PREVISÃO DE TRÁFEGO EM ENLACES DE REDES UTILIZANDO SÉRIES TEMPORAIS

Evandro Luiz Cardoso Macedo

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Luís Felipe Magalhães de Moraes

Rio de Janeiro
Setembro de 2015

PREVISÃO DE TRÁFEGO EM ENLACES DE REDES UTILIZANDO SÉRIES TEMPORAIS

Evandro Luiz Cardoso Macedo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Luís Felipe Magalhães de Moraes, Ph.D.

Prof. Felipe Maia Galvão França, Ph.D.

Prof. Márcio Portes de Albuquerque, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2015

Macedo, Evandro Luiz Cardoso

Previsão de Tráfego em Enlaces de Redes Utilizando Séries Temporais/Evandro Luiz Cardoso Macedo. – Rio de Janeiro: UFRJ/COPPE, 2015.

XVII, 81 p.: il.; 29, 7cm.

Orientador: Luís Felipe Magalhães de Moraes

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2015.

Referências Bibliográficas: p. 75 – 81.

1. redes de computadores. 2. previsão. 3. séries temporais. 4. caracterização. 5. ARIMA. I. Moraes, Luís Felipe Magalhães de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Àqueles que contribuem
infinitamente para que eu cresça
em conhecimento e fé, meus pais
Duilio e Maria Bernadete.

Agradecimentos

Primeiramente, agradeço a Jeová Deus por sua infinita graça, pelo milagre da vida, por seu filho Jesus Cristo e por ser o Melhor Amigo que se pode ter, porque com Ele tudo é melhor, Ele não falha.

Obrigado pela minha família, meu pai Duilio Macedo, minha mãe Maria Bernadete e minha irmã Carla Macedo. Vocês são a melhor família que existe no mundo! Obrigado pelo carinho, apoio e amor. E também aos meus familiares que muito me apoiaram e acreditaram em mim.

Agradeço ao meu orientador Prof. Luís Felipe M. de Moraes, pelas ideias e orientações ao longo do curso, que muito contribuíram para meu crescimento, não se limitando somente ao âmbito acadêmico e profissional, mas também um crescimento pessoal para a vida. Obrigado também pela oportunidade de fazer parte do Laboratório Ravel e pela amizade.

Aos membros da banca, Prof. Felipe França e Prof. Márcio Portes, por aceitarem o convite e proporcionarem relevantes contribuições.

Ao Programa de Engenharia de Computação e Sistemas (PESC/COPPE/UFRJ) pela infraestrutura e suporte que permitiram o desenvolver deste trabalho. Em especial, agradeço aos professores pela excelência na qualidade do ensino, professores Edmundo Silva, Rosa Leão, Cláudio Amorim, José Rezende, Valmir Barbosa, Jayme Szwarcfiter e Aloysio Pedroza. E aos funcionários Gutyerrez da Costa, Cláudia Prata, Josefina Solange, Sônia Regina, Maria Mercedes, Itamar Marques, Adilson Barros, Irmão e Ricardo Cézar, pela prontidão em ajudar.

À equipe da RedeRio de Computadores/FAPERJ, professor Nilton Alves, Marita Maestrelli, Sandro Pereira, Jaime Fernandes e Marcelo Portes, por fornecer os dados para realização do trabalho, além de dicas e sugestões.

Às agências de fomento CAPES e FAPERJ pelo suporte financeiro.

Aos professores da UERJ que contribuirão indiscutivelmente para a construção das bases para o conhecimento que tenho hoje. Em especial, os professores Alexandre Sztajnberg, Paulo Eustáquio, Eduardo Galúcio, Rosa Costa, Leandro Marzulo, Maria Alice, Roseli Wedemann, Marcelo Schots e Cristina Waga.

Aos meus amigos de laboratório, Vander Proença, Renato Silva, Marco Coutinho, José Barbosa, Felipe Espósito e Cláudia Lima, pelo companheirismo, ideias

debatidas e sugestões. Obrigado pelas dicas e conversas na sala de “conivência” e nas confrarias.

Aos amigos da UFRJ e UERJ que muito me fazem agradecido por tê-los conhecido: Israel “Zinc”, Denilson Tavares, Diogo Souza (23), Diogo Silva (não 23), David Barreto, Moysés Sampaio, Brunno Goldstein, Bruno Brazil, Ingrid Boesing, Wellington Rodrigo, Natália Pedroza, Daniela Lübke, Danielle Castelo, Rebeca Motta, Renan Vicente, Renan Spencer, Joanna Manjarres, Gabriel Mendonça e Gustavo Santos.

E aos membros da TitanWings, Felipe Marx, Johnny Vice e Gabriel Ferreira, por contribuirem para meu crescimento e evolução. *PEACE*!

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

PREVISÃO DE TRÁFEGO EM ENLACES DE REDES UTILIZANDO SÉRIES TEMPORAIS

Evandro Luiz Cardoso Macedo

Setembro/2015

Orientador: Luís Felipe Magalhães de Moraes

Programa: Engenharia de Sistemas e Computação

O volume de tráfego da Internet apresenta uma tendência de crescimento incontestável e diversos estudos confirmam este fato. Para atender às demandas geradas pelo aumento do tráfego observado, faz-se necessário realizar atualizações dos recursos de rede de forma a manter índices de desempenho previamente estabelecidos dentro de limites aceitáveis. Se as necessidades de atualizações são identificadas antes que os recursos de rede se esgotem, então ganhos significativos podem ser alcançados em termos da qualidade do serviço prestado. Assim, propõe-se uma ferramenta de auxílio na previsão do tráfego, que seja capaz de capturar informações estatísticas deste, para inferir sobre condições futuras dos enlaces de rede. Este trabalho apresenta uma metodologia baseada em séries temporais, com uso do modelo matemático de previsão ARIMA e sua vertente sazonal SARIMA, bem como a caracterização do tráfego observado. O trabalho utiliza dados reais coletados a partir de enlaces da RedeRio de Computadores/FAPERJ e Redecomep-Rio. Os resultados obtidos através dos testes com diferentes granularidades (escalas de tempo) mostram que a abordagem utilizada é válida, apresentando a previsão para um período de seis meses e para um período de um mês, sobre o tráfego no enlace internacional da rede em estudo. O trabalho ressalta a importância da necessidade de manter um histórico das medidas de rede, para que previsões em janelas maiores de tempo possam ser realizadas com mais precisão.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

NETWORK TRAFFIC FORECAST USING TIME SERIES

Evandro Luiz Cardoso Macedo

September/2015

Advisor: Luís Felipe Magalhães de Moraes

Department: Systems Engineering and Computer Science

Internet traffic volume presents an undeniable growth tendency and lots of studies support this. In order to keep on attending the demands generated by increasing in the observed traffic, it makes necessary to update network resources so as to maintain performance levels previously defined within acceptable limits. If those resource update necessities are identified before their exhaustion, so significant benefits would be achieved in terms of quality of service. By and large, it is proposed a tool that would help forecasting network traffic, which would be able to capture statistics from it so as to infer about future conditions of particular links. Therefore, in this work is presented a methodology based on time series, utilizing the forecast model ARIMA and the seazonal version SARIMA, as well as a characterization of observed traffic. The study uses real data collected from both RedeRio de Computadores/FAPERJ and Redecomep-Rio links. The results obtained from tests with different granularities (time scales) indicate that the approach is viable, presenting forecast results for a period of six months and also for a period of one month of the traffic of international link within the network in study. The work highlights the importance for the need of keeping track of network measurements in order to have more accuretely forecasts in larger time windows.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

D_n	Valor da estatística Kolmogorov-Smirnov, p. 18
F_n	Uma função cumulativa de probabilidade, p. 18
H	Parâmetro de Hurst, p. 16
I	Intervalo de confiança, p. 10
L	Um enlace de rede, p. 9
N	Número de amostras, p. 10
N_i	Amostra observada, p. 18
Np_i	Média esperada, p. 18
S	Desvio padrão amostral, p. 10
S_X	Erro padrão, p. 10
V	Máxima verossimilhança, p. 19
X	Uma variável aleatória qualquer, p. 10
Y_t	Série temporal estacionária, com média zero, p. 16
Z	Distribuição Normal, p. 10
Γ_i	Coeficientes de Autorregressão Sazonal, p. 39
Θ_i	Coeficientes de Média Móvel Sazonal, p. 39
α	Nível de significância, p. 11
ϵ_t	Ruído aleatório, p. 37
γ_L	Tráfego em um enlace L, p. 9
γ_i	Coeficientes de Autorregressão, p. 37

$\hat{x}_i$	Corresponde a um valor estimado, p. 10
λ_{ij}	Fluxo de interesse de uma origem i para um destino j, p. 9
μ_e	Valor estimado de média, p. 11
μ_h	Valor de média sob hipótese, p. 11
$\overline{x}$	Média amostral da variável aleatória X, p. 10
θ_i	Coeficientes de Média Móvel, p. 37
d_n	N-ésima operação de diferença, p. 13
k	Número de parâmetros de um modelo, p. 19
m	Quantidade de séries, p. 16
p_L	Probabilidade deste fluxo passar pelo enlace L, p. 9
x_i	Corresponde a um valor real, p. 10

Lista de Siglas

AIC	*Akaike Information Criterion* – Critério de Informação de Akaike, p. 19
ANN	*Artificial Neural Network*, p. 22
ARIMA	*Autoregressive Integrated Moving Average* – Modelo de Média Móvel Integrada Autorregressivo, p. 3
CDF	*Cumulative Distribution Function* – Função de Distribuição de Probabilidade Cumulativa, p. 32
ECDF	*Empirical Cumulative Distribution Function* – Função de Distribuição de Probabilidade Cumulativa Empírica, p. 32
EQM	Erro Quadrático Médio, p. 21
HMM	Hora de Maior Movimento, p. 31
IP	*Internet Protocol*, p. 9
MAPE	*Mean Average Percentage Error* – Porcentagem do Erro Médio Absoluto, p. 11
MIB	*Management Information Base*, p. 28
MRTG	*Multi Router Traffic Grapher*, p. 28
PDF	*Probability Density Function* – Função de Densidade de Probabilidade, p. 32
PoP	*Point of Presence* – Ponto de Presença, p. 22
RFC	*Request for Comments*, p. 9
SARIMA	*Seazonal Autoregressive Integrated Moving Average* – Modelo de Média Móvel Integrada Autorregressivo Sazonal, p. 3
SDN	*Software-Defined Networking*, p. 70

SNMP *Simple Network Management Protocol*, p. 22, 28

TCP *Transmission Control Protocol*, p. 3

Capítulo 1

Introdução

Através da inclusão digital, o número de pessoas com acesso à Internet segue uma tendência de crescimento, que pode ser observada em pesquisas recentes [1]. Segundo o Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br) [1], em 2013, quase metade das residências no Brasil (49%) possuía ao menos um computador. Já a proporção de domicílios com acesso à Internet cresceu de 24% em 2009 para 43% em 2013. Por conta do aumento da acessibilidade a diversas aplicações emergentes (transmissão de vídeo em alta definição, computação na nuvem e redes sociais, por exemplo), observa-se um crescimento também no volume de tráfego nas redes [2]. Consequentemente, isso resulta em uma maior utilização dos recursos de comunicação, tais como enlaces e roteadores.

Um estudo feito pela *Cisco Systems* [2] aborda a previsão do tráfego global em redes IP para o período de 2012 a 2017. É estimado que em 2017 o tráfego IP global anual alcance os 1,447 *zettabytes* (1,447 ZB), o que significa 120,6 *exabytes*[1] por mês, sendo o patamar de 1 ZB já alcançado em 2015. As redes IP em escala global transmitirão 13,8 *petabytes* a cada 5 minutos, o que seria o equivalente em *gigabytes* a todos os filmes feitos até o ano de 2012 sendo gerados a cada 3 minutos. O estudo não utiliza dados reais coletados, mas os estima com base em outros fatores, tais como o número de usuários de uma determinada classe de aplicação, a adoção e a taxa de *bits* estimada para a aplicação, dentre outros.

O estudo também aponta um crescimento significativo em relação ao tráfego gerado por equipamentos móveis (smartphones, tablets e notebooks) devido às aplicações que essas plataformas permitem, tais como vídeos sob demanda, jogos *on-line*, serviços de mensagens, redes sociais, entre outros. Esses e outros fatores consequentemente contribuem para o aumento no tráfego de dados nas redes.

Como pode ser visto, o aumento na demanda por recursos de rede é uma tendência claramente observada. Com isso, à medida que os recursos vão sendo

[1] 1 zettabyte = 1000 exabytes; 1 exabyte = 1000 petabytes; 1 petabyte = 1000 terabytes

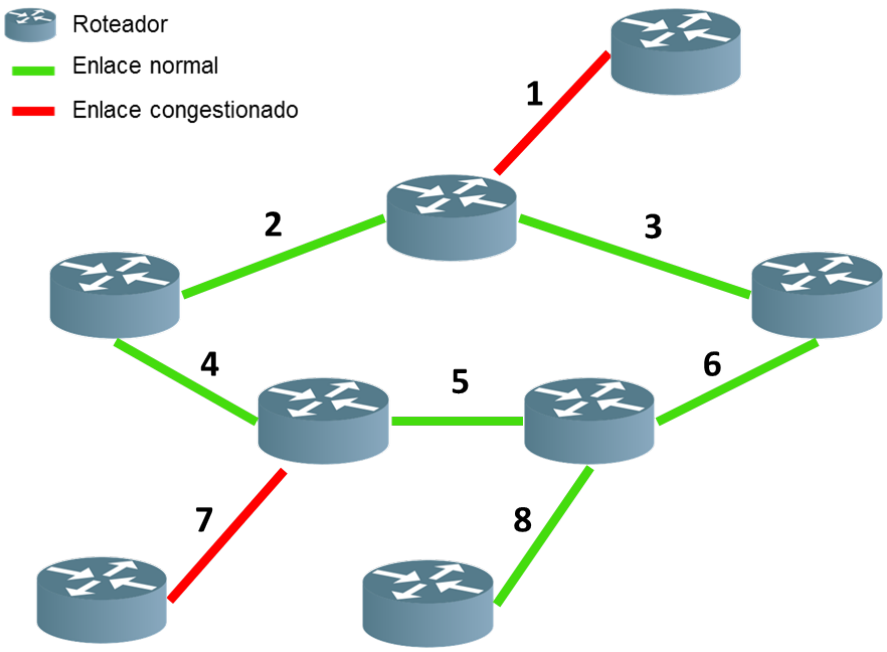


Figura 1.1: Exemplo de Rede de um Provedor de Acesso

consumidos, é necessário que sejam feitas atualizações na infraestrutura de rede para suportar tal demanda. Contudo, a fim de manter a qualidade do serviço e o bom funcionamento da rede, é preciso que as atualizações sejam realizadas antes que a demanda alcance níveis próximos à capacidade dos recursos. Porém, é difícil afirmar o momento em que um determinado recurso não será mais capaz de atender a esta demanda. Em outras palavras, saber com antecedência até quando haverá disponibilidade de largura de banda, de acordo com a demanda, torna-se um desafio. Sendo assim, um bom planejamento de infraestrutura de rede é de suma importância para o funcionamento adequado da rede.

Como um exemplo prático, considere uma rede de um provedor de acesso (um *backbone*) formada pelos roteadores e os enlaces de conexão entre eles, como ilustrado na Figura 1.1. Nesta rede, suponha que um administrador tenha interesse em acompanhar a utilização de um ou mais enlaces específicos, que são críticos para o desempenho da rede por serem enlaces de saída para a Internet, por exemplo, conforme ilustrado na Figura 1.1 pelos enlaces 1 e 7. Se esses enlaces apresentarem congestionamento, todos os usuários que tiverem uma conexão que passe por esses enlaces, terão a qualidade do serviço prejudicada, ocasionada pela lentidão na rede. Assim, busca-se prevenir que estas condições de congestionamento se configurem na rede.

Conforme mencionado anteriormente, coloca-se como desafio saber antecipadamente quando um recurso estará próximo do esgotamento, neste caso a capacidade de transmissão do enlace.

Outro aspecto que pode ser destacado é em relação ao intervalo de tempo entre uma solicitação de um novo recurso e a efetiva implantação do mesmo. No cenário real de uma rede, em face a uma demanda, é comum que ocorram atualizações de recursos, como o aumento da velocidade de transmissão em um enlace, por exemplo. Contudo, o recurso solicitado geralmente não é fornecido no momento da requisição, por questões de logística ou burocracia. Por conta disso, a demanda que por hora era aceitável, passa a ser excessiva, afetando na qualidade do serviço prestado. Desta forma, saber com antecedência o estado em que se encontra a utilização de um recurso, contribui para que a qualidade do serviço não seja impactada. Isso porque, a solicitação poderá ser feita em tempo hábil para que o recurso esteja disponível.

Sendo assim, colocam-se como objetivos desafiadores a previsão do tráfego em enlaces de rede, o planejamento da capacidade dos recursos de forma antecipada e a caracterização dos dados.

Esses objetivos são motivadores para o estudo e construção de uma ferramenta que realize a previsão da demanda de tráfego em um horizonte de previsão específico. Para isso, é necessário que seja definida uma metodologia a ser seguida, a fim de guiar a construção dessa ferramenta.

O presente trabalho apresenta a utilização de uma metodologia de previsão de tráfego em uma avaliação, utilizando o modelo matemático de séries temporais ARIMA (*Autoregressive Integrated Moving Average* – Modelo de Média Móvel Integrada Autorregressivo) e sua variante sazonal SARIMA (*Seazonal* ARIMA).

Também é apresentada a caracterização do tráfego em enlaces de um provedor de serviço. O trabalho mostra a importância da disponibilidade de dados históricos de medições de rede, de forma a prover confiabilidade ao processo de análise, assim como os cuidados na etapa de filtragem e coleta dos dados em relação à granularidade destes.

1.1 O Problema Abordado

Diversos fatores podem influenciar negativamente no desempenho geral de uma rede, como ruído e perdas no canal de transmissão, má configuração de equipamentos, ou ainda uma superutilização de um determinado enlace. Em outras palavras, quando a utilização do enlace chega próximo à capacidade de transmissão deste, configura-se uma condição de congestionamento.

Apesar de protocolos de rede apresentarem mecanismos para controlar condições de congestionamento, como é o caso do TCP (*Transmission Control Protocol*) [3], o enlace sempre tem seu limite de utilização físico. O controle que o protocolo TCP realiza é fim-a-fim, o que significa que não há uma visão de quais enlaces fazem parte do caminho entre a origem e o destino por parte do protocolo, não sendo possível

de serem gerenciados pelo TCP, mas somente pelo gestor do enlace, que detém o monitoramento deste.

Uma metodologia geralmente aplicada para redes tradicionais de comunicação (telefonia, por exemplo) para saber a utilização de um determinado enlace, é a aplicação de uma matriz de tráfego [4, 5]. Juntamente com o modelo de geração de tráfego, que caracteriza as requisições dos clientes, é feito o planejamento de capacidade utilizando uma distribuição de probabilidade conhecida na literatura [6].

Através da matriz de tráfego é possível saber exatamente o quanto de "carga" está sendo levado por um enlace entre uma origem e um destino, permitindo realizar simulações baseadas em cenários reais.

Contudo, com o advento da Internet, o número de requisições nas redes de computadores e o volume de dados que é transitado é de uma escala muito maior do que nas redes tradicionais de telefonia (vide as tendências em [2]). Sendo assim, poucas técnicas e trabalhos relacionados têm sido explorados, como pode ser visto em [7], no que tange a estimação da matriz de tráfego de uma rede nas proporções de um *backbone* de um provedor. Isto é devido à complexidade de construir uma matriz que represente o volume de tráfego entre cada origem e destino possível em uma rede de dimensão mundial.

Outro tipo de abordagem utilizada é a de séries temporais, a fim de acompanhar o histórico de utilização de um enlace, na qual a coleta de dados é realizada ao longo do tempo, tendo assim amostras de tráfego alinhadas na forma de uma série temporal. Com isso, uma variedade de estudos pode ser aplicada a essa série com o intuito de compreender o comportamento futuro do tráfego.

Alguns trabalhos importantes em relação a esse problema apresentam métodos para análise de séries temporais, nos quais é possível encontrar diversas abordagens para serem aplicadas na previsão. Na Seção 2.5 são mostrados os trabalhos relacionados a esse tipo de problema.

Sendo assim, o problema tratado nesta dissertação compreende a previsão de tráfego de rede em um determinado enlace, a fim de saber com um período de antecedência predefinido, qual a quantidade de tráfego (dado em *bytes/s* ou *bits/s*) que passará no enlace em questão, usando uma abordagem de séries temporais. Com isso, é possível fazer o provisionamento de recursos de rede antes que estes se esgotem.

Uma definição mais formal para este problema pode ser apresentada utilizando a perspectiva de variáveis aleatórias [10]. Considere X uma variável aleatória que representa a quantidade de *bytes* que passam por segundo em um determinado enlace (o tráfego de um enlace). Com essa formulação, entende-se uma série temporal como um processo estocástico sobre a variável X, ou seja, uma família de variáveis

aleatórias X indexadas no tempo – $X(t)$. Mais da definição formal para séries temporais pode ser encontrado na Seção 2.3.

O estudo de séries temporais é bem estabelecido e já foi aplicado em diferentes escopos de problemas, inclusive na compreensão do tráfego em Redes de Computadores, como pode ser visto nos trabalhos [11–14].

1.2 Proposta

Tendo em vista o problema apresentado anteriormente, coloca-se como proposta para o presente trabalho a descrição de uma metodologia de previsão de tráfego, utilizando a abordagem de séries temporais, que sirva como ferramenta de auxílio na tomada de decisões em relação ao provisionamento de recursos de rede. Com isso, objetiva-se antecipar a atualização na capacidade de enlaces de rede, a fim de evitar condições de esgotamento.

Além disso, o estudo apresenta resultados em relação à caracterização dos dados coletados do roteador de borda da RedeRio de Computadores/FAPERJ, utilizada como estudo de caso. Tais dados foram fundamentais para o desenvolvimento deste trabalho, permitindo que padrões na utilização dos recursos de rede fossem identificados, possibilitando um planejamento mais acurado e antecipado dos recursos necessários.

Sendo assim, a pesquisa tem por objetivo:

- Descrever a metodologia de previsão de tráfego utilizada;
- Coletar dados através de uma estrutura de monitoramento, com o intuito de utilizar dados reais para avaliação do estudo;
- Gerar as Funções de Distribuição Empíricas do tráfego observado;
- Realizar o Ajuste de Distribuição para caracterização do tráfego;
- Realizar a Previsão de Tráfego utilizando um modelo de análise de séries temporais, a saber, o modelo ARIMA e sua vertente sazonal SARIMA.

1.3 Estudo de Caso

Como estudo de caso, utilizou-se a RedeRio de Computadores/FAPERJ (RedeRio), tendo esta fornecido os dados que foram primordiais para evolução do trabalho.

Inaugurada em 1992, a RedeRio é um dos principais instrumentos de desenvolvimento científico do Estado do Rio de Janeiro, interconectando os mais avançados

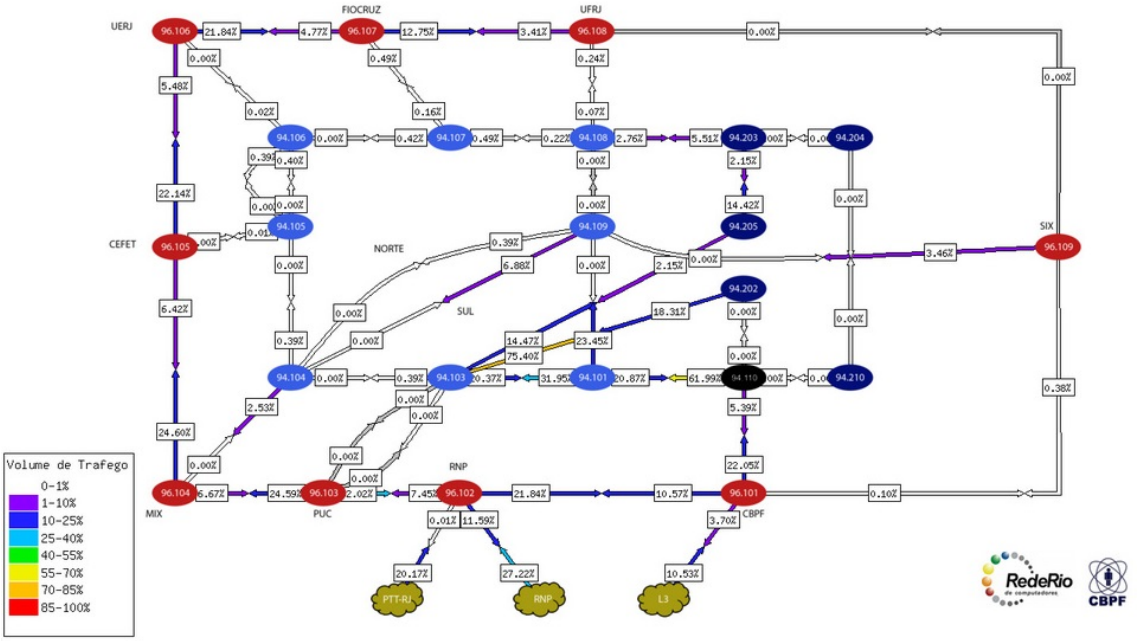


Figura 1.2: Mapa de rede da RedeRio de Computadores/FAPERJ e Redecomep-Rio [15]

centros de pesquisa do país, sediados nas universidades e nas empresas públicas e privadas do Estado [15].

O estudo de caso também envolve a infraestrutura da Redecomep-Rio [16], que é um braço do projeto de Redes Comunitárias de Educação e Pesquisa (Redecomep). O objetivo principal das Redecomep é promover a implantação de redes metropolitanas comunitárias nas capitais dos estados brasileiros, atendendo principalmente às instituições de ensino, pesquisa, ciência e tecnologia.

