = pd.read_excel("pfg_export_medias.xlsx")
df = tratamento_df(df.copy())
srvcs = list(df.CD_SRVC.unique())
datas = list(df.DT_CSM.unique())
df = riscos(df)
df = capacity(df)
df = risc_capacity(df)
df.to_excel("pfg_export_2.xlsx")
return df.copy()
if __name__ == "__main__":
try:
df = script_pfg_risco()
except Exception as e:
print("Erro: {}".format(e))
input("\nAperte ENTER para encerrar. \n>>> ")
```