O Mapa da topologia da RedeRio e Redecompe-Rio pode ser visto na Figura 1.2.

Nestas redes, um dos enlaces de maior importância é o acordado com a *Level 3 Communications* [17], que fornece acesso através de um canal internacional. Assim, este enlace será utilizado para estudo, conforme será visto no Capítulo 4.

1.4 Contribuições

As contribuições do trabalho podem ser destacadas pelo(a):

- Desenvolvimento de um interpretador automatizado para os dados coletados;
- Caracterização dos dados que possibilita a inferência de qual distribuição mais se assemelha ao tráfego de cada interface monitorada;
- Ajuste de distribuição com o objetivo de aproximar um modelo matemático de geração dos dados coletados;

- Apresentação da metodologia de previsão de tráfego aplicada ao contexto de um provedor utilizando dados reais;
- Discussões sobre como lidar com diferentes granularidades a fim de obter uma previsão aceitável;
- Os dados coletados utilizados para elaboração do estudo.

1.5 Organização do Texto

No Capítulo 2 alguns conceitos básicos são apresentados visando trazer definições fundamentais para o melhor entendimento no decorrer do trabalho. Além disso, diversos trabalhos relacionados são abordados para contextualização da dissertação no estado da arte.

No Capítulo 3 a metodologia utilizada é apresentada e detalhada em cada uma de suas etapas.

No Capítulo 4 são apresentados os resultados e discussões das análises realizadas sobre uma das interfaces da RedeRio de Computadores/FAPERJ, utilizada como estudo de caso, aplicando a metodologia apresentada.

O Capítulo 5 encerra com as conclusões e sugestões de trabalhos futuros.

Capítulo 2

Fundamentos e Estado da Arte

Uma vez que a proposta desta dissertação utiliza conceitos de diversas áreas, é fundamental que estes sejam explicados, já que serão de grande importância para nivelar o entendimento ao longo do trabalho. Sendo assim, este capítulo se propõe a tratar destes respectivos conceitos.

Também neste capítulo é feita uma revisão bibliográfica, trazendo diversos trabalhos relacionados, para contextualizar a dissertação no estado da arte.

2.1 Fluxo, Tráfego e Medições

Três conceitos fundamentais em Redes de Computadores são fluxo, tráfego e os tipos de medições que podem ser feitas nas redes. Essas definições ajudam a entender praticamente a maioria dos problemas encontrados em redes e são base também para a compreensão de outros conceitos.

2.1.1 Fluxo

Considera-se um fluxo em Redes de Computadores como o deslocamento de pacotes de dados entre um dispositivo de origem e um destino. Um fluxo é estabelecido através dos respectivos endereços IP[1] (*Internet Protocol*) dos dispositivos, a porta da aplicação e o protocolo de transporte utilizado. A conexão estabelecida entre os dispositivos não necessariamente precisa ser feita de forma direta, com o dispositivo de origem diretamente conectado por um meio físico ao dispositivo de destino. Essa conexão pode ser dada através de uma sequência de outros dispositivos de rede intermediários, que serão transparentes para esse fluxo.

Um exemplo que pode ser dado é no caso de um computador que estabeleça uma conexão com um servidor Web para abrir uma página de um *site*, criando assim

[1] Uma sequência de números que identifica um dispositivo em uma rede

um fluxo de dados entre estes dispositivos. Fazendo uma analogia, é como um tubo virtual interligando os dispositivos, sendo os dados "escoados"por esse tubo.

Algumas RFCs (*Request for Comments*) apresentam definições para fluxo, como a RFC 2722 [18] que define o fluxo como o equivalente a uma chamada ou uma conexão. A RFC 3917 [19] define fluxo como um conjunto de pacotes que passam por um ponto de observação da rede, dentro de um intervalo de tempo. Esta definição é próxima a que foi colocada no início desta seção e será está a utilizada ao longo do texto.

2.1.2 Tráfego

Em Redes de Computadores, o tráfego é compreendido como a quantidade de pacotes de dados transmitidos por unidade de tempo multiplicado pelo tempo de transmissão por pacote. O tráfego é formado pela agregação de um ou mais fluxos de rede e é geralmente quantificado em *bytes/s* ou *bits/s*. Ou seja,

$$R = \lambda \bar{s}$$

onde R é o tráfego, λ é a taxa de chegada de pacotes e $\bar{s}$ é o tempo de transmissão por pacote.

Uma forma de representar a quantidade de tráfego (γ_L) em um enlace L é dada pela seguinte fórmula:

$$\gamma_L = \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_{ij} p_L$$

sendo γ_L o tráfego em um enlace L qualquer, λ_{ij} o fluxo de interesse de uma origem i para um destino j e p_L a probabilidade deste fluxo passar pelo enlace L, para todas as N fontes de tráfego de uma rede.

2.1.3 Medições

Para realizar estudos utilizando o comportamento real de uma rede, é preciso que sejam feitas medições sobre as características da rede em estudo. Desta forma, para obter os dados de interesse, existem basicamente dois tipos de medição: ativa e passiva.

Medição Ativa

São utilizadas ferramentas de rede que enviam pacotes pela rede, a fim de coletar informações de acordo com a resposta que esses pacotes enviados geram em relação a uma métrica analisada. Um exemplo de medição ativa é o uso da ferramenta *ping*

[20], que insere pacotes na rede a fim de coletar o tempo de ida e volta entre uma origem e um destino. Esse tipo de medição geralmente é feito com cautela para que os pacotes inseridos na rede não gerem um tráfego que influencie no comportamento da rede de fato.

Medição Passiva

Neste tipo, um "escutador" (*probe*) realiza a captura dos pacotes de rede sem a necessidade de envio de novos pacotes, o que torna as medições livres de interferência de pacotes que não do próprio tráfego. A ferramenta NetFlow [21] embarcada em roteadores Cisco permite que o tráfego das interfaces de um elemento de rede seja exportado sem que o tráfego em si seja impactado pelo tráfego de monitoramento. Essa medição é mais utilizada quando se procura coletar informações em termos de fluxo.

2.2 Tipos de Erro e P-valores

Quando se realiza um estudo de previsão de valores é comum que tais previsões apresentem divergência em relação aos valores reais. Assim, as previsões apresentam erros.

Nesta seção é explicada a diferença entre erro absoluto e erro relativo. Também é introduzido o conceito de p-valor, que auxilia na decisão sobre uma hipótese levantada.

2.2.1 Erros e Resíduos

O erro é uma medida intuitiva que indica o quanto um valor estimado ($\hat{x}_i$) se difere de um valor real (x_i), como na Equação (2.1). Este é o chamado erro absoluto, ou ainda resíduo.

$$ErroAbsoluto = |x_i - \hat{x}_i| \tag{2.1}$$

O erro relativo é calculado por um percentual em relação ao erro absoluto, como na Equação (2.2):

$$ErroRelativo = \frac{|x_i - \hat{x}_i|}{x_i} \cdot 100 = \frac{ErroAbsoluto}{x_i} \cdot 100 \tag{2.2}$$

O desvio padrão é uma medida de dispersão dos valores amostrados em relação à média, enquanto que o erro padrão é uma medida de variabilidade da média. Ou seja, o erro padrão define o quanto a média pode variar, para mais ou para menos.

O erro padrão faz parte então do cálculo do intervalo de confiança, sendo calculado em função do tamanho da amostra [22]. O desvio padrão é dado pela Equação (2.3).

$$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}} \tag{2.3}$$

onde μ é a média da população.

Para o caso do desvio padrão amostral, tem-se a Equação (2.4). Vale notar que, para o desvio padrão amostral, o denominador é $N - 1$ quando se deseja que a variância amostral seja um estimador não tendencioso da variância populacional.

$$S = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{N - 1}} \quad \text{e} \quad \overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.4}$$

onde S é o desvio padrão amostral, X é uma variável aleatória qualquer e $\overline{x}$ é a média amostral de X.

O erro padrão amostral (S_X) é baseado no conceito de desvio padrão amostral e no número de amostras (N), como na Equação (2.4).

$$S_X = \frac{S}{\sqrt{N}} \tag{2.5}$$

Assim, para o cálculo do intervalo de confiança I considerando um nível de confiança de 95%, tem-se:

$$I = (\overline{x} - z_{0,025} \cdot S_X ; \overline{x} + z_{0,025} \cdot S_X)$$

onde $z_{0,025}$ é o valor correspondente na tabela da distribuição Normal Z para o nível de confiança de 95% [10], a saber, $z_{0,025} = 1,96$.

Uma das formas utilizadas para avaliar o desempenho da previsão realizada é com base no MAPE (*Mean Absolute Percentage Error* – Erro Percentual Médio Absoluto) [23], definido pela Equação 2.6.

$$MAPE = \frac{1}{N} \sum_{t=1}^{N} ErroRelativo \tag{2.6}$$

O modelo que apresenta o menor MAPE é o modelo considerado mais adequado para a previsão de uma determinada série temporal.

2.2.2 P-valor

Quando se está trabalhando com testes de hipótese, uma das características buscadas é saber com que probabilidade um dado valor estimado se distancia em uma

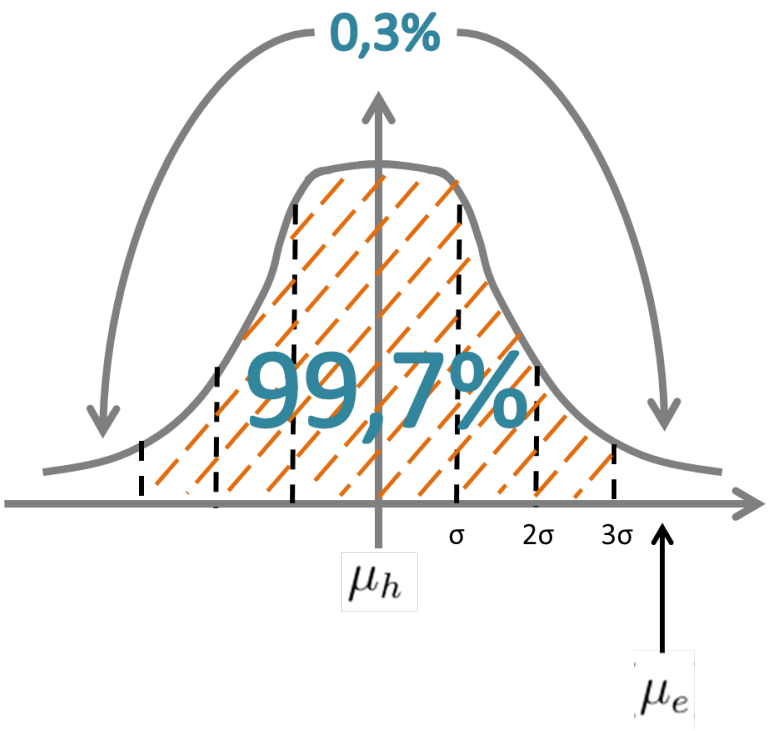


Figura 2.1: Exemplo para p-valor

determinada quantidade de desvios padrão do valor usado na hipótese. Se esta probabilidade for menor que o nível de significância (α), então isso sugere que a hipótese nula é falsa, podendo rejeitá-la e aceitar a hipótese alternativa. A essa probabilidade dá-se o nome de p-valor.

A Figura 2.1 ilustra o conceito de p-valor. No caso, a hipótese nula consiste em saber se um valor estimado de média (μ_e) pode ser entendido como uma aproximação do valor hipotético μ_h, com três desvios padrão de defasagem.

O valor obtido μ_e está a três desvios padrão distante da média μ_h. Qual a probabilidade desse caso ocorrer, dado que a hipótese nula é verdadeira, isto é, que $\mu_h = \mu_e$? Consultando a tabela de Z (distribuição Normal), a probabilidade de ter três desvios distante da média é de 0,003. Ou seja, o p-valor neste caso é 0,003, o que indica a rejeição da hipótese nula ($\mu_h = \mu_e$), sendo então a hipótese alternativa aceita ($\mu_h \neq \mu_e$).

2.3 Séries Temporais e Suas Características

Em termos práticos, uma série temporal nada mais é que uma sequência de observações ordenadas no tempo. Quando se tem um experimento, por exemplo, em que uma determinada variável é coletada a cada minuto ao longo de uma semana, ou um mês, ou um ano, qualquer que seja o período, tem-se uma série temporal indexada neste período. No caso, a periodicidade em que as amostras são realiza-

das define a granularidade das amostras, o que também é um fator importante no estudo de séries temporais. Isso porque, de acordo com a granularidade, um comportamento da variável analisada pode ser camuflado ou exposto se utilizada uma granularidade baixa ou alta, respectivamente.

Um exemplo mais próximo do usual pode ser a quantidade de chuva diária que precipita durante um ano. Neste caso, a variável é dada em milímetros, somando 365 amostras, com granularidade de um dia.

Pode-se dizer também que o conceito de séries temporais remete ao conceito de processo estocástico, definido como uma família de variáveis aleatórias indexadas por outra variável, geralmente o tempo [24]. Assim, em relação ao tráfego de rede coletado ao longo do tempo, pode-se compreender este como uma realização de um processo estocástico sobre uma variável dada em *bytes/s*, coletada segundo uma granularidade específica.

Uma característica intrínseca das séries temporais é a forte dependência das observações adjacentes, no que tange aos níveis de autocorrelação que decaem lentamente para zero. Portanto, o estudo de séries temporais envolve técnicas para lidar com essas dependências, o que recorre ao desenvolvimento de modelos que considerem tais características.

Diversos modelos são encontrados na literatura para estudo de séries temporais [8, 9, 24], contudo, é necessário que sejam levados em consideração alguns fatores que influenciam na escolha de modelos, como a estacionariedade, sazonalidade, entre outros, que são listados nas seções a seguir.

2.3.1 Estacionariedade

Um processo é estritamente estacionário quando a distribuição conjunta para qualquer configuração de amostras é independente do tempo, ou seja, todos os momentos da distribuição conjunta não variam com deslocamento no tempo. Essa é a forma mais forte de estacionariedade [10].

Um processo pode ser também estacionário em sentido amplo (ou estacionário de segunda ordem). A diferença é que a estacionariedade em sentido amplo é mais fraca, requerendo apenas que os dois primeiros momentos sejam constante, isto é, média constante, variância constante e autocorrelação dependente somente da defasagem em relação aos instantes de tempo e não dos instantes de tempo em si [10]. Assim, neste texto, entenda-se estacionariedade por estacionariedade em sentido amplo.

Uma série temporal se encontra em estado estacionário quando seus valores são gerados aleatoriamente ao redor de uma média constante, apresentando um comportamento de equilíbrio estável, o que significa que a média da série não varia com o tempo. Entretanto, a maioria das séries temporais apresentam componentes que

as tornam não estacionárias, como é o caso do componente de tendência.

Assim, para obter dados estacionários, é preciso aplicar algumas transformações. Dentre elas, a forma mais comum é o processo de diferenças sucessivas, onde a primeira diferença é obtida da seguinte forma [9]:

$$d_1 = X(t) - X(t-1)$$

e a segunda diferença definida como

$$d_2 = d(d_1) = d(X(t) - X(t-1)) = X(t) - 2X(t-1) + X(t-2).$$

Assim,

$$d_n = d(d_{n-1}). \tag{2.7}$$

Para saber se a série está em estado estacionário, alguns testes podem ser aplicados. Na literatura são encontrados vários testes tais como:

- Teste de Dickley-Fuller Aumentado
- Teste de Phillips-Perron
- Teste KPSS

A descrição detalhada de cada teste foge ao escopo deste trabalho. Em [9], mais detalhes sobre os testes podem ser vistos. Neste trabalho foi utilizado apenas o Teste de Dickley-Fuller Aumentado.

2.3.2 Sazonalidade

A componente de sazonalidade é caracterizada quando a série temporal apresenta comportamento recorrente em um período de tempo, ou seja, fenômenos que ocorrem com uma repetição em seu padrão, com intervalos de tempo iguais entre suas ocorrências [9].

Existem dois tipos de sazonalidade:

- **Sazonalidade determinística:** quando a sazonalidade é definida como um padrão regular e estável no tempo;
- **Sazonalidade estocástica:** quando a componente sazonal varia com o tempo.

Alguns exemplos de sazonalidade podem ser pontuados, como é o caso da agricultura, que tem os períodos de estiagem com ciclos anuais; no trânsito das cidades,

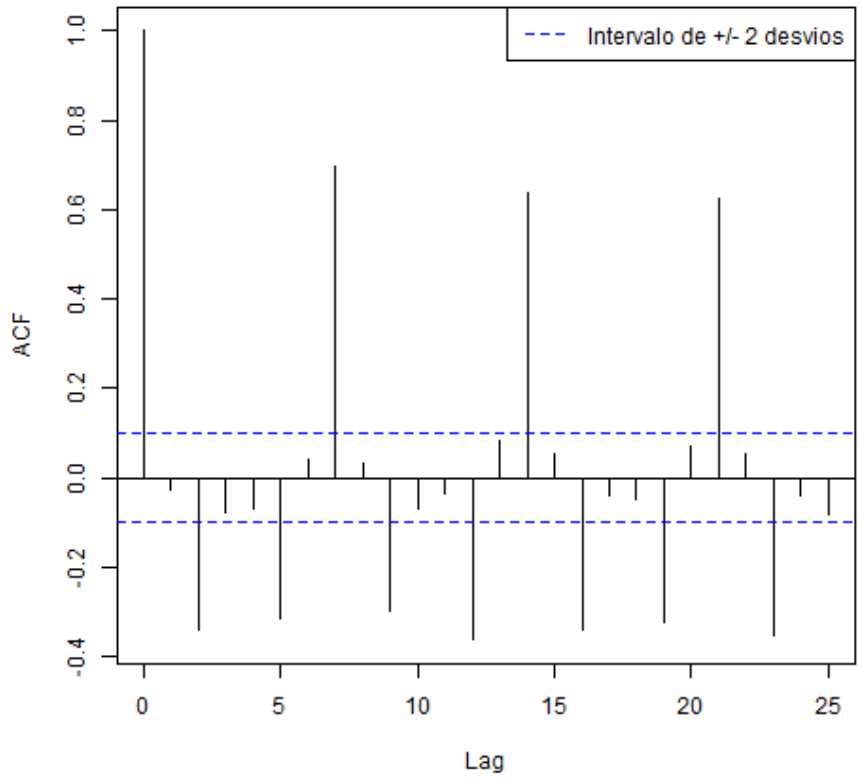


Figura 2.2: Exemplo de gráfico de autocorrelação

com a hora do *rush*; das estações do ano e até mesmo o aumento na procura por roupas de frio no inverno.

Em Redes de Computadores, apesar dessa periodicidade geralmente não ser determinística, ainda assim é possível identificar períodos de repetição, como é o caso dos finais de semana que apresentam uma utilização mais baixa da rede para diversas aplicações.

É importante perceber essa característica, pois com ela é possível identificar a necessidade de se utilizar um modelo que a considere, permitindo uma análise mais precisa do objeto de estudo.

Uma das formas para identificar a presença de sazonalidade é observar os gráficos de autocorrelação e autocorrelação parcial, nos quais os padrões de repetição ficam claramente evidentes. Pela Figura 2.2 é possível perceber uma repetição nos valores de autocorrelação a cada 7 defasagens, o que indica uma sazonalidade de período 7. Mais sobre autocorrelações será abordado na Seção 3.2.3.

2.3.3 Tendência

A componente de tendência é responsável por alterar o primeiro momento da distribuição conjunta do processo estocástico em questão, afetando a média do processo, tornando-o não estacionário.

Algumas formas de identificar componentes de tendência são pelo ajuste de polinômios (ou exponencial) ou pelo ajuste por reta de mínimos quadrados [9]. Após estimar o componente de tendência, este pode ser removido da série original. Um processo usual para remoção de tendência é o processo de diferença mencionado anteriormente na Seção 2.3.1.

2.3.4 Autossimilaridade

Visualmente pode-se compreender o conceito de autossimilaridade como sendo um modelo fractal, onde um determinado objeto é formado recursivamente por partes com formas semelhantes a este mesmo objeto, mas em escalas diferentes, também chamado *Noah Effect* [25]. Em termos matemáticos, a autossimilaridade pode ser entendida como a presença de distribuições que apresentam dependência de longa duração, que por sua vez são encontradas nas distribuições de cauda pesada. A Equação (2.8) descreve matematicamente a autossimilaridade [26].

$$Y_t = m^{-H} \sum_{i=(t-1)m+1}^{tm} Y_i \quad \textit{para todo } m \in \mathbb{N} \tag{2.8}$$

Na Equação (2.8), Y_t é uma série temporal estacionária, com média zero. A série Y_t é *H-autossimilar* se a agregação de m séries Y_t não sobrepostas e reescalonadas de m^H apresenta a mesma distribuição que Y_t, para todo m positivo.

O parâmetro H é conhecido como parâmetro de Hurst e está definido no intervalo $(0, 1)$. Quando uma distribuição apresenta um valor para H menor que $0,5$, então trata-se de uma distribuição de dependência de curta duração (calda leve). Se a distribuição apresenta um valor de H maior que $0,5$, então diz-se que a distribuição tem dependência de longa duração (calda pesada), também conhecido como *Joseph Effect* [25].

O parâmetro de Hurst serve ainda para indicar o quanto uma distribuição se assemelha a um tráfego de rajada.

Um exemplo de alta variabilidade em diferentes escalas de tempo, presente nas distribuições que apresentam essa propriedade, pode ser vista na Figura 2.3. Essa característica traduz o que acontece quando se tem tráfego em rajadas. O conceito de rajada é a ocorrência correlacionada entre dois ou mais picos de tráfego sucessivos (alta taxa de transferência), seguidos por um intervalo de tempo longo de ociosidade,

ou baixa taxa de transferência [27].

A Figura 2.3, extraída de [25], ilustra bem o conceito de autossimilaridade. É possível acompanhar, em cada uma das colunas, o comportamento do tráfego gerado em 5 níveis de granularidade (escala de tempo) distintos. A coluna esquerda mostra o tráfego gerado pelo tráfego de uma rede local. A coluna do meio mostra o tráfego sintético gerado a partir de uma distribuição para modelagem tradicional, Poisson por exemplo. A terceira coluna mostra também um tráfego sintético, porém considerando um modelo com característica autossimilar. As regiões destacadas em cada imagem correspondem ao período de tempo observado na granularidade seguinte.

Tomando como exemplo a primeira e a terceira coluna, nota-se que o tráfego mantém as mesmas características, mesmo em diferentes escalas de tempo. Esse comportamento não se verifica para a coluna do meio, cujo tráfego vai se tornando uniforme à medida que a granularidade é diminuída, o que pode ser observado analisando a figura a partir dos últimos gráficos em direção aos primeiros.

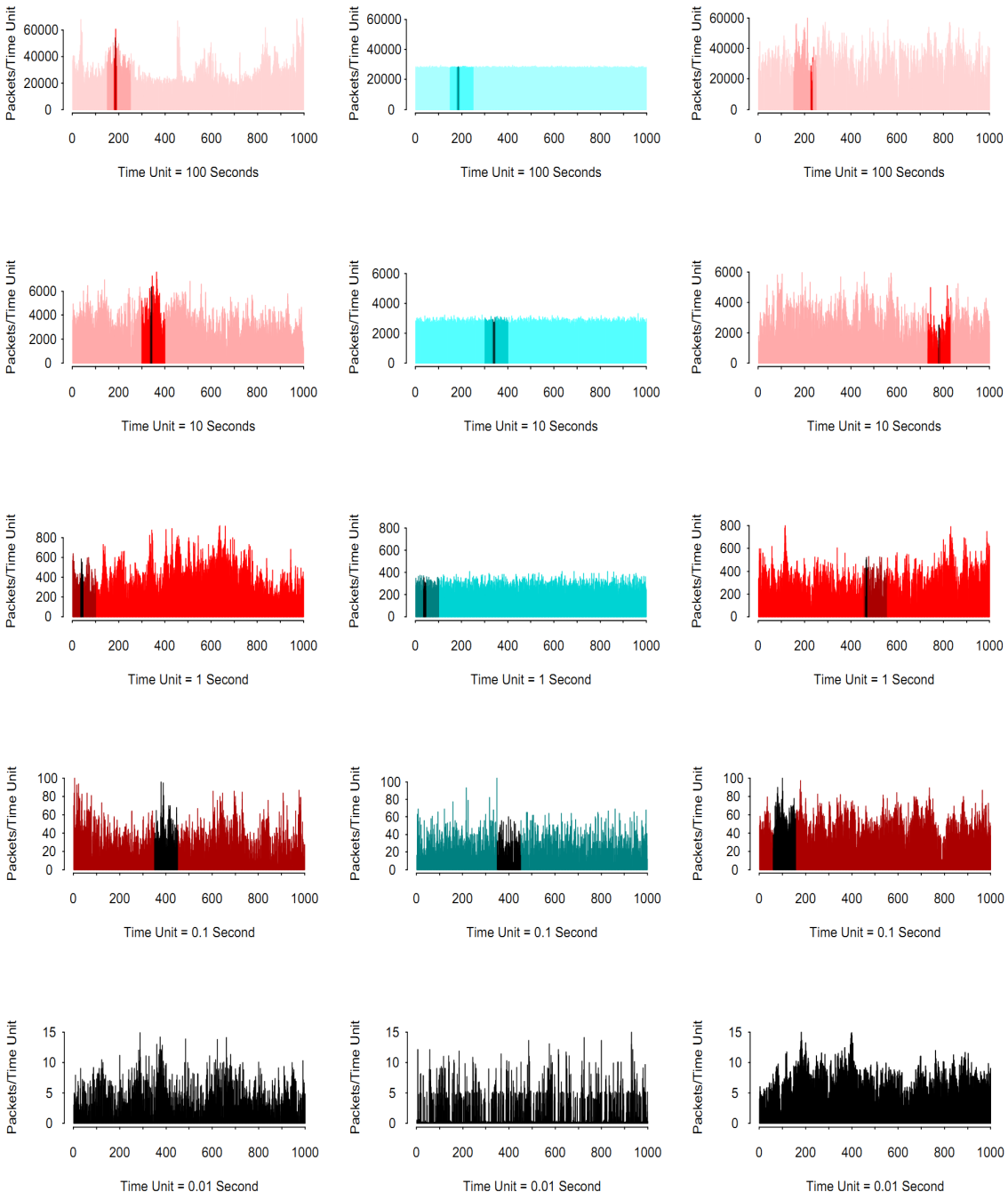


Figura 2.3: Exemplo de tráfego de rede local (coluna da esquerda), tráfego sintético gerado a partir de um modelo tradicional (coluna do meio); e tráfego sintético gerado a partir de um modelo autossimilar (coluna da direita); para 5 escalas de tempo diferentes. Regiões destacadas representam a mesma porção de tráfego para escalas de tempo distintas [25]

2.4 Testes Estatísticos

Quando se deseja realizar o ajuste de uma distribuição conhecida a um conjunto de dados coletados, é necessário que sejam feitos alguns testes estatísticos. Estes testes consistem em comparar dentre diversas distribuições conhecidas, a fim de selecionar a que mais se aproxima das características estatísticas dos dados reais. Na literatura, estes testes são conhecidos como testes de ajuste, ou do inglês, *goodness of fit tests*.

Nesta seção são explicados alguns dos testes mais utilizados para classificar se uma distribuição se ajusta bem ou não.

2.4.1 Teste Qui-quadrado

O teste do Qui-Quadrado [10, 22] é um dos mais utilizados para determinar se as amostras obtidas seguem uma determinada distribuição. Geralmente este teste é mais aplicado para variáveis discretas, mas também pode ser aplicado para variáveis contínuas. Neste último caso, perde-se um pouco de precisão no teste.

A estatística utilizada no teste do Qui-Quadrado é mostrada na Equação (2.9).

$$X^2_{k-1} = \sum_{i=0}^{k-1} \frac{(N_i - Np_i)^2}{Np_i} \tag{2.9}$$

Na Equação (2.9), N_i é uma amostra observada, N é o número total de amostras observadas, p_i é a probabilidade de ocorrência de uma amostra, k é a quantidade de valores possíveis que a variável em questão pode assumir, Np_i é o valor esperado de N_i e X^2_{k-1} é a estatística Qui-Quadrado com $k-1$ graus de liberdade.

Uma outra forma de escrita que pode ser adotada para melhor compreensão pode ser vista na Equação (2.10).

$$X^2_{k-1} = \sum_{i=0}^{k-1} \frac{(observado - esperado)^2}{esperado} \tag{2.10}$$

2.4.2 Teste Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov [22] é o teste mais aplicado para variáveis contínuas, sendo este mais preciso que o teste do Qui-Quadrado. A Equação (2.11) expressa o teste.

$$D_N = \sup_x |F_N(x) - F_0(x)| \tag{2.11}$$

O valor supremo é um limite superior para uma função em um determinado intervalo. O valor supremo pode ser o mesmo que o valor máximo se este se encontra dentro do intervalo observado, mas não deve ser confundido como tal. Uma função

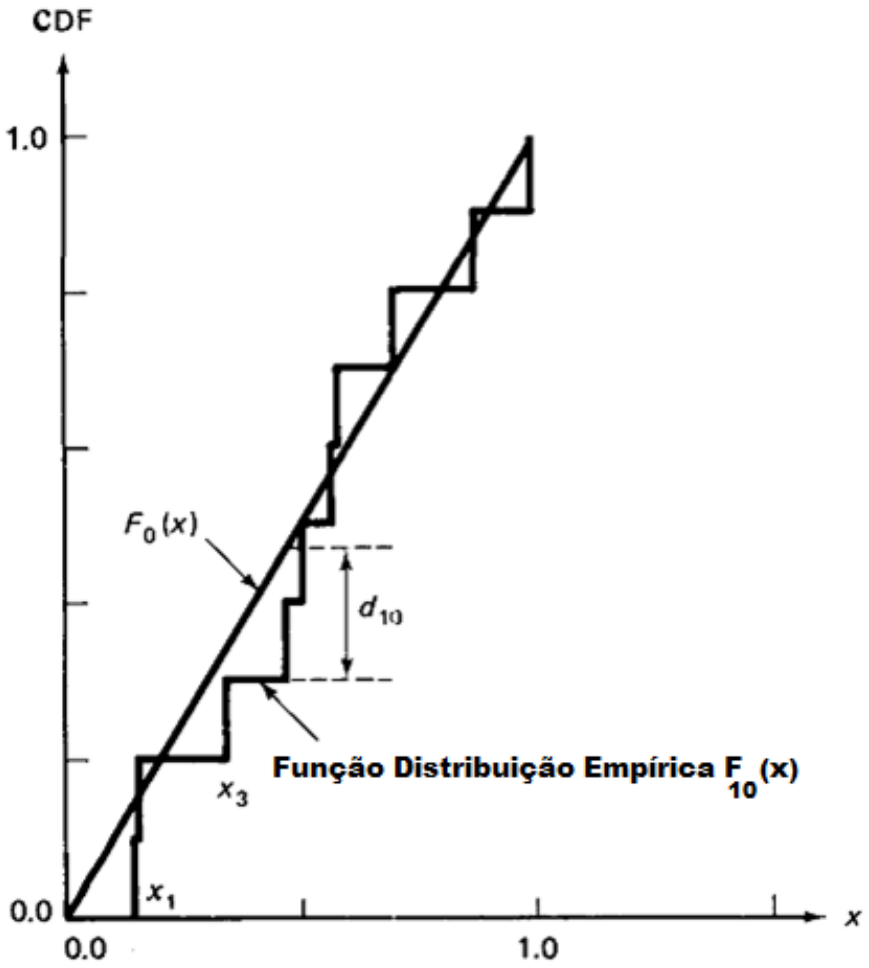


Figura 2.4: Ilustração do cálculo do valor supremo para D_N

sempre tem o valor supremo, ainda que seja infinito, mas nem sempre tem valor máximo, como nesse mesmo exemplo de ser infinito. A Figura 2.4 ilustra a obtenção do valor supremo para o cálculo de D_N, considerando um exemplo com 10 amostras.

O teste pode ser feito utilizando duas abordagens distintas. Uma considera como hipótese nula a comparação entre duas funções empíricas cumulativas de probabilidade, quando se deseja comparar dois conjuntos de amostras. Outra abordagem é ter como hipótese nula apenas uma distribuição empírica utilizando N amostras ($F_N(x)$) e outra distribuição teórica ou conhecida na literatura ($F_0(x)$), que é o caso utilizado neste trabalho.

Para comparar as duas distribuições, as probabilidades cumulativas são calculadas e comparadas através da distância absoluta entre elas. O valor supremo que for encontrado através do cálculo dessa distância será o valor obtido para a estatística do teste. Então, o valor da estatística é comparado com o valor tabelado e decide-se se a hipótese nula deve ser aceita, indicando o ajuste da distribuição empírica, ou não.

2.4.3 Critério de Informação de Akaike

O *Akaike Information Criterion* – AIC (Critério de Informação de Akaike) é um teste que fornece uma medida relativa à qualidade do modelo, levando em consideração não só o ajuste aos dados (*goodness of fit*), mas também o número de parâmetros que o modelo apresenta, penalizando os que requerem mais parâmetros.

Em [28], o autor sugere a escolha de um modelo que minimize o critério, levando em consideração os k parâmetros estimados, como na Equação (2.12).

$$AIC = -2 \log V + 2k \tag{2.12}$$

onde V é máxima verossimilhança para o modelo em questão e k é o número de parêmtros ajustados no modelo.

Mais informações sobre o AIC podem ser encontradas em [28], [9] e [29].

2.5 Estado da Arte

Diversas abordagens podem ser encontradas na literatura envolvendo o problema de previsão de tráfego. Um dos primeiros trabalhos, e também um dos mais citados, que aborda este tipo de problema é o trabalho de Groschwitz e Polyzos [30]. Neste trabalho, os autores utilizam como estudo de caso a NSFNET, que há duas décadas atuava como provedor acadêmico, também sendo um dos primeiros *backbones* da Internet.

O trabalho de Groschwitz e Polyzos reforça a ideia do crescimento constante no volume de tráfego ser uma característica dominante em *backbones*. A ferramenta matemática utilizada pelos autores foi o Processo Autorregressivo Integrado de Média Móvel (ARIMA), o mesmo utilizado nesta dissertação.

Ainda em [30], os autores utilizam um conjunto de dados correspondente à quantidade total de pacotes diária em todos os *links*, por um período de 4 anos e 11 meses, o que resultou em 1794 amostras. Porém, para simplificar a análise em termos de poder de computação na época em que o trabalho fora realizado, os autores optaram por colapsar os dados diários em semanas, reduzindo o conjunto de dados para 256 amostras.

No caso, utilizando a granularidade de semanas, para prever 1 ano, são 52 valores a serem previstos. Tendo em vista que o trabalho propõe uma previsão de 1 ano (e também de 2 anos), a quantidade de dados prevista corresponde aproximadamente a um quinto do número de amostras. O trabalho conclui com as previsões mencionadas, porém não apresenta o intervalo de confiança, o que contribuiria para uma melhor avaliação dos resultados obtidos. A Figura 2.5 apresentada no referido trabalho ilustra a previsão realizada pelos autores.

A proposta apresentada nesta dissertação tem a referência [30] como base, utilizando também o modelo ARIMA. Porém, os dados coletados e analisados são relativos à vazão no enlace, em vez do número de pacotes, como abordado pelos autores. Utilizar a vazão pode dar mais credibilidade à análise, devido ao fato da quantidade de pacotes em si não refletir uma característica mais próxima do real em relação ao tráfego, por conta da variação no tamanho dos pacotes.

Outro trabalho relevante é o de Willinger *et al.* [25], que aborda a questão da existência da propriedade estatística de autossimilaridade no tráfego de redes. Os autores buscam responder a duas perguntas: uma explicação física para tal propriedade estar presente no tráfego; e qual o impacto essa mudança teria nos protocolos, em termos de avaliação de desempenho. O trabalho identifica a alta variabilidade, também conhecida como *Noah Effect*, como fator essencial para a presença de autossimilaridade, através de análise estatística sobre amostras de tempo real, observadas em fontes individuais de tráfego Ethernet em redes locais. Willinger

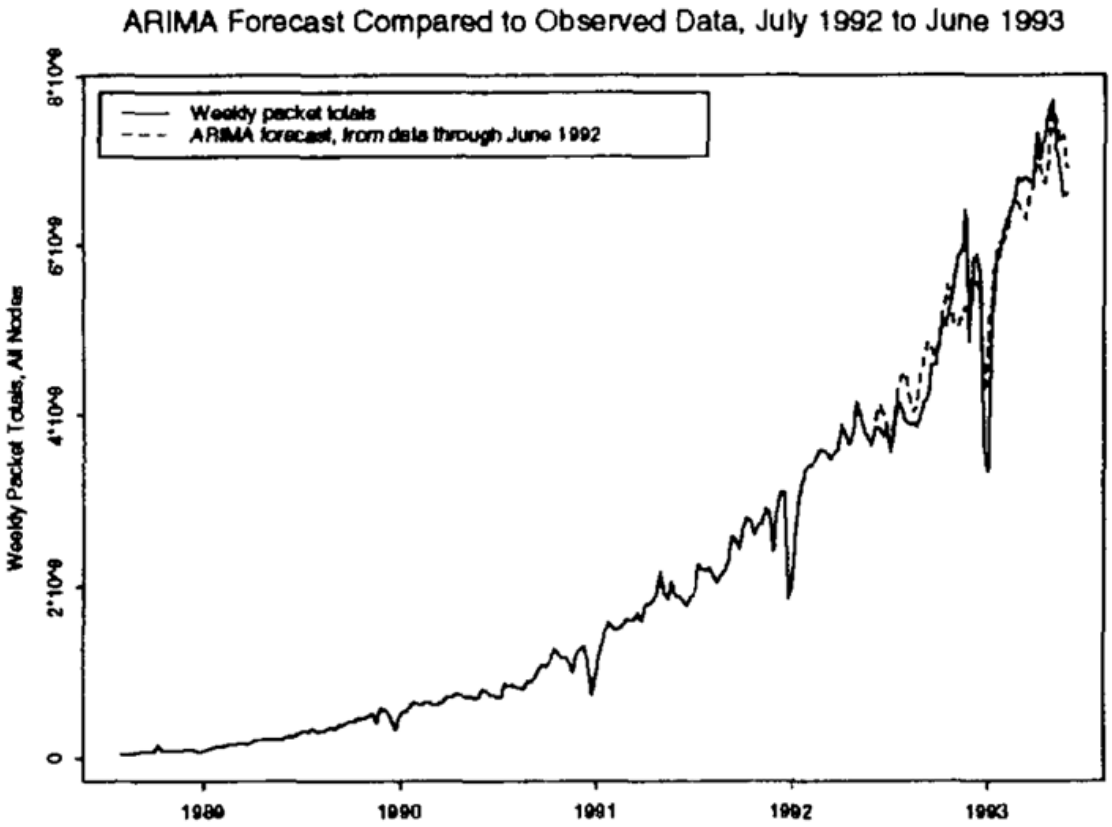


Figura 2.5: Previsão realizada no trabalho de Groschwitz e Polyzos [30]

et al. consideram fontes de dados *On-Off*, o que significam fontes que passam um período gerando tráfego (*On*) e outro período sem geração de tráfego (*Off*). A propriedade de autossimilaridade será abordada com mais detalhes na Seção 2.3.4.

Apesar de [25] considerar apenas tráfego de rede local (e não de *backbone*), a autossimilaridade, no contexto desta dissertação, é importante ser considerada, pois esta característica explica determinados comportamentos do tráfego que podem surgir, conforme analisado em outros trabalhos [26, 31–33].

Feng e Shu [34] apresentam as técnicas mais comuns para realização de previsão de tráfego, mostrando um comparativo entre previsores paramétricos e não-paramétricos, com relação à complexidade computacional. Os autores chegam à conclusão de que os modelos ARIMA e FARIMA (ARIMA Fracionário) são complexos computacionalmente ($O(n^2)$) em comparação a outros modelos analisados. Em contrapartida, o modelo com base no ARIMA é mais preciso, em termos de Erro Quadrático Médio (EQM), que os modelos baseados em *Artificial Neural Networks – ANN* (Redes Neurais Artificiais).

De acordo com os autores, a técnica de *wavelet* é a de menor custo computacional, seguida de ANN, ARIMA e FARIMA. Porém, os modelos baseados no ARIMA suportam tanto tráfego de curta dependência, quanto de longa dependência. Com o

modelo ARIMA ganha-se precisão em detrimento ao custo computacional. Sugere-se então que o ANN seja um bom modelo para previsão *on-line* [2], enquanto que o ARIMA funcione bem como previsor *off-line* [3].

O trabalho [11] apresenta uma metodologia para decidir *quando* e *onde* devem ser feitas atualizações em relação aos enlaces de um *backbone*. Os autores utilizam análise *wavelet* multirresolução, junto com modelos de séries temporais lineares como o ARIMA. O trabalho utilizou informações de três anos de coletas através de SNMP, realizando previsões de pelo menos 6 meses no futuro, com uma granularidade de 12 horas. Para decidir *onde* a atualização de um recurso deve ser executada, as análises foram feitas entre PoPs (*Point of Presence* – Ponto de Presença) adjacentes. Decide-se *quando* a atualização deve ser executada através das previsões do tráfego com o modelo escolhido. O trabalho não leva em consideração o uso de modelos que compreendam sazonalidade, como o SARIMA, e também não apresenta os intervalos de confiança.

O trabalho de Maia e Filho [35] leva em consideração a utilização de aprendizado de máquina (*machine learning*) para previsão de tráfego, utilizando uma metodologia baseada em tráfego origem-destino multidimensional, com foco na previsão de curto prazo. O trabalho usa uma abordagem de fluxos para reduzir o número de variáveis e o processamento computacional, fazendo uso de uma técnica de redução de dimensionalidade, a Análise de Componentes Principais. Esta técnica reduz o número de componentes, considerando apenas aqueles que mais contribuem nas variações nos fluxos. Para realização da previsão e análise de tendência, os autores utilizaram o algoritmo *K-means*. Apesar do trabalho apresentar resultados satisfatórios em termos de margem de erro, as técnicas de *machine learning* são mais indicadas em trabalhos de previsão de curto prazo, quando a proposta desta dissertação se diferencia.

No trabalho [14], os autores mostram que o tráfego de *BitTorrent* também pode ser previsto através de uma análise de séries temporais usando o modelo ARMA. O foco do trabalho se dá em proporcionar uma ferramenta para que os provedores de serviço (*Internet Service Provider - ISP*) consigam gerenciar a largura de banda, em termos do tráfego gerado sob o protocolo *BitTorrent*, atuando na qualidade do serviço prestado. Isso porque o protocolo *BitTorrent* tem a característica de aumentar a utilização da largura de banda, tanto no momento em que um cliente faz o *download* das partes de um arquivo, quanto no momento em que este faz *upload* das partes que já tem. Isso contribui para um aumento no volume de tráfego, o que gera latência e perda de pacotes.

[2] Previsões que fornecem resultados em tempo real, ou próximo ao instante de coleta dos dados

[3] Previsões que fornecem resultados depois de certo tempo (dias ou meses, por exemplo) em que os dados foram coletados

O objetivo dos autores em [14] é prever o tráfego gerado pelas aplicações *BitTorrent*, a fim de melhorar os fluxos dentro da rede do ISP, bem como minimizar a troca de pacotes entre ISPs, o que impacta na QoS dos não-clientes *BitTorrent*. O trabalho conclui que o tráfego gerado pelos *seeds*[4] pode apresentar sazonalidade (ou padrões cíclicos) e é passível de ser previsto com o modelo ARMA. Contudo, não fica claro se o trabalho explora outras parametrizações para o modelo, o que possivelmente melhoraria os resultados de previsão, bem como a utilização do modelo SARIMA para a componente sazonal.

Um exemplo de aplicação do modelo ARMA(1,1) para modelar o tráfego gerado por jogos de tiro em primeira pessoa, com dois ou três jogadores, pode ser encontrado em [36]. Os autores buscam extrapolar o modelo para um número maior de jogadores, observando que os modelos podem ser compostos através de modelos mais simples para um número menor de jogadores. O trabalho observa a caraterística desse tipo de tráfego ser assimétrica, semelhante ao tráfego de requisições HTTP [26]. Em outras palavras, o tráfego de um cliente para um servidor é formado por pacotes de menor tamanho, enquanto que o tráfego do servidor para os clientes é maior, variando conforme o número de jogadores. O trabalho mostra que modelos ARMA(1,1) podem ser utilizados na forma proposta, sendo uma alternativa para análise do desempenho em redes.

Outros trabalhos abordam também a existência de classificações quanto aos fluxos de dados na Internet. Por vezes esses fluxos são apelidados com nomes de animais, de acordo com suas características. Os casos mais comum são os fluxos elefante e fluxos rato. Fluxos elefante são aqueles que levam uma grande quantidade de dados, ou que duram um tempo considerável. Já os fluxos rato são aqueles que apresentam curta duração ou levam poucos dados [37].

Os trabalhos que envolvem esse assunto mostram que existe uma proporção entre as classificações, sendo os fluxos rato em maior quantidade que os fluxos elefante [31, 38, 39]. Porém, os fluxos elefante "esmagam" os fluxos rato. Em outras palavras, por terem maior quantidade de dados para trafegar, os fluxos elefante conseguem uma largura de banda maior, dado que estes ocupam por mais tempo o canal de transmissão. Ao mesmo tempo, os fluxos rato não conseguem usufruir da largura de banda por serem de menor duração. Uma das causas para este fato que as referências pontuam é a larga utilização do protocolo TCP, que tem como um dos mecanismos de controle o *slow-start* [3], um crescimento exponencial da janela de congestionamento. Ressaltar esses resultados permite ter uma ideia mais abrangente do comportamento do tráfego, que pode influenciar consideravelmente durante o processo de previsão.

O trabalho de Cortez *et al.* [40] apresenta três métodos de previsão aplicados ao tráfego de dois provedores, inclusive o método ARIMA. O foco do trabalho é

[4] *seeds* são usuários que possuem o arquivo completo para compartilhamento

na previsão de curto prazo e, para tal propósito, é obtido o resultado que usar a abordagem de Redes Neurais ou Holt-Winters é melhor, neste caso. Os resultados obtidos em relação ao ARIMA não foram tão satisfatórios, nos exemplos mostrados, mas vale ressaltar que foram utilizados no máximo 2 meses de amostras, o que impacta consideravelmente nos resultados para tal modelo.

A diferença do trabalho apresentado nesta dissertação para o de Cortez *et al.* é o foco na previsão de longa duração, o que permite que decisões estratégicas sejam tomadas para um período mais longo no futuro, no qual o modelo ARIMA apresenta melhores resultados.

A abordagem escolhida para ser utilizada nesta dissertação se baseia em análise de séries temporais (Seção 1.2), utilizando métodos estatísticos para resolução de problemas deste tipo, como é o caso do modelo ARIMA. Esta abordagem proporciona acurácia em tráfego de curta e longa dependência [34]. Além disso, as previsões são realizadas de maneira *off-line*, visto que o objetivo é o planejamento e o provisionamento de recursos de rede de forma antecipada.

Sendo assim, pelo crescimento indiscutível do tráfego nas grandes redes e pela diversidade de trabalhos relacionados, fica claro ser uma área de interesse, que ainda tem bastante atenção pela comunidade de Redes de Computadores, como pode ser visto nos trabalhos discutidos e também em outros [12, 14, 41, 42].

Capítulo 3

Metodologia

Estudar o comportamento do tráfego que passa por uma rede é extremamente importante, pois este comportamento influencia diretamente no desempenho geral da rede, o que impacta na qualidade do serviço prestado. Com isso, analisar o tráfego de forma antecipada contribui para que a infraestrutura de rede sempre tenha recursos disponíveis para suportar as novas demandas.

Neste trabalho é apresentada uma metodologia baseada nos trabalhos [8], [11] e [30], que auxilia no processo de caracterização do tráfego em redes de *backbone*, bem como da previsão do comportamento do tráfego em um período de tempo futuro, utilizando a abordagem por séries temporais. O trabalho [43] é uma excelente referência em relação à caracterização de tráfego em *backbones*, também servindo de base para o presente trabalho.

A metodologia apresentada a seguir tem como premissas a simplicidade, a independência de tecnologias proprietárias, o desempenho computacional e a otimização de custos financeiros.

Sendo assim, uma das contribuições deste trabalho é o estabelecimento de uma metodologia que sirva de arcabouço para o desenvolvimento de um sistema de previsão de tráfego. O intuito é atuar de forma simplificada no planejamento e provisionamento de recursos para a infraestrutura de rede necessária para suportar uma demanda prevista.

Na Seção 3.1, algumas considerações são feitas com o intuito de elencar alguns cenários possíveis para introdução da metodologia. Na Seção 3.2 são descritas as etapas que envolvem a metodologia, detalhando cada uma das etapas.

3.1 Considerações Iniciais

Diversos cenários são explorados no contexto de previsão quando se trata de Redes de Computadores. Por exemplo, se um novo cliente deseja ter acesso através de um novo *link*, as seguintes perguntas podem surgir:

- Qual a demanda que este novo cliente traz para a rede?
- Qual o impacto desta demanda no que tange ao consumo dos recursos de rede?
- Como esta demanda influencia na qualidade do serviço dos clientes que já fazem parte da rede?

Essas perguntas revelam uma necessidade intrínseca de conhecer, antecipadamente, qual o comportamento das possíveis fontes de dados envolvidas nos processos de entrada de novos clientes (novas fontes de dados), bem como a mudança de comportamento ou uma variação na demanda de um ou mais clientes.

Deve ser tomado o cuidado para que novos clientes não prejudiquem o desempenho de clientes já existentes, bem como o cuidado para que um novo cliente não entre na rede tendo uma percepção ruim desta. Assim, é razoável fazer um planejamento dos recursos a fim de antecipar condições da rede para prever problemas de congestionamento.

Para atender às necessidades citadas, realiza-se a previsão sobre as amostras coletadas ao longo do tempo. Através dessa previsão, identifica-se um comportamento que pode revelar uma condição de congestionamento, permitindo uma atuação próativa no que diz respeito à atualização dos recursos utilizados.

Paralelamente à previsão, realiza-se a caracterização dos dados. O objetivo é revelar qual padrão estatístico mais se assemelha à realidade de uma fonte de dados específica. Em outras palavras, através da caracterização é possível identificar uma massa de dados como uma aproximação para amostras geradas a partir de uma distribuição de probabilidade conhecida, como por exemplo, a distribuição de Poisson, a distribuição Normal, Exponencial, dentre outras. Desta forma, permite-se compreender e simular o comportamento de tais fontes, podendo assim ter uma base para responder às questões anteriormente levantadas.

Contudo, alcançar esses objetivos para uma topologia completa de rede pode ser uma tarefa árdua e até mesmo desnecessária. Isso porque, considerando uma metodologia simples, levando em conta a perspectiva de análise orientada a enlaces, alcançar tais objetivos se torna uma tarefa de menor complexidade, já que nem toda a rede será analisada.

Sendo assim, neste trabalho é apresentada uma metodologia que tem foco na previsão de tráfego aplicada à infraestrutura de um provedor de acesso, na qual objetiva-se estudar o comportamento futuro do tráfego na topologia de rede atual. Considera-se uma perspectiva de análise de enlaces individuais, alcançando os resultados esperados em relação à previsão, consequentemente, expandindo os resultados para a topologia como um todo.

3.2 Etapas da Metodologia

Para tornar a metodologia mais clara e objetiva, abaixo são apresentadas as etapas que a compõem. Esta metodologia se baseia em [11, 30] que abordam análise de séries temporais.

As etapas listadas abaixo são ilustradas na Figura 3.1.

1. **Coleta de Dados:** Esta etapa considera a obtenção dos dados de interesse;
2. **Filtragem dos Dados e Extração de Medidas:** Nesta etapa é feito o tratamento dos dados coletados;
3. **Análise Estatística e Escolha de Modelos:** Nesta etapa é feita a escolha de modelos a serem utilizados para previsão e caracterização dos dados;
4. **Previsão e Cálculo de Desvio:** Nesta etapa são realizadas as previsões e os ajustes necessários para reduzir os desvios de previsão;
5. **Tomada de Decisão:** Nesta etapa são decididas quais ações serão executadas com base nos resultados obtidos.

A seguir, cada etapa da metodologia é detalhada. A metodologia em si é o componente mais importante, formando as bases que direcionam todo o estudo.

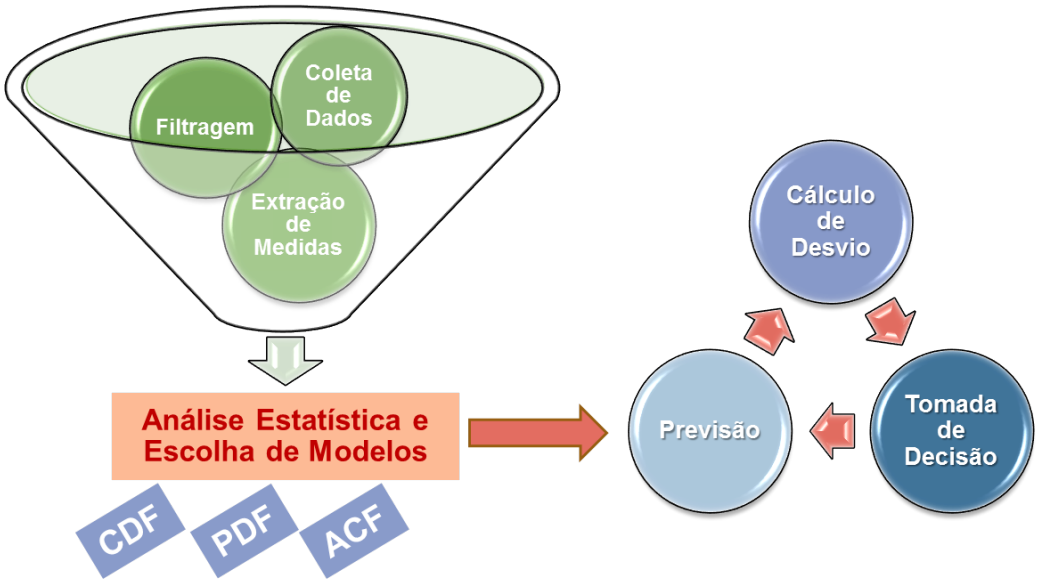


Figura 3.1: Metodologia para previsão de tráfego

3.2.1 Coleta de Dados

Algumas definições básicas devem ser feitas para que não haja divergência no entendimento ao longo do texto.

Enlace (ou *link*) é uma ligação entre dois equipamentos de rede na qual é possível a troca de dados em ambos os sentidos (*full-duplex*).

Interface é um enlace entre um equipamento de rede do provedor e um equipamento de rede do cliente do provedor. Este conceito também é conhecido como **salto de última milha**, o qual efetivamente liga o provedor ao cliente. Ou seja, toda interface é um enlace, mas nem todo enlace é uma interface. Um exemplo prático seria a interface de um roteador da RedeRio com um roteador da UFRJ, o que significa o ponto de rede em que a UFRJ se conecta à rede da RedeRio.

Outro conceito importante é o **sentido do tráfego**, que tem duas direções. O tráfego de entrada de uma interface, é aquele com origem em uma rede externa qualquer, destinado à rede da instituição associada àquela interface, conhecido como tráfego de *downlink*. Já o tráfego de saída da interface é o tráfego enviado pela instituição para outras redes externas a essa, também conhecido como tráfego de *uplink*. A Figura 3.2 ilustra esse conceito.

Dito isto, também é importante frisar os detalhes da coleta de dados em si. Neste trabalho é utilizada a coleta de dados ativa através do protocolo SNMP (*Simple Network Management Protocol*), mais especificamente, utilizando o programa MRTG (*Multi Router Traffic Grapher*) [44]. O uso deste protocolo facilita a obtenção dos dados sobre o tráfego de cada interface, através de ferramentas de monitoramento dispostas nos roteadores de borda de um provedor. Como estudo de caso, será apresentado um exemplo utilizando uma rede acadêmica de computadores, a saber, a RedeRio de Computadores/FAPERJ.

Os dados de interesse coletados para este trabalho são:

- A quantidade média de *bytes* entrando em cada interface;
- A quantidade média de *bytes* saindo de cada interface;
- Os clientes associados a cada interface.

O SNMP foi escolhido por ser um protocolo padronizado, independente de tecnologia proprietária, o que é uma das premissas deste trabalho. Além disso, a rede utilizada no estudo de caso (RedeRio) já apresenta infraestrutura de monitoramento com SNMP montada, realizando a coleta dos dados de cada interface com as instituições associadas a esta.

Durante a coleta, a granularidade das amostras também é relevante. Como a coleta foi feita utilizando o MRTG, este calcula médias para diferentes escalas de tempo. As amostras diárias são obtidas através de médias dos valores coletados,

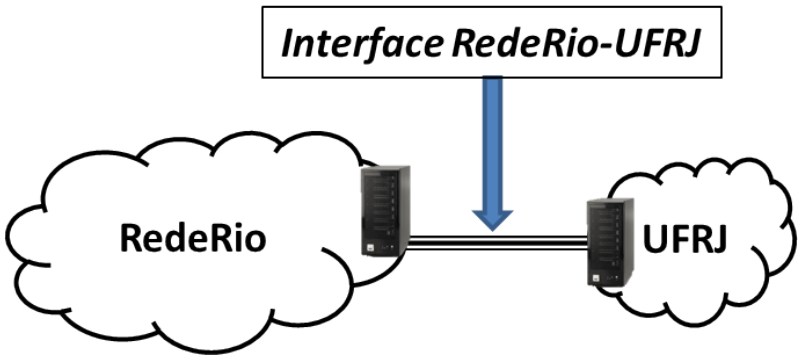


Figura 3.2: Exemplo do conceito de interface

calculadas em diferentes granularidades. Primeiramente, as amostras são coletadas com granularidade de 5 minutos e são mantidas por uma janela de tempo de um dia, ou seja, são mantidas 288 amostras com granularidade de 5 minutos. A partir dessa granularidade, outras médias são calculadas. A média com granularidade de 30 minutos é calculada a partir de 6 amostras de 5 minutos, sendo mantidas durante uma janela de tempo de uma semana. Depois é calculada a média a cada 2 (duas) horas, com 4 amostras de 30 minutos, mantidas por uma janela de tempo de um mês. Por fim, as médias de 1 (um) dia são calculadas com 12 amostras de 2 (duas) horas cada uma, as quais compõem um intervalo de amostras de um ano, correspondendo a 365 amostras.

As diferentes granularidades das amostras têm impacto no momento da previsão em relação a qual horizonte de previsão é desejado e qual a quantidade de amostras no passado são necessárias para a realização de tal previsão. Por exemplo, dado um ano de amostras de médias diárias (365 amostras), qual seria o tamanho da janela de previsão possível de ser alcançado, respeitando limites de erros previamente estabelecidos? A resposta exata para esta pergunta pode ser um tanto quanto difícil de ser encontrada, porém é intuitivo sugerir que quanto mais dados se tem, mais precisas as previsões se tornam. Em [45] esta questão fica mais clara através da discussão sobre quantos dados são necessários para que modelos de previsão sazonais possam ser utilizados. O trabalho conclui que a quantidade de dados depende do modelo usado e que o número mínimo de dados necessários seria dado pela quantidade de parâmetros que devem ser estimados para o modelo.

3.2.2 Filtragem dos Dados e Extração de Medidas

No processo de coleta de dados, em geral, os dados capturados vêm num formato que por vezes necessitam de tratamento. Esse processo de tratamento, comumente chamado de *parse*, é realizado para que se possa extrair as medidas de interesse.

A etapa de filtragem consiste também em remover alguns dados que não são rele-

vantes para a análise em questão. Ou seja, determinar quais os dados serão levados em consideração para estimação dos parâmetros do modelo na etapa de previsão. Esta etapa é de grande importância, pois influencia diretamente na acurácia das previsões, como será visto nos resultados do Capítulo 4.

Sendo assim, esta segunda etapa da metodologia envolve um *parse* das informações de *log* obtidas, bem como a remoção de amostras de interfaces que apresentam comportamento indesejado e que podem vir a tornar os resultados tendenciosos. É considerado um comportamento indesejado de uma interface quando esta apresenta um histórico de tráfego irrelevante, ou seja, tráfego nulo, o que pode ser causado por uma interface que tenha ficado muito tempo desativada ou não monitorada.

Para execução do *parse* é utilizada uma ferramenta de *software* de apoio, desenvolvida para este trabalho, que agiliza este processo. Essa ferramenta extrai as medidas de interesse, deixando-as prontas para a próxima etapa de análise estatística.

Também nesta etapa são utilizados *scripts* desenvolvidos para este trabalho, que automatizam o processo de geração dos gráficos para a caracterização dos dados, utilizando a ferramenta estatística R [46].

De acordo com a filtragem realizada, a granularidade das amostras pode ser afetada. Conforme visto na fase de coleta, os dados capturados apresentam diversas granularidades, ou seja, diversas escalas de tempo em que são caracterizados. Existem os dados coletados na maior granularidade, a cada 5 minutos, assim como os de menor granularidade, compostos pela média diária, como na Figura 3.3.

A filtragem pode ser realizada de diversas formas, dependendo do problema. No caso de previsão de tráfego para provisionamento de recursos de rede, são utilizadas amostras decorrentes da média de tráfego por dia em um determinado enlace.

Outra forma que pode ser utilizada é considerar as medidas de tráfego durante os momentos de maior utilização, descartando períodos da noite, por exemplo. Essa é

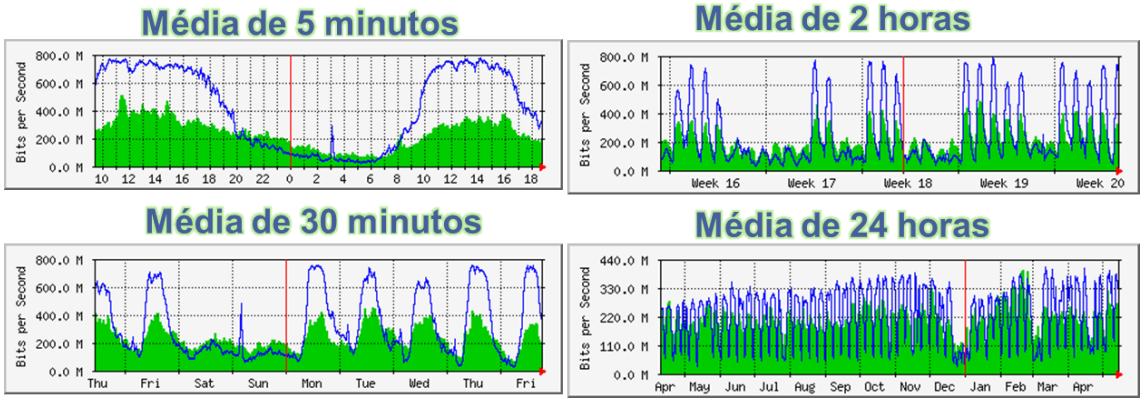


Figura 3.3: Diferentes granularidades dos dados coletados [15]

uma abordagem conhecida como hora de maior movimento (HMM), muito utilizada nas redes de telefonia e comunicação, onde se define uma faixa de horário em que as amostras serão consideradas. Um contraponto dessa abordagem seria a definição dessas faixas de horário, que é dependente do enlace sendo analisado, já que o perfil do tráfego é variável. Ou seja, o HMM de um enlace não necessariamente será o mesmo para outro enlace, o que dificulta uma análise em termos de *backbone*. A ideia é que se tenha um sistema que possa atuar de forma escalável, realizando a previsão, tanto a respeito de um enlace, quanto a um conjunto destes.

Uma terceira abordagem é a utilização de valores de pico diário, a fim de trazer para o modelo valores "pessimistas" sobre o tráfego, no sentido de que as amostras utilizadas se referem aos momentos em que o enlace estava em sua maior utilização possível, dentro daquela janela de tempo de coleta. Esta abordagem pode até ser mais adequada em alguns casos, visto que as previsões realizadas sob a perspectiva de valores de pico já cobririam os casos das medidas de tráfego inferior. Porém, é desejada também uma previsão que reduza custos por superestimação de recursos, o que sugere o uso de valores de média como uma melhor forma de aproximar a utilização do enlace.

Os valores de pico podem trazer informações tendenciosas devido a fatos isolados, como falhas de segurança, ataques, medições erradas, dentre outros fatores que acabam por tornar os valores de pico não tão representativos quanto os valores de média, quando a previsão realizada fica mais ajustada para o comportamento médio do tráfego no enlace, o que fica mais próximo da realidade.

Neste trabalho utilizam-se amostras decorrentes da média de tráfego diária e mensal em um determinado enlace. É importante mencionar que, como será visto no Capítulo 4, para o propósito de uma ferramenta que possibilite uma previsão para fins de provisionamento de recursos de rede, é razoável que essa previsão seja feita para um horizonte de meses. Em contrapartida, para ter uma quantidade de amostras significativa dos valores de médias com a granularidade de meses, é preciso ter uma quantidade maior ainda das amostras de 5 minutos para que as médias possam ser calculadas.

3.2.3 Análise Estatística

Nesta seção são descritas as etapas de caracterização dos dados, o ajuste de uma distribuição e o cálculo das autocorrelações, que são importantes para interpretação dos dados e para a etapa de identificação dos parâmetros dos modelos de previsão, como será visto na Seção 3.2.5.

Caracterização e Ajuste de Distribuição por PDF e CDF

No processo de Análise Estatística, as funções que caracterizam os dados, a saber, a Função de Densidade de Probabilidade (*Probability Density Function* – PDF) e a Função de Distribuição de Probabilidade Cumulativa (*Cumulative Distribution Function* – CDF), são de extrema importância para um melhor entendimento sobre o comportamento dos dados, permitindo que inferências sobre estes possam ser realizadas.

Como os experimentos realizados geralmente consideram dados empíricos, as funções de probabilidade acima mencionadas são dadas de forma empírica, realizando-se posteriormente um ajuste de alguma distribuição conhecida que mais se aproxime no sentido estatístico. Isto é, que segundo algum teste de ajuste (ex. Kolmogorov-Smirnov), aquela distribuição empírica pode ser entendida como um tipo da distribuição real.

Como um exemplo prático, suponha que seja realizada uma coleta de pacotes de dados numa rede local durante um período arbitrário de tempo. Se o intervalo de geração de pacotes apresentar um tempo exponencial entre as gerações, então quando for feito o ajuste de uma distribuição conhecida sobre os dados empíricos relativos à quantidade de pacotes que chegam até um determinado instante, isso resultará num ajuste que se aproxima de uma distribuição de Poisson, com uma taxa correspondente ao inverso do intervalo médio entre as gerações de pacotes. Mais sobre o modelo de Poisson será apresentado na Seção 3.2.4.

A Figura 3.4 ilustra um ajuste de uma distribuição de *Poisson* para dados gerados sinteticamente segundo tal distribuição. Através do histograma mostrado no primeiro gráfico, as amostras geradas formam as barras verticais e a curva ajustada representa uma distribuição de *Poisson* com parâmetro de média $\lambda = 100$. O segundo gráfico corresponde a CDF empírica (ECDF) para o mesmo exemplo.

Os testes de ajuste de distribuição utilizados na metodologia são os testes do Qui-Quadrado e de Kolmogorov-Smirnov, descritos na Seção 2.4.

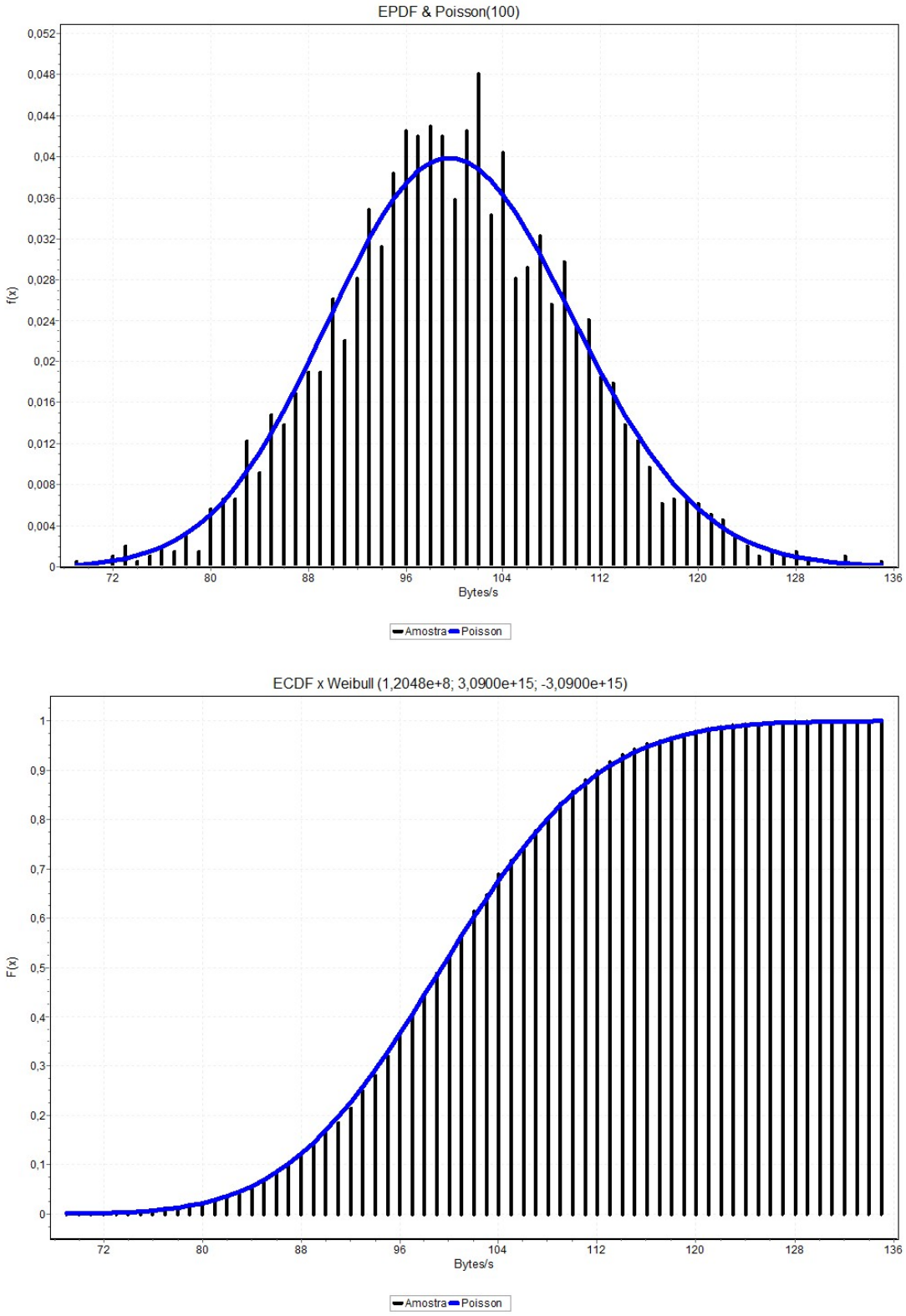


Figura 3.4: Exemplo de Ajuste de Distribuição – $Poisson(\lambda = 100)$

Autocorrelação

Também nesta etapa são gerados os gráficos do coeficiente de autocorrelação, a fim de identificar se a série temporal se encontra, ou não, em estado estacionário, além de auxiliar na estimação dos parâmetros para os modelos de previsão utilizados e identificar a presença de sazonalidade.

A autocorrelação utilizada é baseada na implementação encontrada no pacote *stats* do *software* R versão 2.15 [46], que se baseia em [47].

Para melhor compreensão do coeficiente de autocorrelação, tome $X(t)$ como um processo estocástico e t_1 e t_2 como diferentes instantes no tempo [10].

A média do processo $X(t)$ é dada por:

$$\mu_X(t) = E[X(t)] = \int_{-\infty}^{\infty} x f_{X(t)}(x) dx \tag{3.1}$$

A autocorrelação do processo $X(t)$ é definida como:

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{X(t_1),X(t_2)}(x_1, x_2) dx_1 dx_2 \tag{3.2}$$

De modo geral, a autocovariância é definida como:

$$C_X(t_1, t_2) = E[X(t_1) - \mu_X(t_1)(X(t_2) - \mu_X(t_2)] \tag{3.3}$$

O coeficiente de autocorrelação é dado pela equação abaixo:

$$\rho_X(t_1, t_2) = \frac{C_X(t_1, t_2)}{\sqrt{C_X(t_1, t_1)C_X(t_2, t_2)}} \tag{3.4}$$

Supondo um processo estacionário em sentido amplo, temos que o valor esperado do processo não depende do tempo [10], ou seja,

$$E[X(t)] = \mu \tag{3.5}$$

Considerando $\tau = |t_2 - t_1|$, o valor esperado $E[X(t)X(t + \tau)]$ é tido como a função de autocorrelação [10] e é descrito como abaixo:

$$R_X(\tau) = E[X(t)X(t + \tau)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{X(t)X(t+\tau)}(x_1, x_2) dx_1 dx_2 \tag{3.6}$$

Já a autocovariância [10] de um processo é denotada como:

$$COV(t, t+\tau) = E[(X(t) - \mu)(X(t+\tau) - \mu)] \tag{3.7}$$

onde μ é a média estatística de X (independente de t).

A fórmula que define o coeficiente de autocorrelação é apresentada na Equação (3.8) abaixo e pode ser encontrada em [8, 24, 47, 48].

$$\rho_{XX}(\tau) = \frac{E[(X(t) - \mu)(X(t+\tau) - \mu)]}{\sigma_x^2} = \frac{COV(t, t+\tau)}{\sigma_x^2} \tag{3.8}$$

Nos casos extremos tem-se:

- Quando $\tau = 0$, a amostra está totalmente correlacionada com ela mesma (mesmo instante);
- Quando $\tau \to \infty$, as correlações se aproximam de zero à medida que a distância entre as amostras no tempo aumenta.

A função de autocorrelação é utilizada como forma de modelar as dependências existentes entre as amostras [24].

Devido ao fato de se utilizar dados empíricos, é necessário utilizar uma estimativa para o cálculo do coeficiente de autocorrelação. Essa estimativa pode ser vista na equação abaixo [8]:

$$\widehat{\rho}_{xx}(\tau) = \frac{c_x(\tau)}{c_x(0)} \quad onde \quad c_x(\tau) = \frac{\sum_{i=1}^{N-k} [x(i) - \overline{x}][x(i+k) - \overline{x})]}{N-k} \tag{3.9}$$

Geralmente, tem-se valores de autocorrelação menores à medida que se aumenta a defasagem (*lag*) τ, o que indica que, caso exista dependência temporal entre as amostras, esta dependência vai diminuindo com o passar do tempo. Em outras palavras, amostras próximas com valor de autocorrelação maior que amostras mais afastadas umas das outras indicam maior dependência temporal.

No gráfico de autocorrelação são plotados os limites para se considerar um valor de autocorrelação como sendo zero, com $\pm 2\sigma$ de intervalo de confiança e 95% de significância, como pode ser visto na Figura 3.5.

Valores positivos de autocorrelação indicam que as amostras tendem a se manter em um mesmo nível, acima ou abaixo da média do processo. Ou seja, dado um valor de *lag*, os valores da série ao longo do tempo tendem a se manter acima ou abaixo da média.

Em contrapartida, valores negativos de autocorrelação indicam que as amostras tendem a oscilar em torno da média. Se uma amostra num dado instante t está

acima da média, a amostra em $t + \tau$ terá uma forte tendência de estar abaixo do valor da média, e vice-versa.

Uma analogia pode ser feita para melhor compreensão dos valores negativos de autocorrelação: No caso de um funcionário de banco conseguir realizar um atendimento rápido, o próximo atendimento tem grande chance de ser realizado com mais calma, podendo demorar mais. Por outro lado, se o funcionário faz um atendimento que toma muito tempo, o próximo tende a ser agilizado para que ele possa voltar a sua rotina de horário normal. Ou seja, é como se o tempo de atendimento oscilasse em torno de uma média.

Em Redes de Computadores pode-se dizer que este comportamento se assemelha com as condições de congestionamento, quando há um impacto negativo na taxa de transmissão dos pacotes, com um posterior aumento nessa taxa quando um enlace não congestionado é alcançado.

Outro ponto que pode ser observado no gráfico do coeficiente de autocorrelação é a propriedade de sazonalidade, que se caracteriza pela repetição de padrões comportamentais nos valores de autocorrelação. Como pode ser visto na Figura 3.5, existe um ciclo de 7 *lags*, no qual os valores apresentam comportamentos recorrentes.

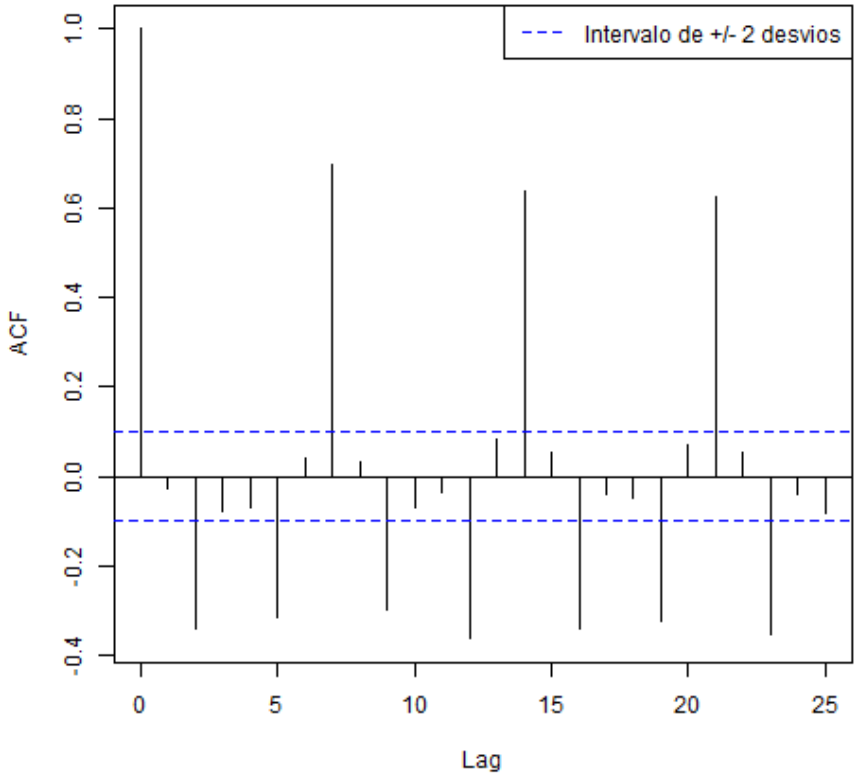


Figura 3.5: Gráfico de autocorrelação

3.2.4 Escolha de Modelos

Um dos principais componentes de impacto em avaliação de desempenho em Redes de Computadores é o modelo de tráfego a ser utilizado a fim de representar um comportamento real [49]. Os modelos de tráfego devem ser precisos ao ponto de capturar as características estatísticas do tráfego atual, para que possa ser preciso o suficiente para as tomadas de decisão. Vários modelos são apresentados na literatura, sendo os mais conhecidos listados abaixo [49]:

- **Modelo de Poisson:** É um dos modelos mais antigos utilizado nas redes de telefonia, conhecido por ser um modelo simples onde o intervalo entre chegadas segue uma distribuição exponencial com parâmetro λ, ou que o número de chegadas em um intervalo de tempo segue uma distribuição de Poisson com taxa λ. Assim, o Modelo de Poisson é expresso como abaixo:

$$P[N_n \leq k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \tag{3.10}$$

- **Modelo de Pareto:** O Modelo de Pareto, tendo como base tal distribuição, apresenta dois parâmetros: α relativo à forma e β como parâmetro de localização. Os parâmetros e a Função de Distribuição Cumulativa estão descritos abaixo:

$$P[X \leq x] = 1 - \left(\frac{\beta}{x}\right)^{\alpha} \quad ; \quad \beta \leq x \tag{3.11}$$

  Este modelo é utilizado para geração de tráfego *On-Off*, no qual o tráfego é gerado apenas em períodos *On*

  O Modelo de Pareto faz parte da classe de distribuições que seguem as Leis de Potência, também conhecidas como distribuições de cauda pesada, que apresentam a forma como na equação abaixo:

$$P[X > x] \sim x^{-\alpha}, \quad com \quad x \to \infty, \quad 0 < \alpha < 2 \tag{3.12}$$

  Distribuições de cauda pesada apresentam como principal característica a média e a variância infinitas para certos valores de parâmetros.

Outro fator relevante na escolha de modelos para previsão utilizando séries temporais, está relacionado a qual modelo de previsão será utilizado. A literatura oferece uma variedade de técnicas que se propõem a trabalhar neste escopo [11, 13, 34, 50].

Para este trabalho foram escolhidos modelos baseados em regressão, devido ao fato destes serem mais utilizados para previsão de longo prazo. Outras abordagens,

como conhecimento de máquina, geralmente são voltadas para previsão de curto prazo, tendo como objetivos a detecção de anomalias [51–53], alocação dinâmica de recursos, entre outros [13], diferentemente dos objetivos que são propostos aqui neste trabalho, como visto no Capítulo 1.

Os modelos para previsão baseados em séries temporais são listados abaixo.

- **Modelo Autorregressivo – AR(p):** Este modelo considera o modo como os valores em um determinado período estão relacionados com valores em um período prévio.

  O modelo é composto por um termo constante que recebe a influência de outros termos de "perturbação". A Equação (3.13) mostra a formação de um modelo autorregressivo.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t \tag{3.13}$$

  Na Equação (3.13), y_t é o valor de uma amostra gerada num instante t, μ é a média constante, γ_i é o i-ésimo coeficiente de autorregressão que será multiplicado pelo valor da amostra no instante $t - i$, e ϵ_t é um ruído aleatório. A presença do componente de ruído é devido ao fato de o modelo em questão ser um filtro linear, o qual supõe que a série temporal seja gerada através de um filtro cuja entrada é o ruído aleatório [54]. O mesmo acontece para os modelos que se seguem.

- **Modelo de Média Móvel – MA(q):** Este modelo considera a relação entre uma variável e resíduos em períodos prévios. A Equação (3.14) mostra a formação de um modelo de média móvel.

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{3.14}$$

- **Modelo Autorregressivo Integrado e de Média Móvel – ARIMA(p,d,q):** Este modelo é baseado no modelo $ARMA(p, q)$, que por sua vez é a combinação dos modelos autorregressivo e de média móvel. A Equação (3.15) descreve o modelo $ARMA(p, q)$.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{3.15}$$

  Porém, o $ARMA(p, q)$ requer uma série temporal em estado estacionário, o que nem sempre é realidade. Para isso, aplicam-se operações de diferença nas amostras, a fim de remover componentes de tendência.

Sendo assim, o modelo ARIMA agrega ao modelo ARMA um parâmetro d relativo à quantidade de operações de diferença necessárias para tornar uma série possível de ser utilizada, sem influência de componentes de tendência.

Para estimar o parâmetro q, utilizam-se os valores de coeficiente de autocorrelação, pelos quais q é estimado ser o valor de *lag* a partir do qual os valores de autocorrelação podem ser considerados zero, ou seja, que ficam dentro do limite de 95% de nível de confiança [9].

Para estimar o parâmetro p, utilizam-se os valores de coeficiente de autocorrelação parcial, pelos quais p é estimado ser o último valor de *lag* maior que o limite de 95% de confiança.

- **Modelo Autorregressivo Integrado e de Média Móvel Sazonal – SARIMA(p,d,q)(P,D,Q):** O Modelo SARIMA segue as mesmas características apresentadas para o modelo ARIMA, com o diferencial de ter adicionalmente uma parte sazonal, com mais três parâmetros: de autorregressão, diferença e média móvel em termos de sazonalidade, respectivamente P, D e Q.

  As operações de diferença para a parte sazonal são análogas à operação definida na Equação 2.7, sendo que, em vez de calcular apenas a diferença com 1 (uma) defasagem, desta vez a operação é realizada de acordo com o período de sazonalidade. Ou seja, se a sazonalidade apresenta um período semanal (7 dias) por exemplo, então a operação de diferença em termos sazonais será feita conforme abaixo:

$$d_1 = X(t) - X(t-7).$$

Genericamente, para uma sazonalidade com período s, tem-se as operações de diferença definidas conforme a Equação 3.16.

$$d_n = d(d_{n-s}) \tag{3.16}$$

A Equação 3.17 mostra o modelo SARIMA [55].

$$SARIMA = ARIMA(p,d,q)(P,D,Q)_s \tag{3.17}$$

onde s é o período de sazonalidade.

Isso resulta em:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{P} \Gamma_i y_{t-i} + \sum_{i=1}^{Q} \Theta_i \epsilon_{t-i} + \epsilon_t \tag{3.18}$$

3.2.5 Previsão Segundo o Modelo Escolhido

A etapa de previsão consiste em submeter os modelos encontrados na etapa anterior a procedimentos que permitam a geração de valores futuros da série, considerando possíveis desvios na previsão. Assim, é necessário fazer a estimação dos parâmetros do modelo escolhido, ajustando-os a fim de estudar diferentes comportamentos, até que seja alcançada uma configuração com resultado satisfatório. Um resultado pode ser considerado satisfatório se os valores previstos, considerando os possíveis desvios, ficam dentro de uma margem de erro aceitável.

Segundo a metodologia em [8], três passos são utilizados para se definir os parâmetros do modelo de previsão ARIMA serão escolhidos. Os passos são:

1. **Identificação:** Esta fase consiste em determinar os valores p, d e q para o modelo ARIMA correspondente. Para isto, utilizam-se os gráficos de autocorrelação (ACF) e autocorrelação parcial (PACF). Nestes gráficos, os valores de autocorrelação considerados acima de zero serão utilizados como estimativa inicial para os parâmetros mencionados.

   Para identificar o parâmetro p utiliza-se o gráfico de ACF. Para identificar o parâmetro q utiliza-se o gráfico de PACF.

   O parâmetro d trata da quantidade de operações de diferença, necessárias para tornar a série estacionária. O teste de Dickley-Fuller Aumentado auxilia nesta fase, respondendo se a série temporal é estacionária, ou não.

   Os processos mencionados também são aplicados para identificar os parâmetros do modelo que contempla sazonalidade, o SARIMA, fazendo as mesmas considerações para os valores de autocorrelação a partir da defasagem sazonal, conforme descrito na Seção 3.2.4.

2. **Estimação:** Esta fase estima os valores para os coeficientes no modelo em questão, conforme visto na Seção 3.2.4. Com auxílio do *software* Gretl [56], foi feita a estimação utilizando o Filtro de Kalman, atrávés da Máxima Verossimilhança Exata. Mais detalhes sobre os métodos de estimação podem ser encontrados nas referências [8, 9].

3. **Verificação:** Nesta fase, compara-se os valores previstos com valores reais, a fim de calcular o desvio de previsão e o intervalo de confiança. Também é feito o teste de Ljung-Box [57] para verificar a existência de autocorrelação no modelo estimado. O teste contempla a hipótese nula de não existência de autocorrelação, sendo esta aceita caso seja obtido um p-valor maior do que o nível de significância, geralmente configurado para 0,05. Caso o teste não seja satisfeito, é necessário voltar a fase de identificação.

Estes passos serão vistos com exemplo de uso durante o estudo de caso apresentado no Capítulo 4.

A etapa de previsão pode sofrer influência de como a coleta de dados foi realizada. Dependendo da granularidade das medidas, um determinado tipo de previsão poderá ser feita. De acordo com [58], as previsões são definidas como descrito abaixo:

- **Tempo real:** relativas a períodos de poucos minutos, geralmente utilizado em sistemas que requerem uma previsão em tempo real (dentro de uma janela de tempo curta), de forma *online*, como detecção de anomalias, por exemplo;
- **Curto prazo:** relativas a períodos de algumas horas, utilizado para escalonamento dinâmico de recursos;
- **Médio prazo:** relativas a períodos de alguns dias, utilizado para planejamento de alocação de recursos;
- **Longo prazo:** relativas a períodos de alguns meses a anos, utilizado para provisionamento de recursos e decisões estratégicas de mudança de tecnologias, com preocupação na alocação de recursos financeiros, por exemplo.

A proposta desta dissertação se enquadra na previsão de longo prazo, também podendo ser estendida para previsões de médio prazo.

Para verificação da previsão, os dados coletados foram utilizados da seguinte forma. De acordo com o período que se propôs prever, uma quantidade correspondente aos dados coletados foi separada para comparação com os resultados da previsão. Deste modo, se é procurado um período de 6 meses de previsão, então os 6 últimos meses de dados coletados são separados dos dados utilizados para definição do modelo de previsão e são utilizados posteriormente para comparação entre os dados reais e os dados previstos.

3.2.6 Decisão da ação a ser executada

Com base nos resultados obtidos pelas saídas das análises de previsão, juntamente com a caracterização do tráfego, é possível decidir sobre quais ações devem ser executadas.

Duas formas das ações serem aplicadas podem ser destacadas:

- **Ações Automáticas:** Esta abordagem torna o sistema de previsão mais completo, no sentido de atuar pró-ativamente, sem a necessidade de intervenção humana. Entretanto, é razoável que o sistema permita que ações críticas devam ser confirmadas por administradores.

- **Alertas de Ações:** Esta abordagem é mais passiva, já que não aplica nenhum tipo de regra baseada nas decisões extraídas do sistema de previsão. Porém, esta é a que mais se adapta a ambientes críticos onde uma ação automatizada seria muito arriscada, sendo de extrema necessidade uma intervenção humana.

Independente de qual abordagem seja utilizada, esta etapa basicamente responde se um determinado enlace deve ou não ser revisto em termos de sua capacidade de transferência de dados. Através dos resultados obtidos durante a análise, decide-se, com um determinado nível de precisão, qual enlace deve ser atualizado e em que escala essa atualização deve ocorrer, isto é, em quanto a capacidade deve ser aumentada. Assim, se é previsto que um enlace irá ultrapassar a capacidade máxima de transferência, os processos que envolvem a atualização do mesmo já podem ser iniciados de forma antecipada.

Diversas técnicas mais sofisticadas podem ser aplicadas para auxiliar no processo de decisão. O aprofundamento destas técnicas foge ao escopo deste trabalho. Porém, para dar um exemplo para futuras diretrizes, pode ser possível aplicar a tomada de decisão com base em ontologias (conhecimento), onde previsões já realizadas podem ser utilizadas para a previsão de uma rede específica. Esta técnica poderia inclusive auxiliar no caso de uma nova rede sendo construída. Como neste cenário não há um histórico de medições que possa ser aplicado aos modelos de previsão, faz-se necessário que outros meios sejam levados em consideração, podendo ser o caso da aplicação de estatísticas de outras redes que já vêm sendo estudadas e que possuem um histórico de medidas registrado. Contudo, é preciso que sejam feitos mais estudos para verificar a viabilidade da aplicação de oncologias para este propósito.

Capítulo 4

Estudo de Caso – RedeRio

Conforme visto no Capítulo 3, foi apresentada uma metodologia baseada em [11], [29] e [30] de forma a auxiliar no processo de caracterização do tráfego em redes, bem como na previsão do tráfego.

Neste capítulo serão apresentados os resultados das análises feitas sobre o estudo de caso do enlace de conexão internacional da RedeRio de Computadores/FAPERJ [15] através da Level 3 Communications [17].

O modelo de previsão escolhido foi o Autorregressivo Integrado e de Média Móvel (ARIMA) e sua variação sazonal, SARIMA. A escolha desse modelo se deu pelo fato deste ser mais preciso para previsões de longo prazo. Apesar da complexidade computacional ser da ordem de N^2 (N é o número de amostras), o modelo é mais adequado para previsões cujo objetivo é o planejamento de ações futuras, não sendo tão relevante a questão da complexidade computacional, como comparado em [34].

Em outras palavras, não é restritivo ter um modelo computacionalmente complexo para fins de uma análise realizada de maneira *off-line*, em termos da existência de flexibilidade de tempo para escolha e execução do modelo. Isso se difere em outros casos, como por exemplo, na realização de uma previsão de curto prazo para alocação dinâmica de recursos ou detecção de anomalias. Nestes casos, a complexidade computacional é um fator relevante.

Alguns programas de apoio foram desenvolvidos para extrair as medidas dos arquivos de registro (*log*) gerados pelo programa de coleta. Esses programas também proporcionaram uma automatização de algumas etapas do processo de geração de gráficos, como mencionado na Seção 3.2.2.

Com os dados já separados e tratados, colocados em um formato compreensível e utilizável, inicia-se o processo de caracterização desses. Este processo compreende a geração de estatísticas descritivas (média, desvio padrão, autocorrelação, entre outros), bem como os gráficos de frequência relativa, representados por histogramas. Analisa-se assim, qual é a frequência de utilização de uma determinada taxa de transferência. Por consequência do histograma, também é obtida a Função de Pro-

babilidade Cumulativa Empírica (ECDF). Essa caracterização permite uma melhor compreensão sobre o comportamento estatístico do tráfego nos enlaces, permitindo a realização de inferências sobre esse comportamento.

Através das ECDF, estima-se qual a probabilidade de uma dada interface apresentar uma taxa de transferência menor ou igual a um valor específico. Tal resultado permite a verificação de um possível esgotamento de um enlace contratado, ajudando na tomada de ações preventivas para que tal fato não ocorra. Essas ações sugerem redução de custos, bem como uma melhor qualidade do serviço prestado.

Vale ressaltar que, através de uma análise mais próxima dos dados coletados, levando em consideração diferentes contextos, infere-se que nem sempre uma alta probabilidade de baixos valores de tráfego, ou até mesmo de zeros, significa necessariamente que o enlace é ruim ou que tem muito tempo em que está fora de operação. Isto porque, condições de baixa utilização, ou falta de monitoramento da interface, também apresentam o mesmo comportamento. Assim, é preciso ter cuidado ao analisar os dados obtidos, a fim de ter uma interpretação correta dos mesmos, evitando assim uma possível tomada de decisão equivocada.

Após isto, o processo de ajuste de distribuição é feito a fim de aproximar uma distribuição conhecida aos dados coletados. Para isso, testes estatísticos de ajuste são feitos, conforme mencionado na Seção 2.4.

Em seguida, é feita a segunda parte do trabalho sobre a previsão do tráfego. Com os parâmetros do modelo ARIMA (ou SARIMA) já definidos, é possível gerar os gráficos contendo as previsões para a interface escolhida.

Além disso, de forma a explorar diferentes granularidades (escalas de tempo), também são apresentados testes utilizando valores calculados a partir das amostras coletadas, para validação da metodologia apresentada nesta dissertação.

Testes com amostras sintéticas também foram feitos para verificar a conformidade das previsões segundo os modelos escolhidos. Através dos testes foi possível perceber a necessidade de se ter amostras suficientes para estimação dos modelos, devido aos intervalos de confiança obtidos terem sido muito amplos.

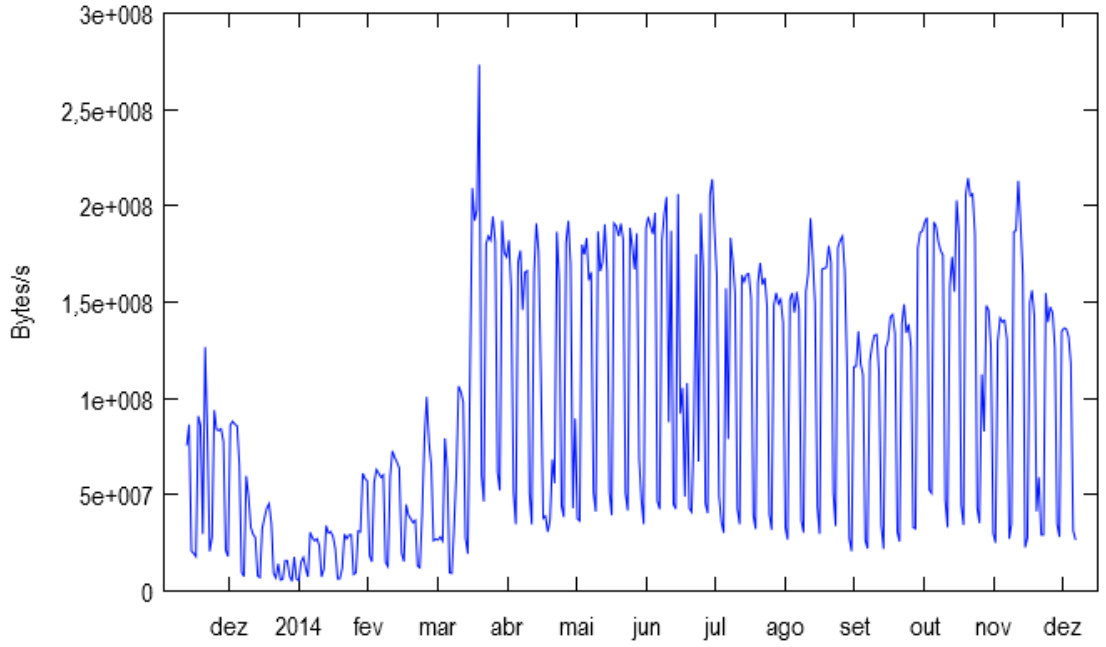
4.1 Cenário

Como dito na Seção 1.3, são utilizados neste estudo os dados coletados a partir dos enlaces da RedeRio. O foco se deu em um dos enlaces de maior importância no que tange à conexão internacional para as instituições associadas, atualmente acordado com a Level 3 Communications. Sendo assim, é importante trazer os resultados da análise desta interface como exemplo prático deste trabalho. A Interface Level 3 disponibiliza um largura de banda de 3 Gbps.

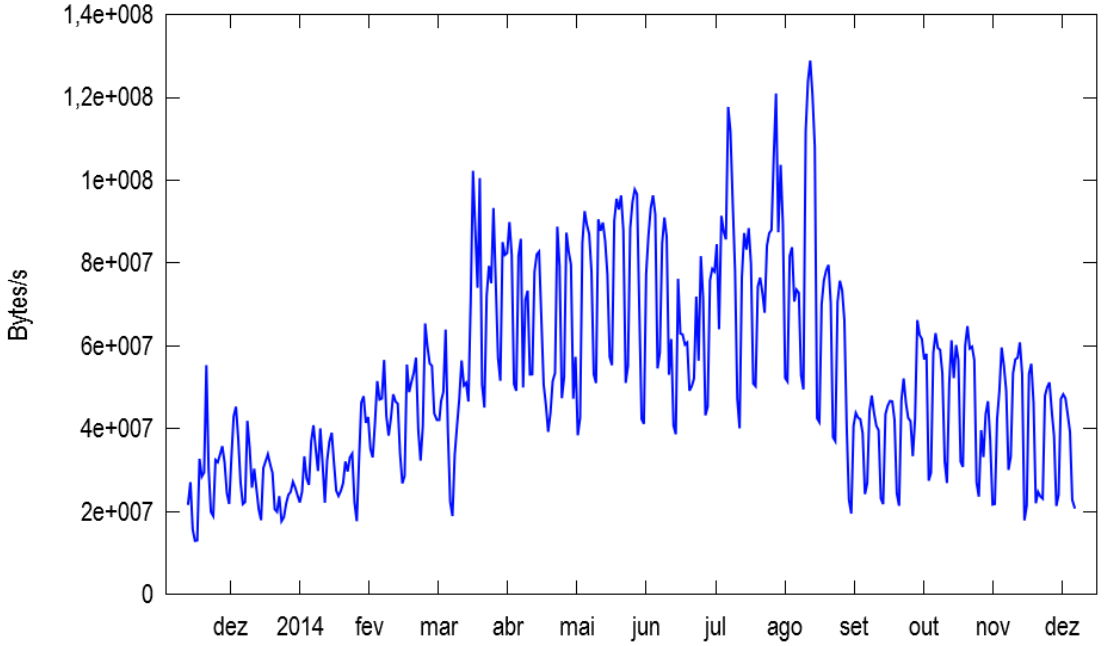
Após as etapas de coleta, filtragem e extração de medidas terem terminado,

obtém-se os dados para serem utilizados nas próximas etapas da metodologia. A Figura 4.1 ilustra o tráfego de entrada e de saída coletados com granularidade de um dia durante o período de 13/11/2013 a 07/12/2014, totalizando 390 amostras válidas.

Nos gráficos da Figura 4.1, no período anterior ao mês de março, o comportamento reduzido foi devido a mudanças de infraestrutura na rede, bem como nos



(a) Tráfego de Entrada



(b) Tráfego de Saída

Figura 4.1: Tráfego da Interface Level 3

processos de coleta, que ainda passavam por ajustes. Contudo, decidiu-se manter tais amostras a fim de acompanhar como os modelos de previsão se ajustariam.

4.2 Caracterização e Ajuste de Distribuição

Apresenta-se inicialmente os resultados da caracterização dos dados com base nos histogramas e ECDFs para o tráfego de entrada, que podem ser visualizadas na Figura 4.2.

Cada barra do histograma é o correspondente à frequência relativa das amostras dentro de um intervalo de 10 MBytes/s. Ou seja, a quantidade de amostras que apresentam valores no intervalo entre 0 e 10 MBytes/s (inclusive), divido pelo total de amostras, corresponde a primeira barra. A segunda barra corresponde ao mesmo cálculo para o intervalo entre 10 MBytes/s e 20 MBytes/s e assim sucessivamente para os outros intervalos subsequentes.

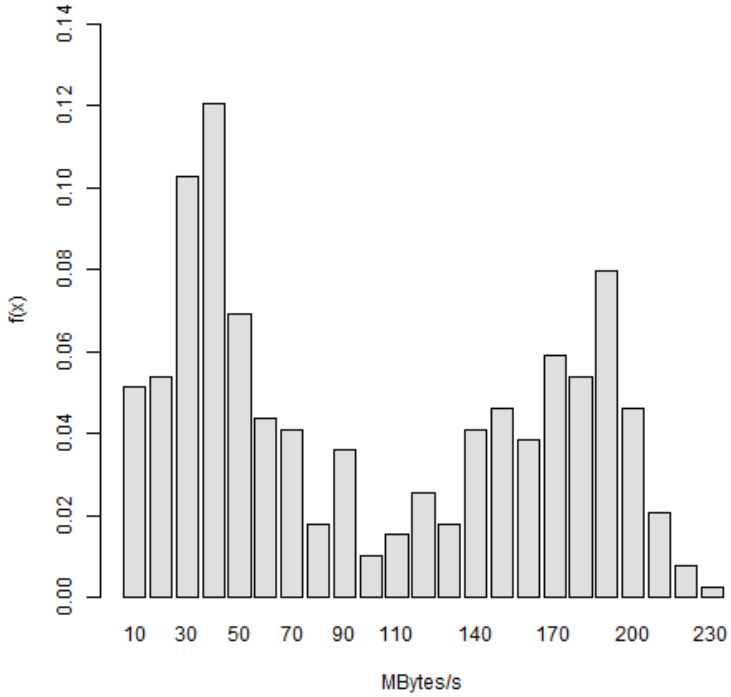
Através dessas funções pode-se inferir que:

- A vazão do tráfego de *download* se acumula basicamente em duas classes de tráfego: (i) uma classe de menor vazão, entre 0 e 90 MB/s; (ii) uma classe de maior vazão, entre 110 MB/s e 230MB/s;
- A probabilidade da vazão do tráfego de *download* exceder 122 MB/s ($\approx$ 0,976 Gbps) é de 40%;
- Pelo menos 97% do tráfego de *download* apresenta vazão máxima de 200 MB/s ($\approx$ 1,6 Gbps).

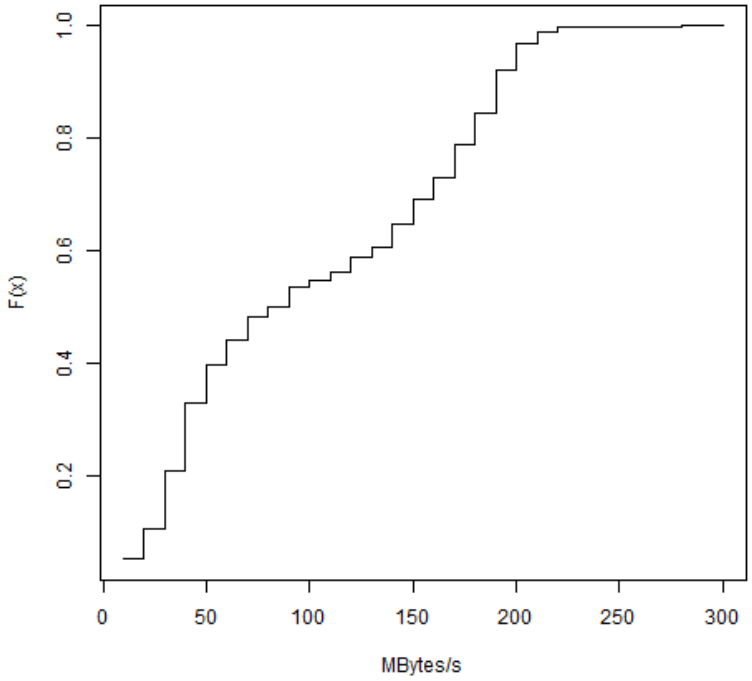
Esses resultados permitem chegar a conclusões importantes, como qual a probabilidade do tráfego em um determinado enlace exceder a capacidade máxima disponível. Isso também auxilia na etapa de tomada de decisão, junto com os resultados da previsão, para uma melhor avaliação do uso do enlace.

Uma análise em relação ao comportamento do tráfego ao longo dos dias da semana também foi realizada. A média do tráfego para cada dia da semana pode ser vista na Figura 4.3. É razoável perceber que, durante a semana, o tráfego exibe sazonalidade devido aos finais de semana e que o tráfego de maior volume médio se encontra nas quartas-feiras. Os outros dias, com exceção do final de semana, têm um comportamento semelhante. Com isso, já há indícios de que um modelo que compreenda a componente de sazonalidade deverá ser utilizado.

Dando sequência à metodologia, os testes de ajuste (*fitting*) de distribuição foram feitos utilizando diversos tipos de distribuição dentre as mais usuais, a fim de identificar qual padrão estatístico mais se assemelha ao padrão de geração dos dados para a interface em estudo, com auxílio do programa *EasyFit* [59].



(a) Histograma



(b) ECDF

Figura 4.2: Caracterização do Tráfego de Entrada da Interface Level 3

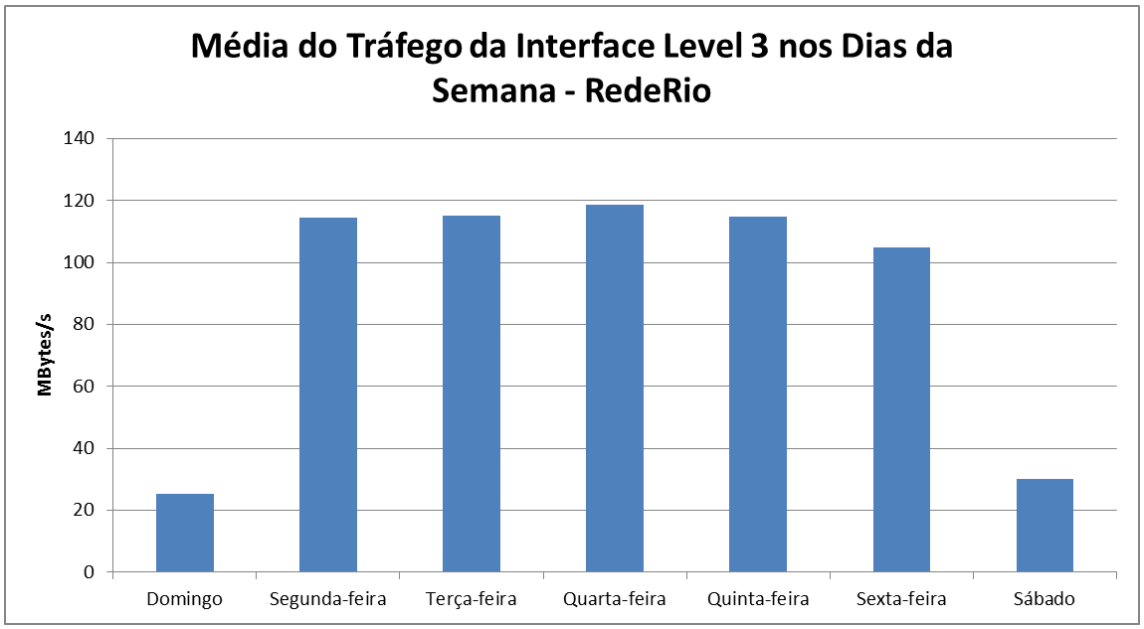


Figura 4.3: Média do tráfego da Interface Level 3 nos dias da semana

Para comparar dentre as distribuições mais conhecidas, são realizados os testes de Kolmogorov-Smirnov e Qui-Quadrado [22], como já explorados na Seção 3.2.3. Os resultados dos testes de ajuste de distribuição para o tráfego de entrada podem ser verificados na Tabela 4.1.

Analisando os p-valores da tabela, nota-se que, tanto para o teste do Qui-Quadrado, quanto para o Kolmogorov-Smirnov, os ajustes encontrados não apresentam p-valores significativos, ou seja, maiores que 0,05.

Assim, o ajuste de distribuição para os dados em análise não apresentou resultados satisfatórios que possam afirmar a aproximação das amostras a alguma distribuição da literatura. A distribuição que mais se aproximou foi a Johnson SB. A distribuição Johnson SB também pode ser compreendida como uma Lognormal de 4 parâmetros (dois parâmetros de forma, um de escala e um de localização), servindo para caracterizar taxas ou variações limitadas a um intervalo definido [60]. Supõe-se que essa aproximação é dada pelo fato do resultado de duas classes de tráfego estarem presentes nos dados coletados. A Figura 4.4 mostra a aderência da

Tabela 4.1: Resultado dos testes de Kolmogorov-Smirnov e Qui-Quadrado para o Tráfego de Entrada na Interface Level 3

Distribuição	**Estatística K-S**	**p-valor**	**Estatística Qui-Quadrado**	**p-valor**
Johnson SB	0,09813	0,00101	–	–
Beta	0,10317	4,5552E-4	74,4	6,5037E-13
Weibull	0,13282	1,8364E-6	93,246	1,1102E-16
Cauchy	0,21178	8,9495E-16	137,64	0

Tabela 4.2: Resultado dos testes de Kolmogorov-Smirnov e Qui-Quadrado para o Tráfego de Entrada em meses na Interface Level 3

Distribuição	**Estatística K-S**	**p-valor**	**Estatística Qui-Quadrado**	**p-valor**
Weibull	0,20111	0,64659	0,90619	0,34113
Cauchy	0,23172	0,47045	0,19469	0,65904
Beta	0,23393	0,45866	–	–
Johnson SB	0,25017	0,37711	–	–

distribuição Johnson SB às amostras utilizadas.

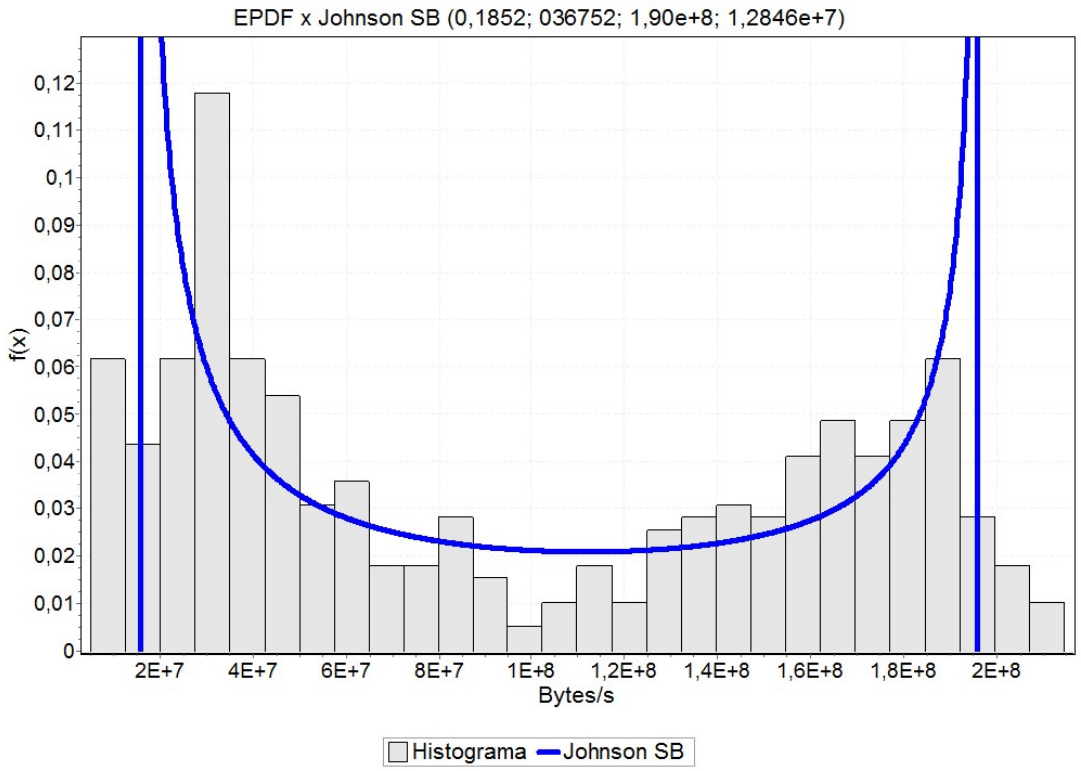
Como pode ser visto, apesar do ajuste da distribuição apresentar resultados não conclusivos em termos dos p-valores para as estatísticas encontradas, a distribuição até se aproxima do histograma e da ECDF das amostras coletadas. Contudo, a fim de buscar um resultado mais adequado, optou-se por realizar testes com outras granularidades dos dados.

Foi feito então o cálculo de médias mensais com base nos valores diários, resultando em apenas 12 amostras. Com essa granularidade de mês, obteve-se os ajustes conforme a Tabela 4.2.

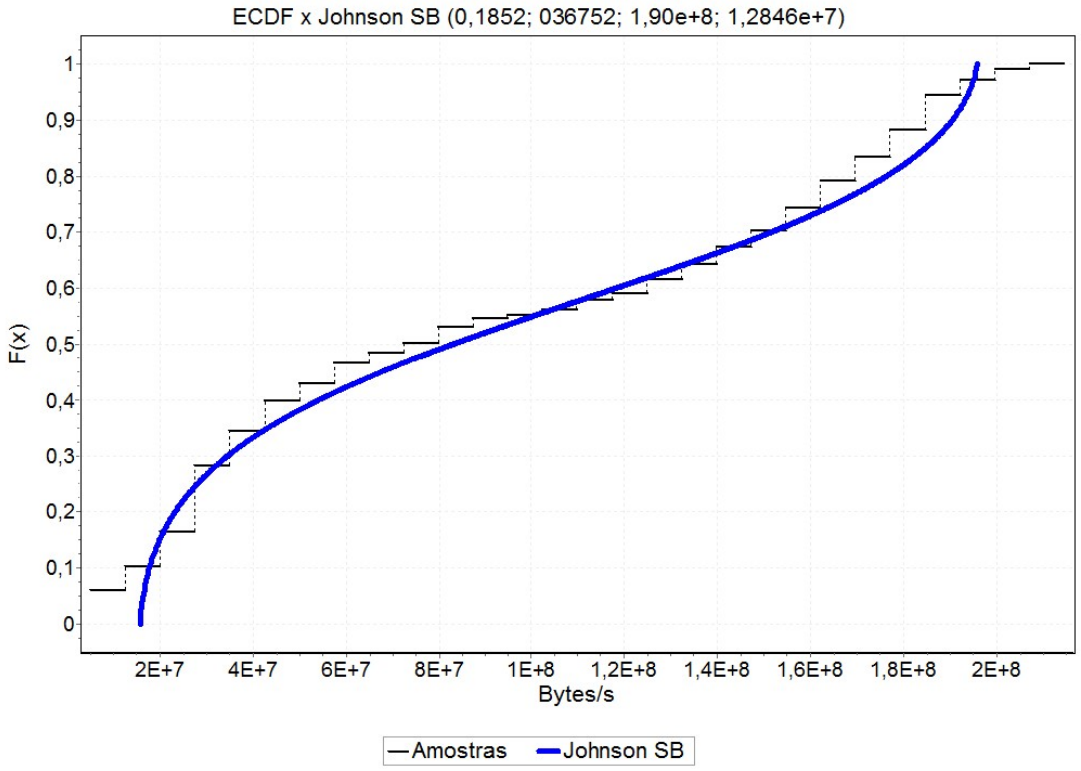
Neste caso, pelos p-valores obtidos, a distribuição Weibull foi a que melhor aderiu aos dados, segundo o teste de Kolmogorov-Smirnov, podendo servir de referência para geração de valores para um tráfego que se assemelhe ao que passa na interface da Level 3. Os parâmetros para tal distribuição são $\alpha = 1,2048 * 10^8$, $\beta = 3,0900 * 10^{15}$ e $\gamma = -3,0900 * 10^{15}$. A Figura 4.5 mostra a distribuição Weibull ajustada aos dados.

Vale perceber no exemplo mostrado que existe a necessidade intrínseca de uma quantidade maior de amostras para que os resultados possam ser mais precisos e ajustados. Para obter um bom estimador, recomenda-se ter pelo menos 50 amostras, conforme [8].

Mais a frente será abordado como que a primeira etapa de caracterização e ajuste de distribuição pode contribuir junto com a etapa seguinte de previsão do tráfego.

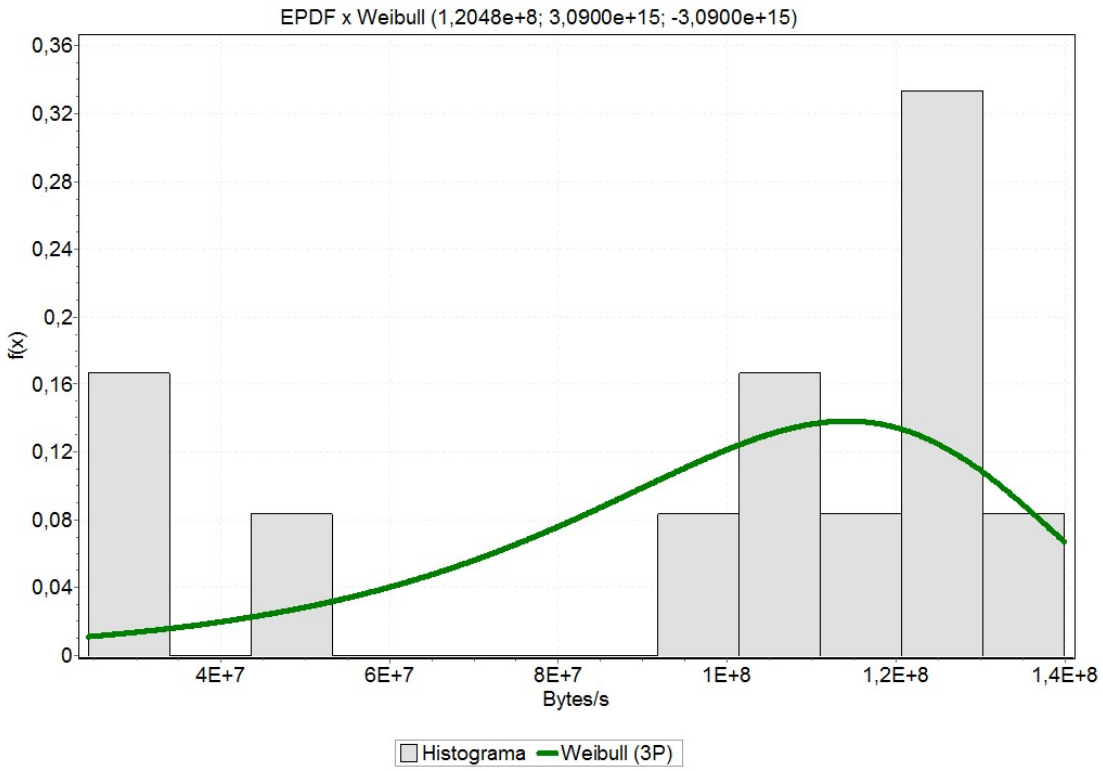


(a) Histograma Ajustado

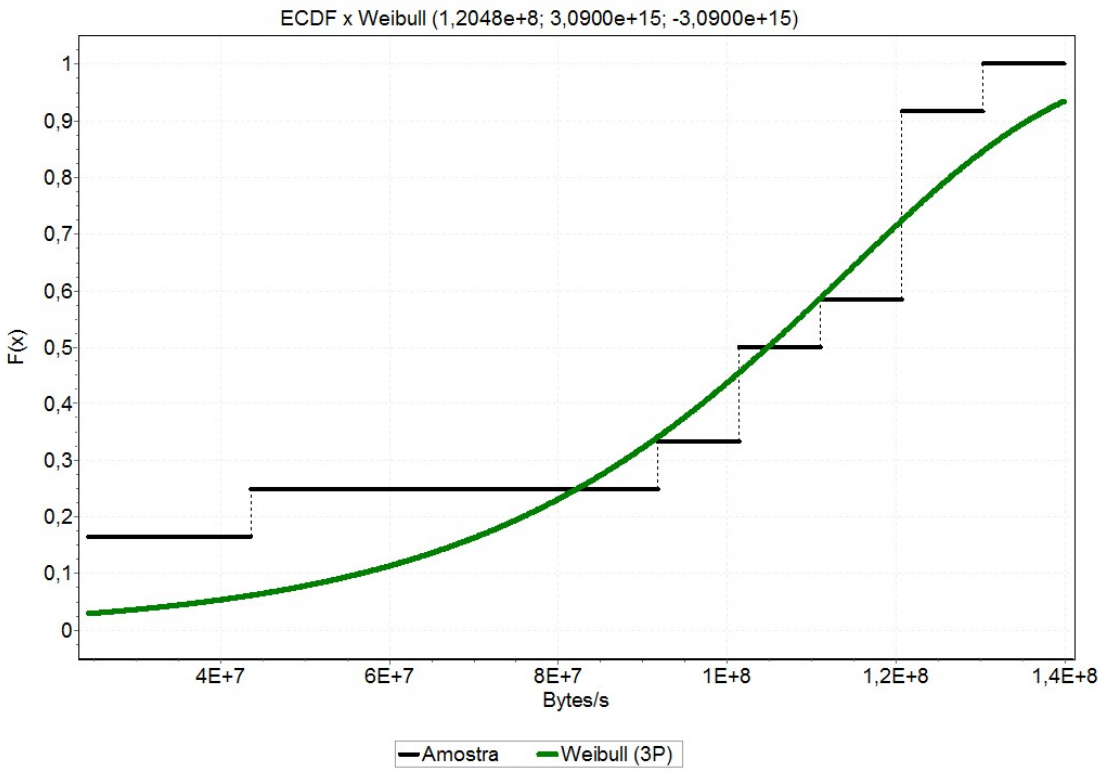


(b) ECDF Ajustado

Figura 4.4: Ajuste de Distribuição Johnson SB do Tráfego de Entrada da Interface Level 3



(a) EPDF Ajustado



(b) ECDF Ajustado

Figura 4.5: Ajuste de Distribuição Weibull do Tráfego de Entrada da Interface Level 3

4.3 Previsão de Tráfego

Na segunda etapa do trabalho, conforme a metodologia apresentada no Capítulo 3, todo o trabalho de estimação dos parâmetros dos modelos gira em torno das funções de autocorrelação (ACF) e autocorrelação parcial (PACF).

Começando pelo gráfico de autocorrelação (Figura 4.6), é possível identificar pelo menos duas características. A primeira é dada pelo fato de que os valores de autocorrelação decaem para zero lentamente (Figura 4.6a). Outra característica é o padrão de repetição nos valores de autocorrelação a cada 7 defasagens. Esses fatores revelam a presença de não estacionariedade e sazonalidade nos dados.

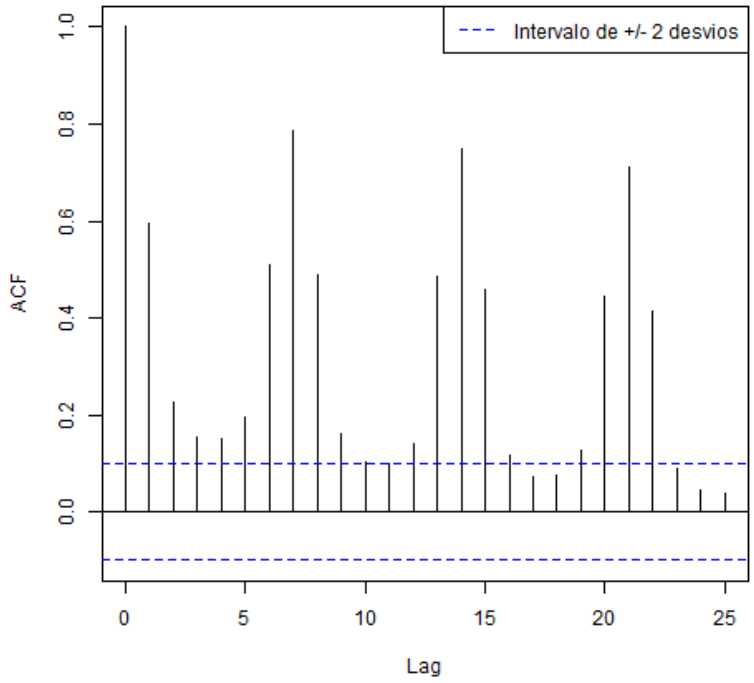
Para comprovar a existência de estacionariedade, aplica-se o teste de Dickley-Fuller Aumentado, como mencionado na Seção 2.3, constatando a não estacionariedade.

Para tornar a série estacionária, são feitas as operações de diferença até que a série passe no teste. Através do gráfico da Figura 4.7a é possível perceber que, mesmo após a série ter sido diferenciada uma vez e o teste de estacionariedade apresentar resultado satisfatório, as autocorrelações ainda apresentam valores significativos nas defasagens múltiplas de 7. Isso sugere a presença de sazonalidade, o que leva à aplicação de operações de diferença sazonal, neste caso semanal (de 7 em 7).

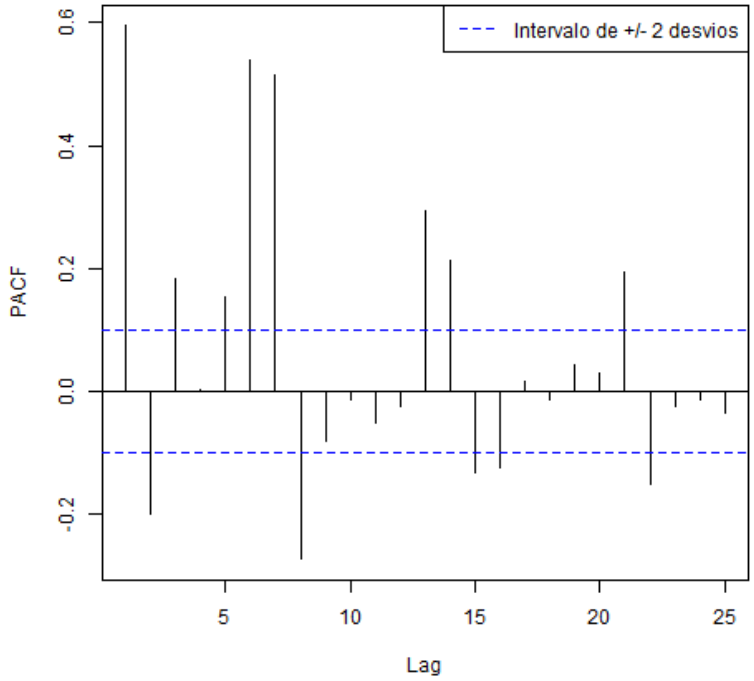
Sendo assim, aplica-se a primeira operação de diferença sazonal, tendo como resultado os valores de autocorrelação da Figura 4.8. Analisando a figura, pode-se iniciar a etapa de identificação de um modelo inicial de estudo.

Os valores das defasagens anteriores ao período de sazonalidade (7 defasagens) no gráfico de autocorrelação sugerem um valor inicial para o parâmetro de média móvel. Ao mesmo tempo, os valores de autocorrelação parcial, auxiliam na identificação dos parâmetro de autorregressão para a componente não sazonal do modelo. Da mesma forma, os valores que excedem os limites calculados a partir do período de sazonalidade fornecem informação sobre quais valores utilizar para os parâmetros da componente sazonal do modelo. Portanto, neste exemplo, pode-se utilizar como parametrização inicial $d = 1$ e $D = 1$, utilizando o modelo sazonal $SARIMA(p, 0, q)(P, 1, Q)$.

Pela Figura 4.8b, o parâmetro p de autorregressão dado pela PACF é $p = 6$. Isso porque na defasagem 6 ocorreu o último valor que excedeu o limite superior, representado pela linha tracejada no gráfico, correspondente a dois desvios ($2/\sqrt{N}$) em relação a valores que podem ser considerados zero. O parâmetro q de média móvel é sugerido pelo ACF como $q = 3$, segundo a Figura 4.8a. Para a parte sazonal, tem-se de maneira semelhante a partir do *lag* 7, os parâmetros $P = 0$ e $Q = 3$. Isso resulta no modelo inicial $SARIMA(6, 0, 3)(0, 1, 3)$. Contudo, o parâmetro de autorregressão para a parte não sazonal ficou acima do que geralmente é encontrado na literatura,

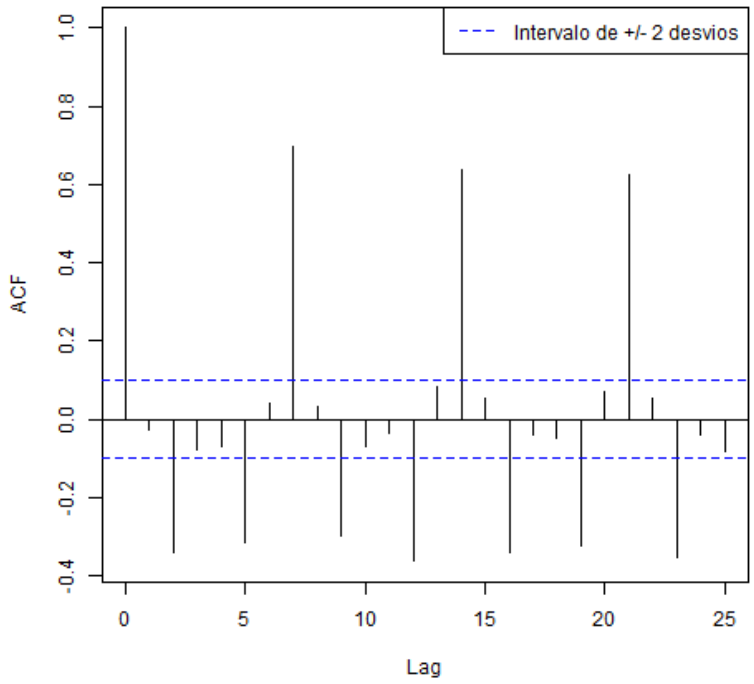


(a) ACF

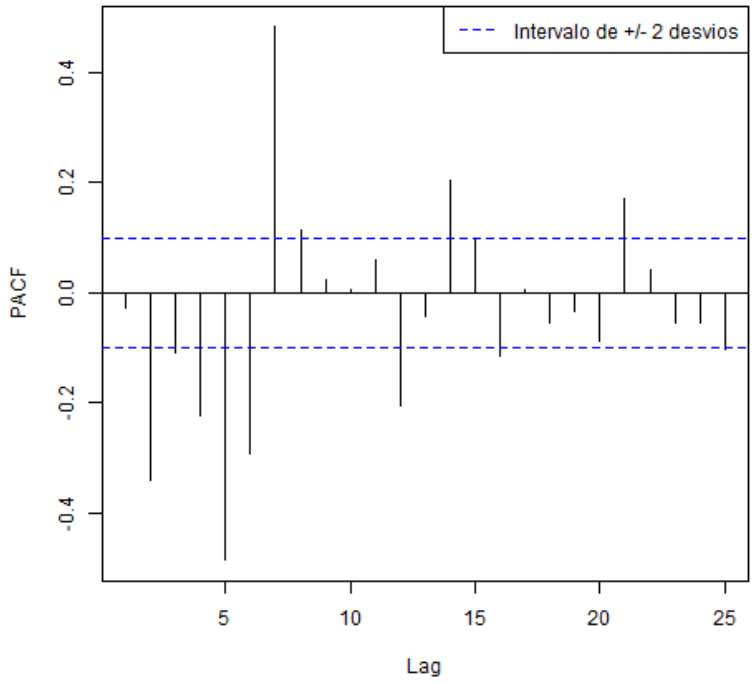


(b) PACF

Figura 4.6: Gráfico de ACF e PACF para o Tráfego de Entrada da Interface Level 3

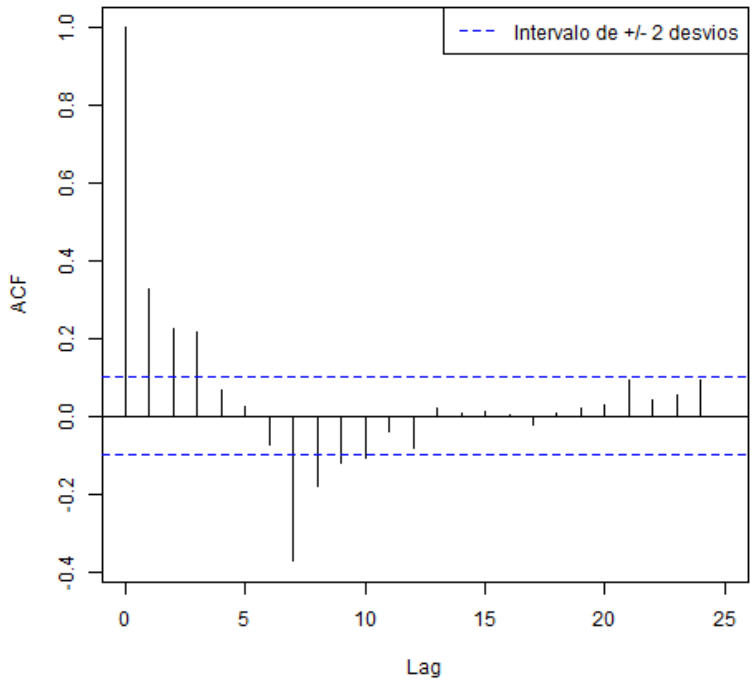


(a) ACF

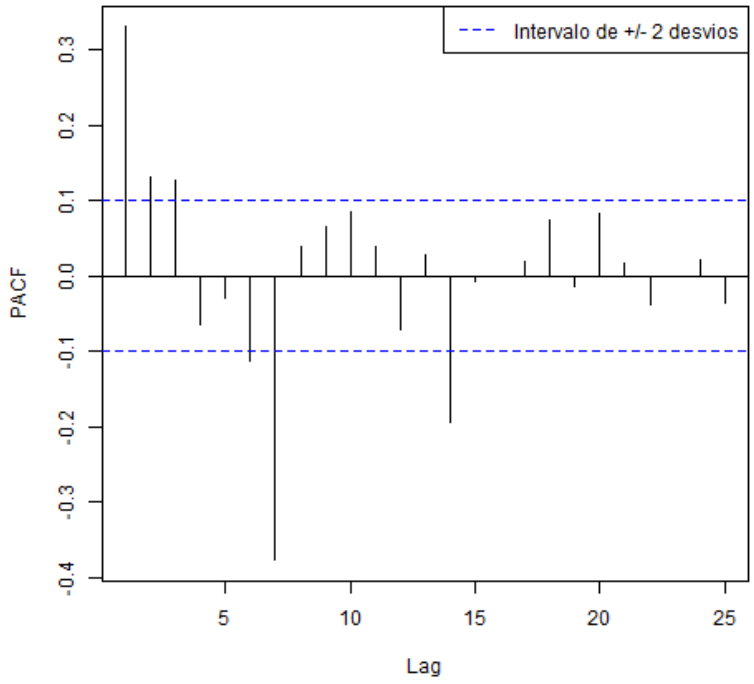


(b) PACF

Figura 4.7: Gráfico de ACF e PACF após uma diferença para o Tráfego de Entrada da Interface Level 3



(a) ACF



(b) PACF

Figura 4.8: Gráfico de ACF e PACF após uma diferença sazonal para o Tráfego de Entrada da Interface Level 3

a saber, no máximo cinco [61]. Vale ressaltar que essa redução no parâmetro de autorregressão ocorreria mesmo que não houvesse a recomendação da referência mencionada, pois a busca por um modelo mais adequado faz parte do processo, o que é conseguido através da remoção de parâmetros que apresentam menor significância. Portanto, o modelo inicial a ser utilizado será o $SARIMA(5,0,3)(0,1,3)$.

Antes de estimar o modelo, foram separadas algumas amostras a fim de serem utilizadas na etapa de sua respectiva validação. Como a granularidade (escala de tempo) das amostras está em uma amostra por dia, foram separadas 180 amostras para estimar o período de 6 meses à frente. Também foram feitos testes com uma janela de previsão menor, separando 90, 60 e 30 amostras, correspondendo a três meses, dois meses e um mês de previsão, respectivamente, como será visto mais adiante.

Utilizando o *software* Gretl [56], é possível realizar as etapas de estimação e verificação dos modelos identificados. Após a execução da estimação, o primeiro teste a ser feito é o de Ljung-Box [57] para verificar a existência de autocorrelação no modelo estimado. Através do teste sob a hipótese nula de não existência de autocorrelação, pode-se aceitar tal hipótese caso seja obtido um p-valor maior do que o nível de significância, geralmente configurado para 0,05. Neste caso, obteve-se um p-valor de 0,99, o que indica que o modelo estimado tem uma probabilidade de 99% de não apresentar autocorrelação.

Executando o modelo inicial escolhido, obtém-se diversas estatísticas. O primeiro item observado é o Critério de Akaike tendo valor de 7588,11, que por ora será apenas reservado para posterior comparação com outros modelos.

Para cada coeficiente do modelo estimado existe um p-valor associado que corresponde à relevância deste coeficiente no modelo. Caso o p-valor esteja acima de 0,05, então sugere-se re-estimar o modelo removendo tal coeficiente. O modelo atual sendo analisado apresenta os resultados conforme descritos na Tabela 4.3.

Pode-se notar que quatro coeficientes apresentam p-valores acima de $0,05$, o que indica que o modelo ainda pode ser trabalhado, ou seja, re-estimado reduzindo os parâmetros correspondentes a estes coeficientes.

O modelo apresenta uma aderência aos dados no período anterior à previsão, como pode ser visto na Figura 4.9.

Realizando a previsão para o período de 6 meses, partindo do mês de junho, obtém-se a Figura 4.10. É possível perceber que, apesar do modelo ter aparentemente se ajustado ao período anterior à previsão, a partir do momento em que esta é realizada, o modelo não consegue acompanhar o tráfego observado, gerando um intervalo de confiança extremamente grande, como representado no gráfico pela região sombreada em verde.

Com o modelo inicial ainda não é possível concluir sobre a previsão de 6 meses

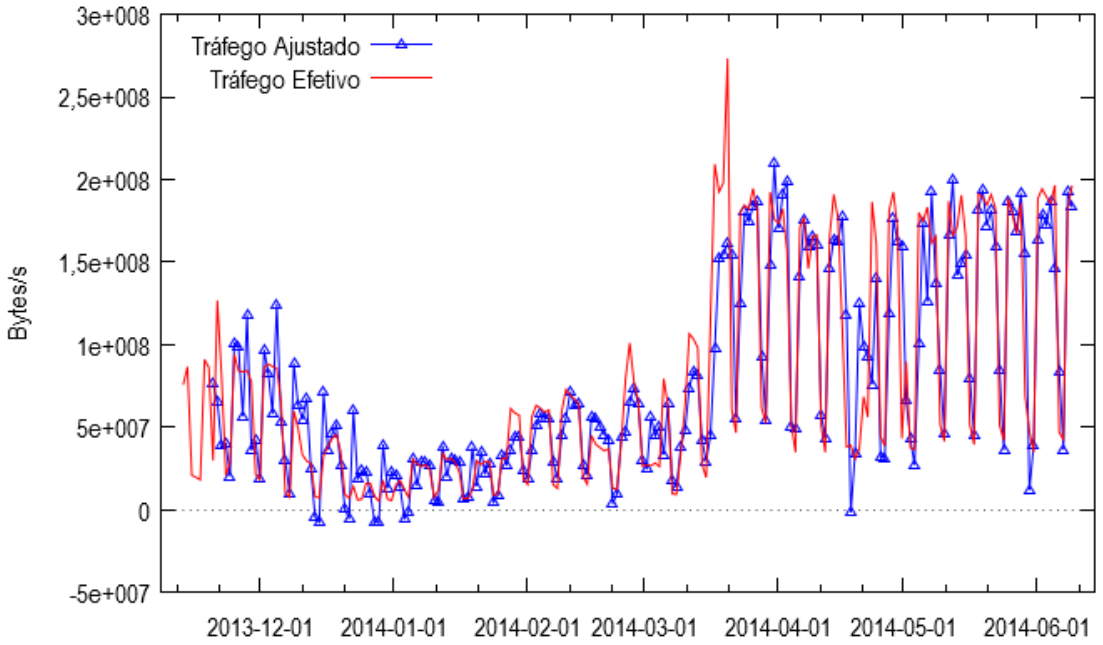


Figura 4.9: Comparação do tráfego ajustado pelo Modelo SARIMA(5,0,3)(0,1,3) com o tráfego observado

com um valor de MAPE que seja satisfatório, abaixo de 10%, dado que para este modelo o MAPE resultou em $35,06\%$.

Contudo, ainda é possível evoluir com o modelo, re-estimando com a redução no número de parâmetros de acordo com os coeficientes que apresentam p-valor acima de 0,05, conforme mencionando anteriormente, na tentativa de encontrar um modelo que apresente um resultado com MAPE menor.

Assim, fazendo primeiramente uma redução no parâmetro de média móvel, tem-se o modelo $SARIMA(5,0,2)(0,1,3)$ a ser estimado. Com esse modelo obteve-se, um MAPE de $35,88\%$ e Critério de Informação de Akaike (AIC) de $7592,42$. Como

Tabela 4.3: Resultado dos coeficientes para o modelo SARIMA(5,0,3)(0,1,3)

	Coeficiente	Erro Padrão	Estatística Z	p-valor
ϕ_1	−0,605109	0,219215	−2,7603	0,0058
ϕ_2	0,771402	0,246426	3,1304	0,0017
ϕ_3	0,595900	0,0968591	6,1522	0,0000
ϕ_4	−0,101738	0,180335	−0,5642	0,5726
ϕ_5	0,0703866	0,114144	0,6166	0,5375
θ_1	1,19844	0,238195	5,0314	0,0000
θ_2	−0,182231	0,356316	−0,5114	0,6091
θ_3	−0,630714	0,187149	−3,3701	0,0008
Θ_1	−0,762393	0,0846908	−9,0021	0,0000
Θ_2	0,231403	0,115569	2,0023	0,0453
Θ_3	−0,161710	0,0894587	−1,8077	0,0707

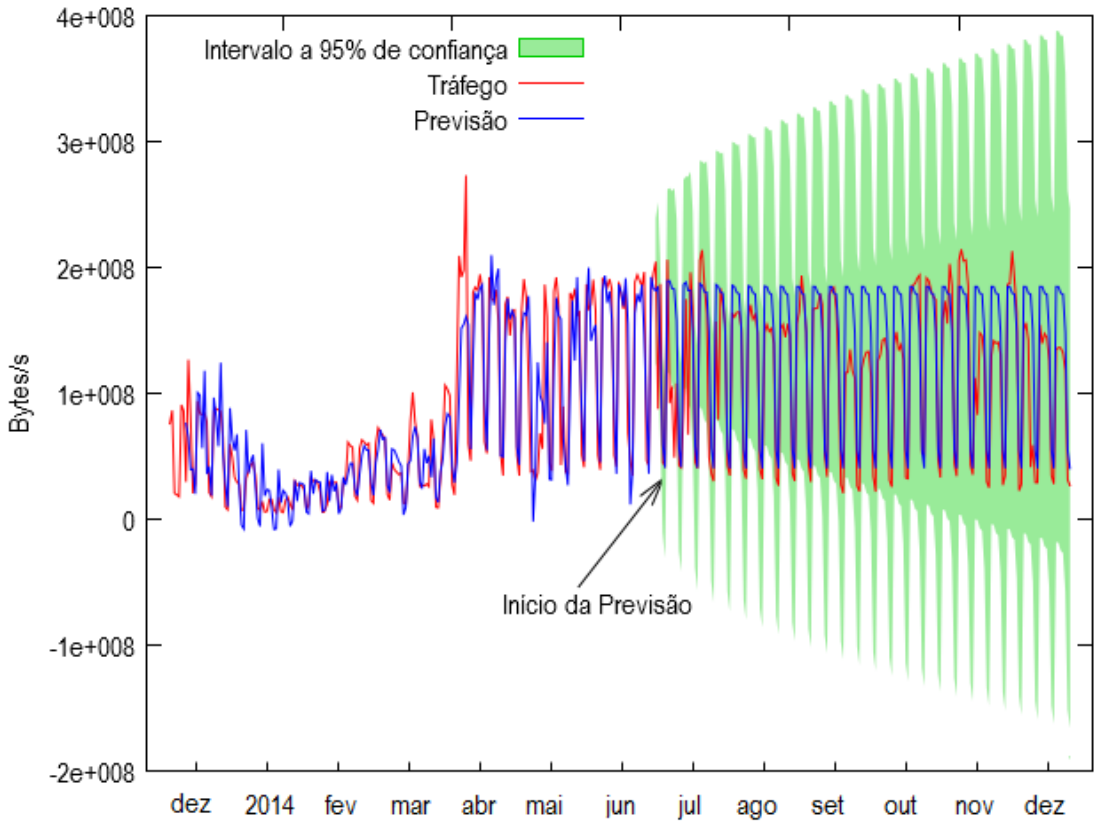


Figura 4.10: Previsão de 6 meses segundo o Modelo SARIMA(5,0,3)(0,1,3)

este resultado não foi melhor que o do modelo anterior, continua-se buscando um modelo mais adequado, desta vez com a redução de um parâmetro de autorregressão.

O próximo modelo, $SARIMA(4,0,2)(0,1,3)$, apresentou o AIC como $7590,46$, mas com o mesmo resultado do modelo anterior para o MAPE, indicando que o parâmetro removido realmente era indiferente no modelo.

Então, removendo mais um componente de autorregressão, tem-se o modelo $SARIMA(3,0,2)(0,1,3)$ com um AIC de $7587,67$, MAPE de $36,64\%$ e todos os coeficientes apresentando p-valor abaixo de 0,05. Isso indica que o modelo já está ajustado, sendo todos os coeficientes relevantes. Como o valor de MAPE não foi melhor que o do primeiro modelo ajustado, então o modelo considerado será o $SARIMA(5,0,3)(0,1,3)$. Vale ressaltar que todos os modelos testados passaram nos testes feitos para o primeiro modelo, como o teste de estacionariedade Ljung-Box, por exemplo.

Assim, para apresentar um resultado mais conclusivo, optou-se por reduzir o horizonte de previsão para uma janela de 30 dias, a fim de obter um intervalo de confiança reduzido. A Figura 4.11 mostra o resultado da previsão.

Nota-se que, para a previsão de 30 dias, o intervalo de confiança (região sombreada em verde a partir do início da previsão) ficou reduzido, atingindo um valor máximo próximo dos 250 MB/s, enquanto que a previsão anterior apresentou um

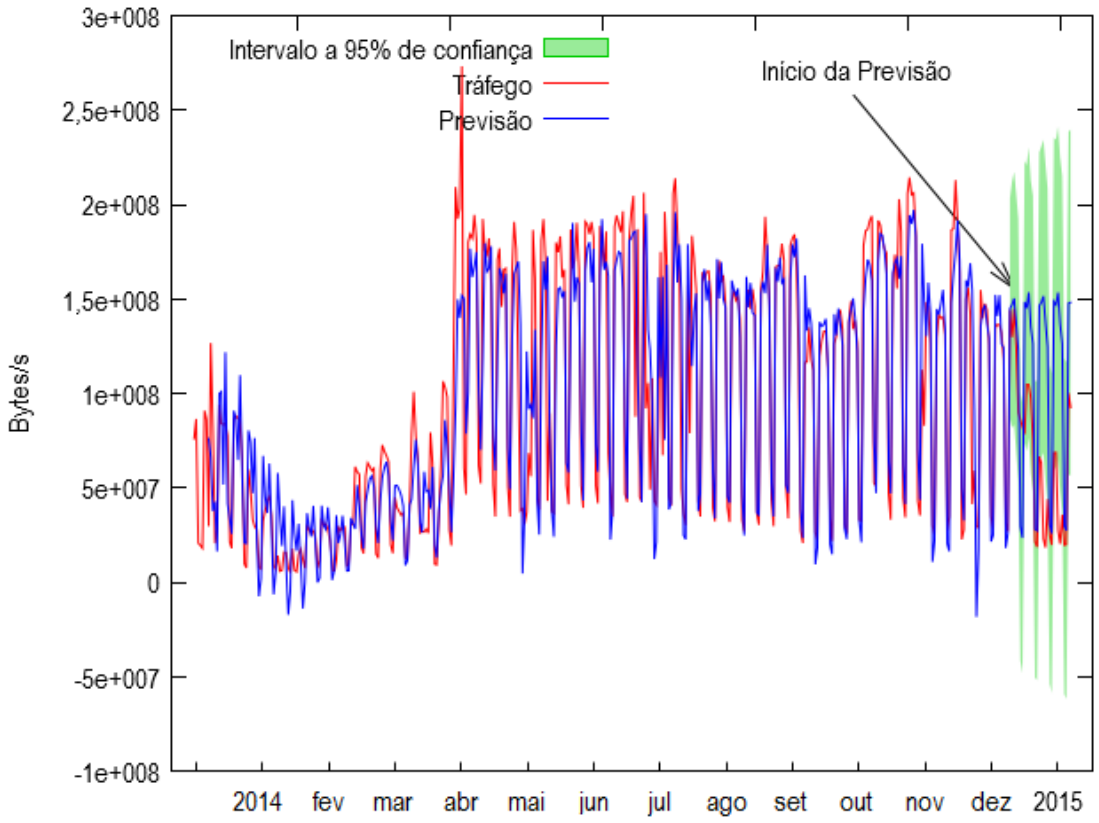


Figura 4.11: Previsão de um mês segundo o Modelo SARIMA(5,0,3)(0,1,3)

intervalo próximo dos 400 MB/s. Assim, tem-se um resultado mais apropriado para a tomada de decisão. Contudo, o ideal é que seja feita uma previsão para um mínimo de 6 meses. Para isso, alguns testes experimentando outras granularidades dos dados foram feitos e são mostrados na seção seguinte.

4.4 Testes com Diferentes Granularidades e Dados Sintéticos

A acurácia da previsão está associada também à granularidade dos dados. Em outras palavras, com uma granularidade de dias, as previsões serão tomadas também na escala de tempo de dias. Porém, para o modelo isso é transparente, pois a previsão é feita com base na quantidade de amostras que serão previstas. Se é desejado realizar uma previsão de 6 meses, isso significa prever 180 valores, já que as amostras são diárias. Contudo, se a quantidade de valores a serem previstos é muito grande, então o erro de previsão tende a piorar.

Portanto, a granularidade é trabalhada, construindo-se valores baseados nas amostras coletadas. Assim, outro teste feito foi construir valores de médias de escalas de tempo de meses e comparar os resultados de previsão.

O teste foi feito com base nos dados coletados, gerando médias mensais sobre eles, resultando em uma amostra para cada mês, totalizando 12 amostras. A pergunta que surge é: será que a previsão realizada para um horizonte de um mês utilizando a granularidade mensal apresentará um resultado melhor (MAPE menor) do que na granularidade diária?

Para responder essa pergunta, é necessário que um número maior de amostras esteja disponível, como visto em [45], que mostra um estudo sobre o número mínimo de amostras necessário para realizar a previsão com modelos sazonais, como o SARIMA utilizado nesta dissertação. Nesse trabalho, os autores sugerem que o número mínimo de amostras é dado pela quantidade de parâmetros a serem estimados no modelo em questão.

No caso anterior, no modelo $SARIMA(5,0,3)(0,1,3)$ tem-se $5+3+1+3=12$ coeficientes a serem estimados, que é exatamente a quantidade de valores mensais disponíveis. Porém, ainda é necessário separar uma parte das amostras para a etapa de estimação e outra parte para verificação, como foi feito anteriormente. Portanto, não há uma quantidade de dados suficiente para realizar a previsão com base na granularidade de meses.

Entretanto, para contornar essa limitação a fim de avaliar se a adoção do modelo SARIMA é relevante para o estudo em questão, foram também utilizados dados sintéticos, tendo como base as amostras já disponíveis, buscando previsões mais acuradas, certificando que o modelo é adequado.

Considerou-se os valores mensais relativos à média do tráfego na Interface Level 3. Com base nesses valores, três séries foram utilizadas. Uma série considerando alguns valores discrepantes (*outliers*), com o intuito de observar o comportamento do modelo sob tais condições; outra série apresentando uma média constante; e na terceira série foi adicionada uma componente de tendência, baseada no crescimento do tráfego observado, através das medidas do ano coletado.

As três séries foram geradas com um período de cinco anos de medidas, sendo que um ano de medidas, correspondente às últimas 12 amostras como na Figura 4.12, provém de medições reais, tendo em vista os dados do ano de 2014. Os outros quatro anos sintéticos foram baseados no ano posterior. Isso significa que, os dados para o ano de 2013 foram gerados a partir dos dados de 2014; os dados do ano de 2012 foram criados com base em 2013; e assim por diante. Conforme pode ser visto na Figura 4.12, alguns valores mais altos em relação à média (apresentando uma série com valores discrepantes) foram deixados para perceber como o modelo se comporta.

Utilizando dos mesmos procedimentos descritos pela metodologia apresentada no Capítulo 3 e comentados no exemplo da Level 3, considerou-se para esta série o modelo inicial $ARIMA(1,1,2)$, tendo em vista a análise da evolução deste modelo

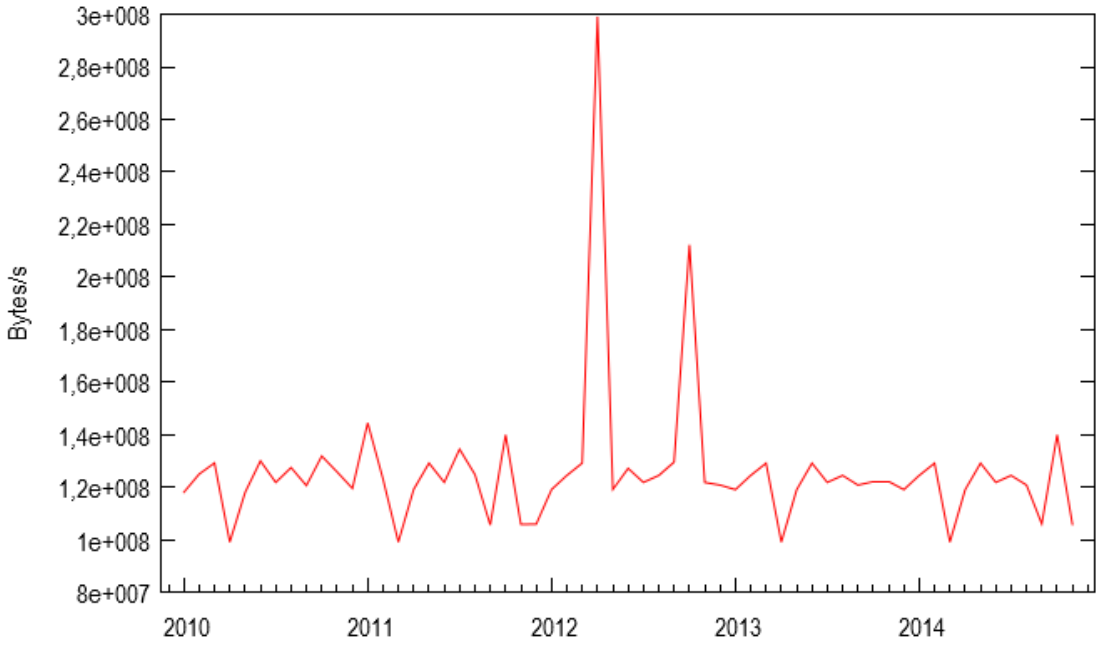


Figura 4.12: Tráfego sintético com picos gerado a partir dos dados de 2014

posteriormente para o modelo SARIMA mais robusto.

Pelo resultado da previsão neste caso, é possível perceber a influência dos valores de pico sobre o intervalo de confiança da previsão, como na Figura 4.13.

Analisando a Figura 4.13, nota-se que a partir dos valores de pico, o modelo de previsão utilizado sofre uma perturbação, gerando um intervalo de confiança significativo para o período de um ano de previsão, com valor de MAPE de $9,21\%$.

Buscando outro modelo com o objetivo de obter um melhor ajuste, o comportamento ainda assim se repete. O modelo ARIMA(2,2,1) até acompanha os picos, porém, apesar de apresentar um valor de MAPE de $7,22\%$, acaba resultando em um intervalo de confiança ainda maior, conforme pode ser observado na Figura 4.14.

Na tentativa de eliminar o efeito dos valores de pico, realizou-se o ajuste do modelo de previsão após os valores discrepantes, gerando a Figura 4.15. Neste caso, o MAPE ficou em $7,20\%$ e o intervalo de confiança foi bastante reduzido, porém, dada a pouca quantidade de amostras disponíveis, o intervalo não considerou todos os valores reais, não fornecendo confiabilidade à previsão, o que corrobora com as considerações feitas anteriormente.

Em sequência, alterando-se os valores discrepantes, foi utilizada a segunda série de estudo, resultando na Figura 4.16. Essa série já apresenta variância menor, sendo semelhante a um ruído branco, o que sugere estado estacionário, contribuindo para um modelo mais preciso.

Fazendo a previsão partindo de um modelo simples (ARIMA(0,0,1)), obtém-se ainda assim um resultado inconsistente, tendo em vista que o intervalo de confiança

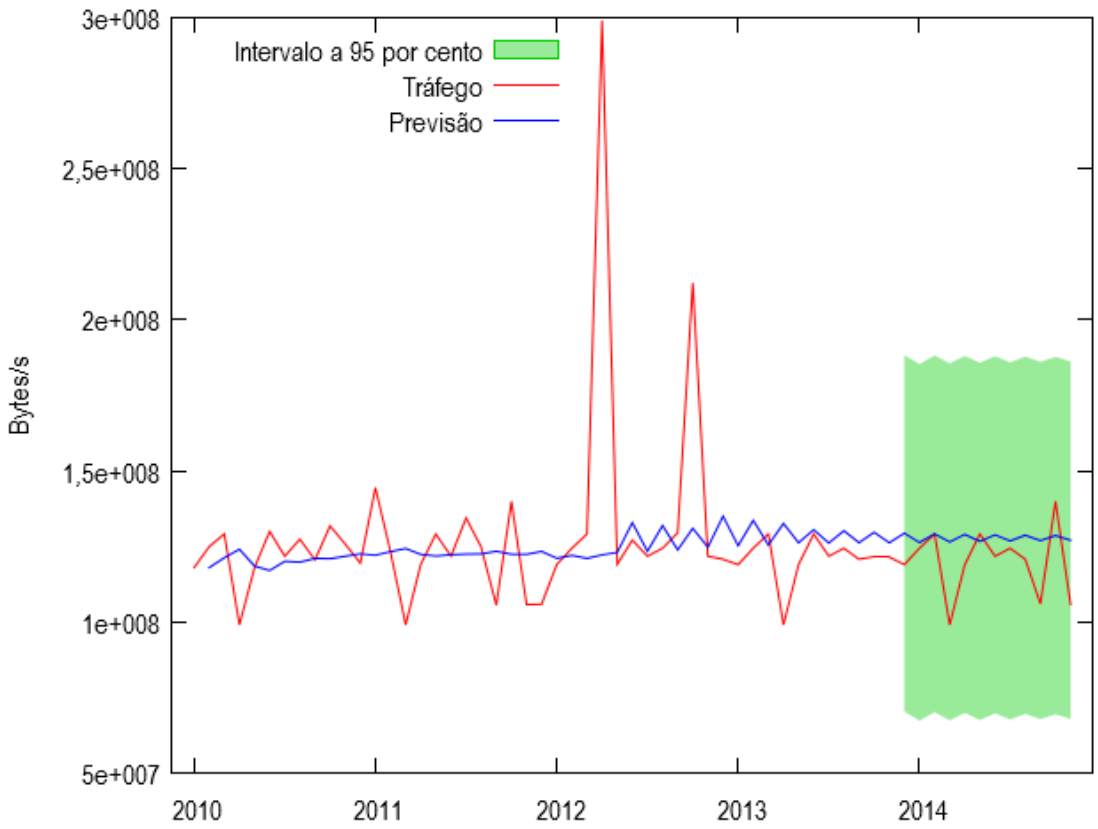


Figura 4.13: Previsão de um ano utilizando ARIMA(1,1,2)

desta vez não contempla todos os valores reais da série, como na Figura 4.17. O MAPE neste caso ficou em $7,48\%$. Sendo assim, deslocando-se o início da previsão para depois do primeiro valor que saiu do intervalo, o modelo acaba por se ajustar, conforme a Figura 4.18, com MAPE um pouco maior $7,83\%$.

Porém, a fim de alcançar um modelo mais preciso, parte-se para um modelo um pouco mais complexo, levando em consideração um componente de sazonalidade. Neste modelo, apresentado na Figura 4.19, o intervalo de confiança acompanha os dados reais de maneira mais próxima do comportamento destes, com MAPE de $9,70\%$. Todavia, nota-se que, como se trata de um modelo mais complexo, mais amostras se tornam necessárias.

Neste caso, utiliza-se o correspondente a um ano de amostras (12 amostras) devido a componente de autocorrelação sazonal. Conclui-se assim que, quanto mais complexo for o modelo, mais amostras são necessárias para alcançar uma previsão aceitável.

A fim de aproximar os testes a um comportamento de tráfego mais realista, o último teste considera a terceira série de teste gerada com base no crescimento médio do tráfego no enlace da Interface Level 3. Este crescimento foi observado utilizando os dados coletados ao longo do ano de 2014, resultando em um crescimento de 10% no volume do tráfego. Essa forma de geração criou uma componente de tendência

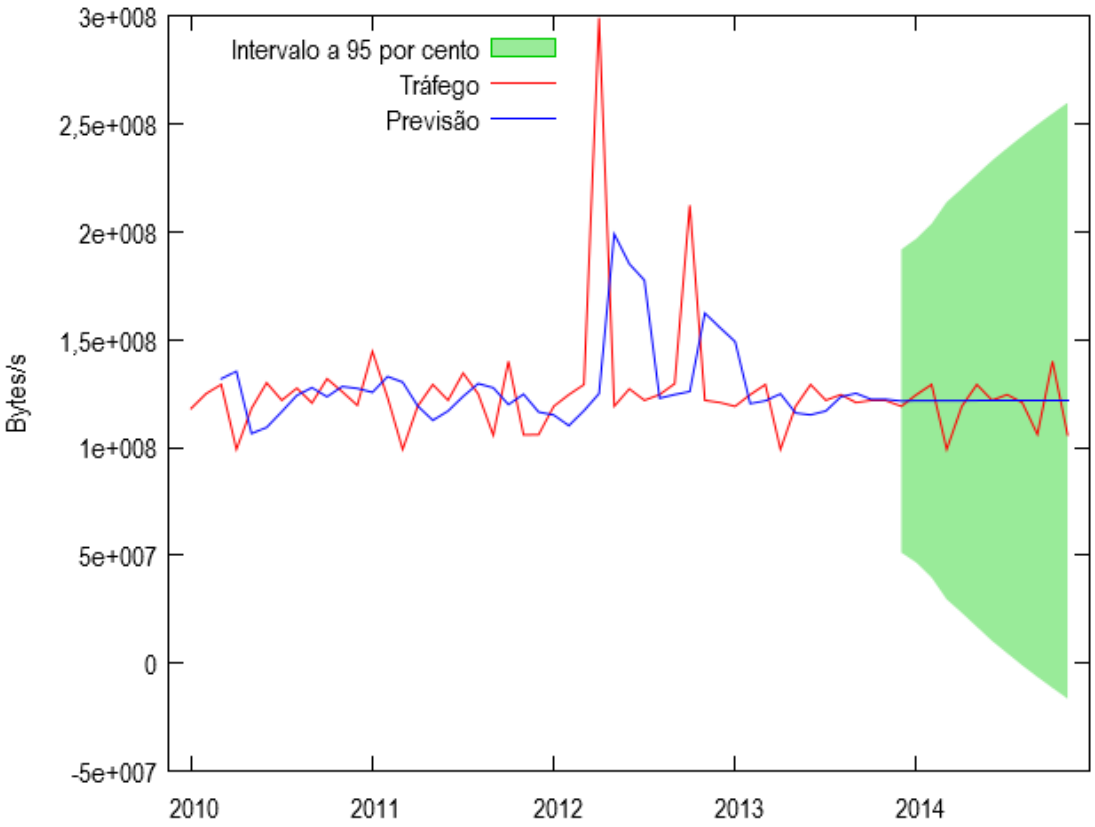


Figura 4.14: Previsão de um ano utilizando ARIMA(2,2,1)

no tráfego.

A previsão neste caso, utilizando o modelo estimado ARIMA(2,1,2), já aproxima bem o intervalo de confiança ao tráfego estimado, com MAPE de $7,75\%$, conforme a Figura 4.20. Utilizando um modelo sazonal SARIMA(1,1,1)(0,1,0), a previsão fica praticamente sem erros, como pode ser visto na Figura 4.21. Neste caso, o valor do MAPE ficou em $0,6789$.

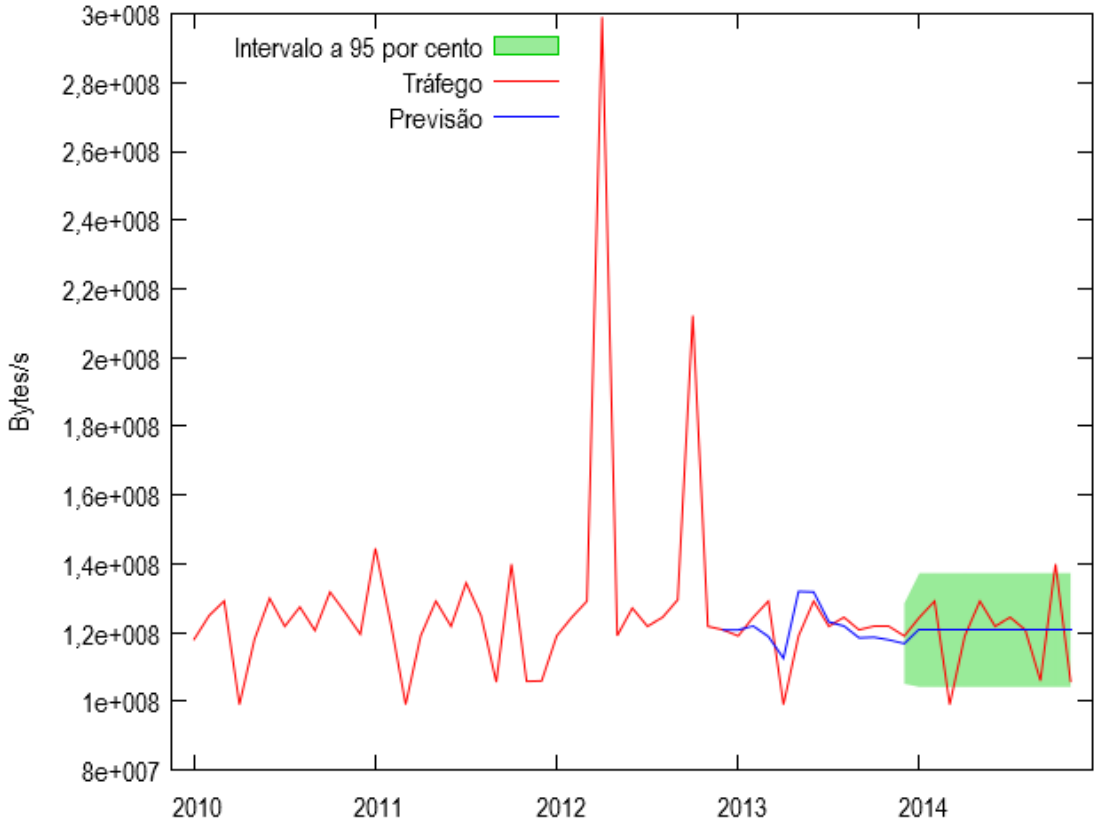


Figura 4.15: Previsão de um ano utilizando ARIMA(0,0,1)

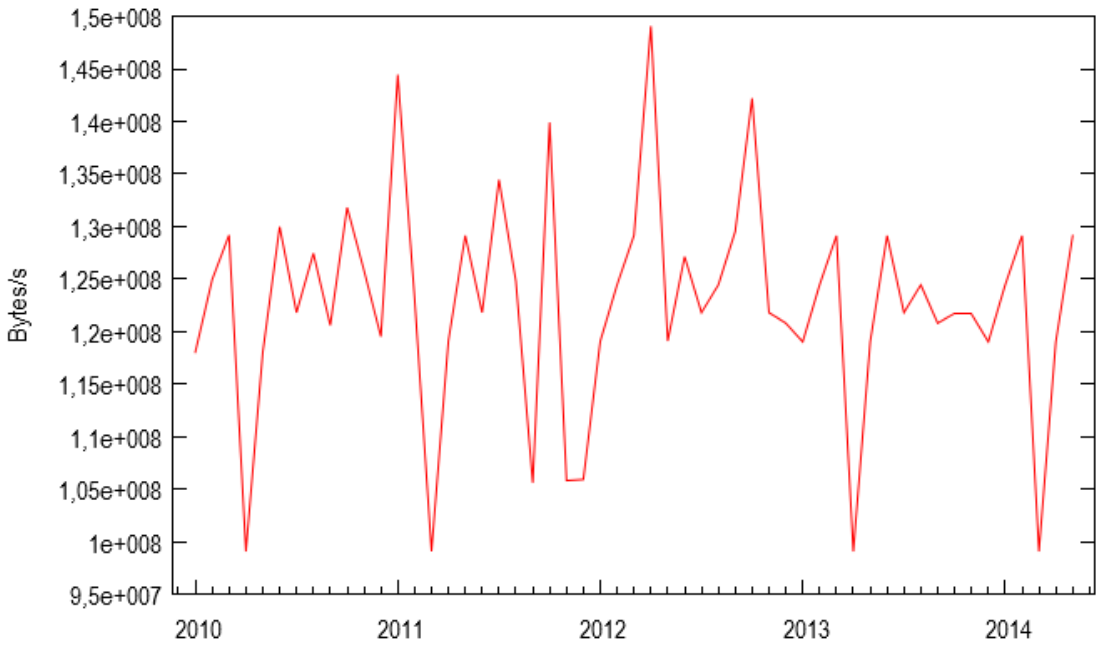


Figura 4.16: Série com valores de pico suavizados

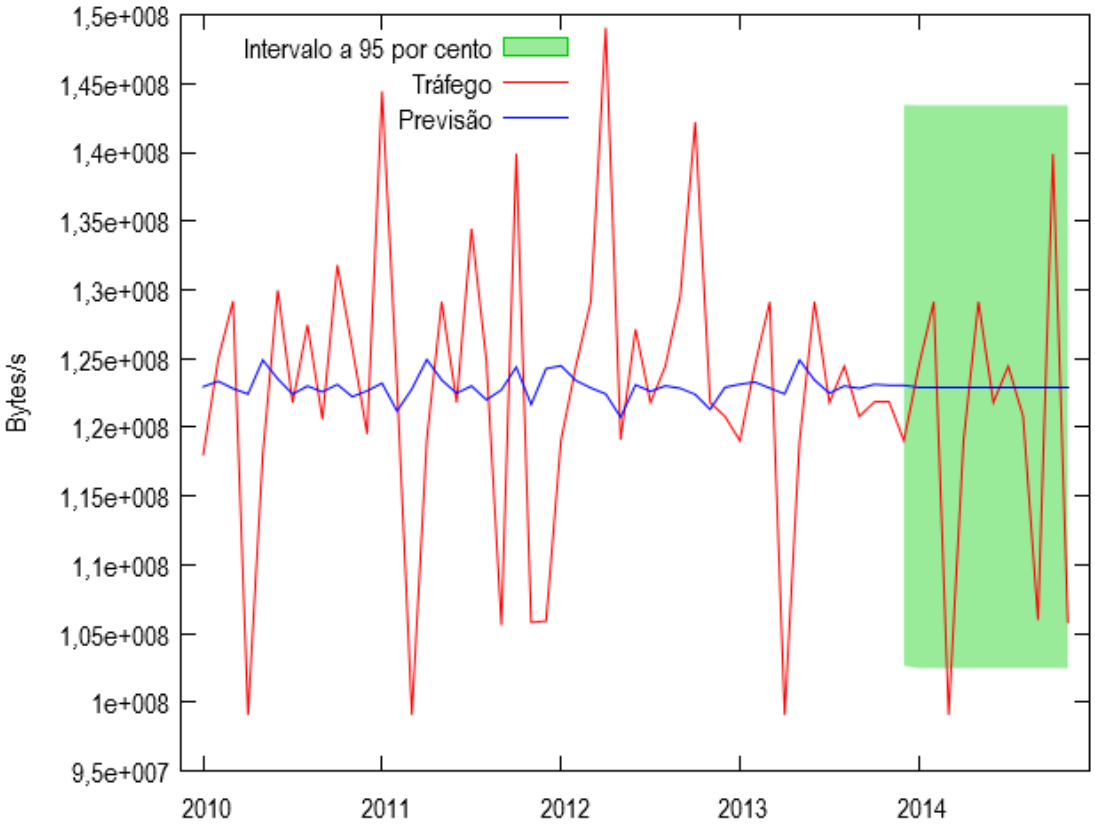


Figura 4.17: Previsão de um ano utilizando ARIMA(0,0,1)

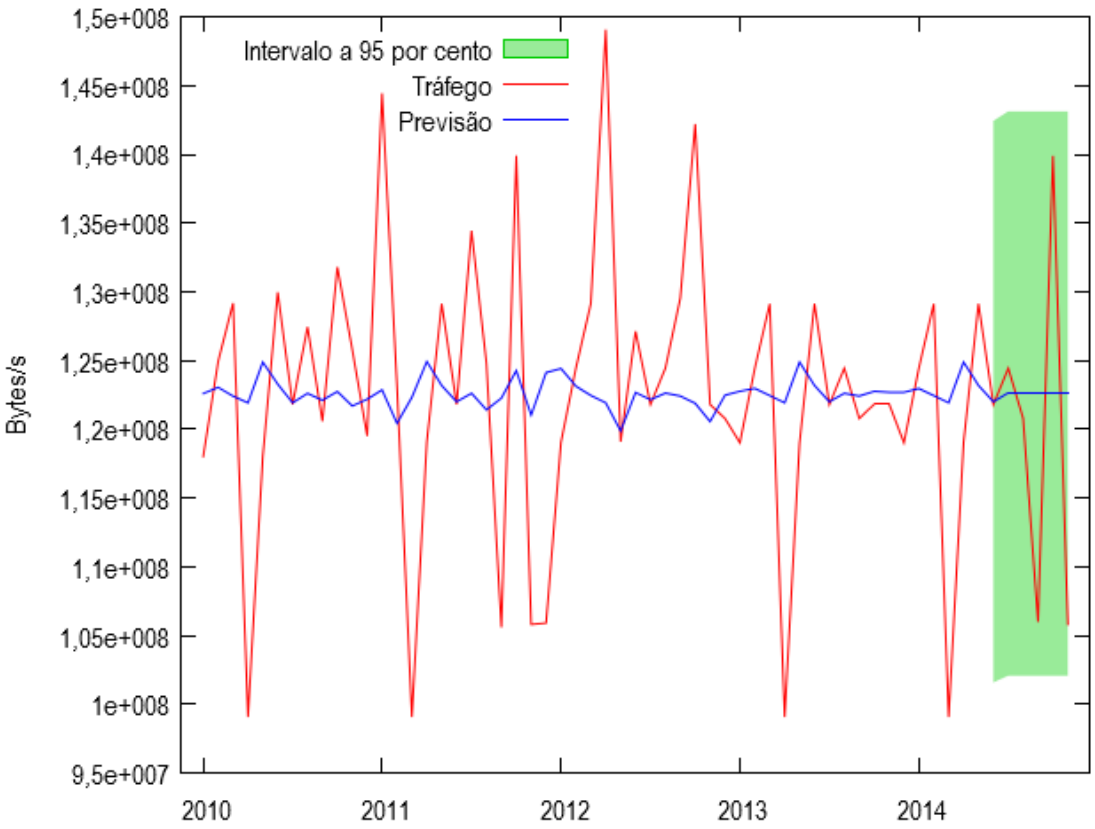


Figura 4.18: Previsão de 6 Meses utilizando ARIMA(0,0,1)

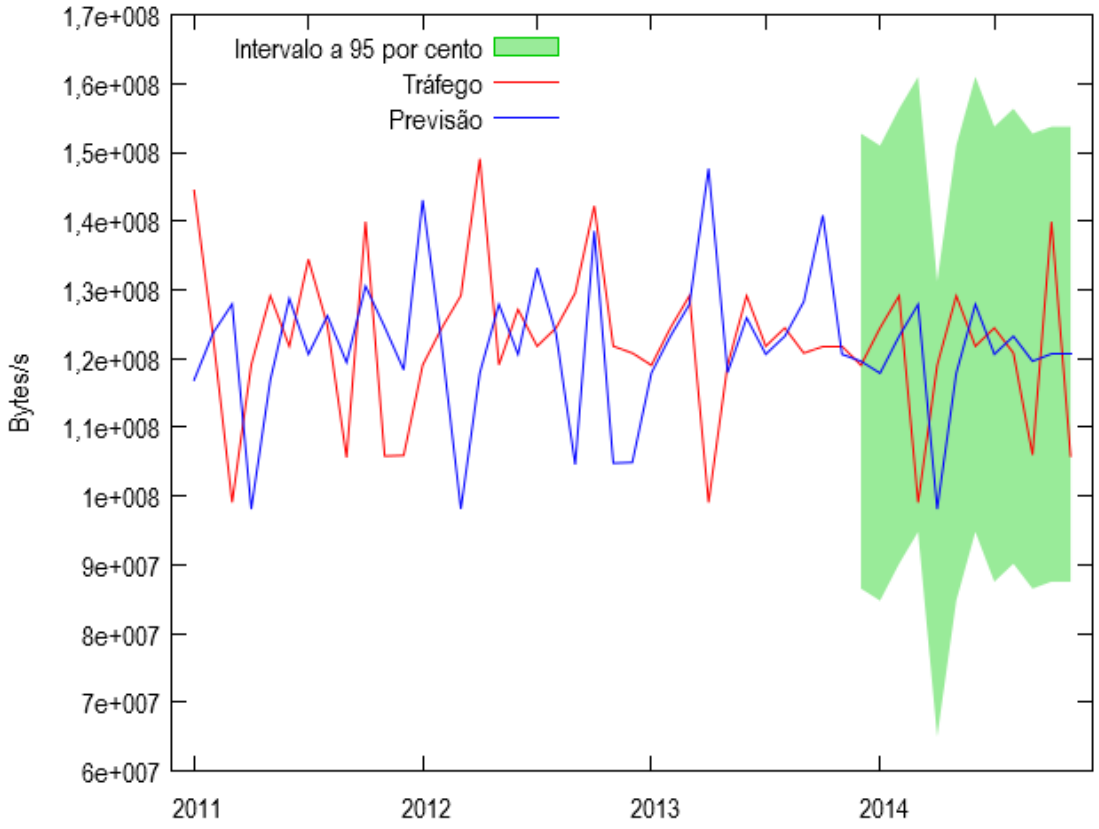


Figura 4.19: Previsão de um ano utilizando SARIMA(0,0,0)(1,0,0)

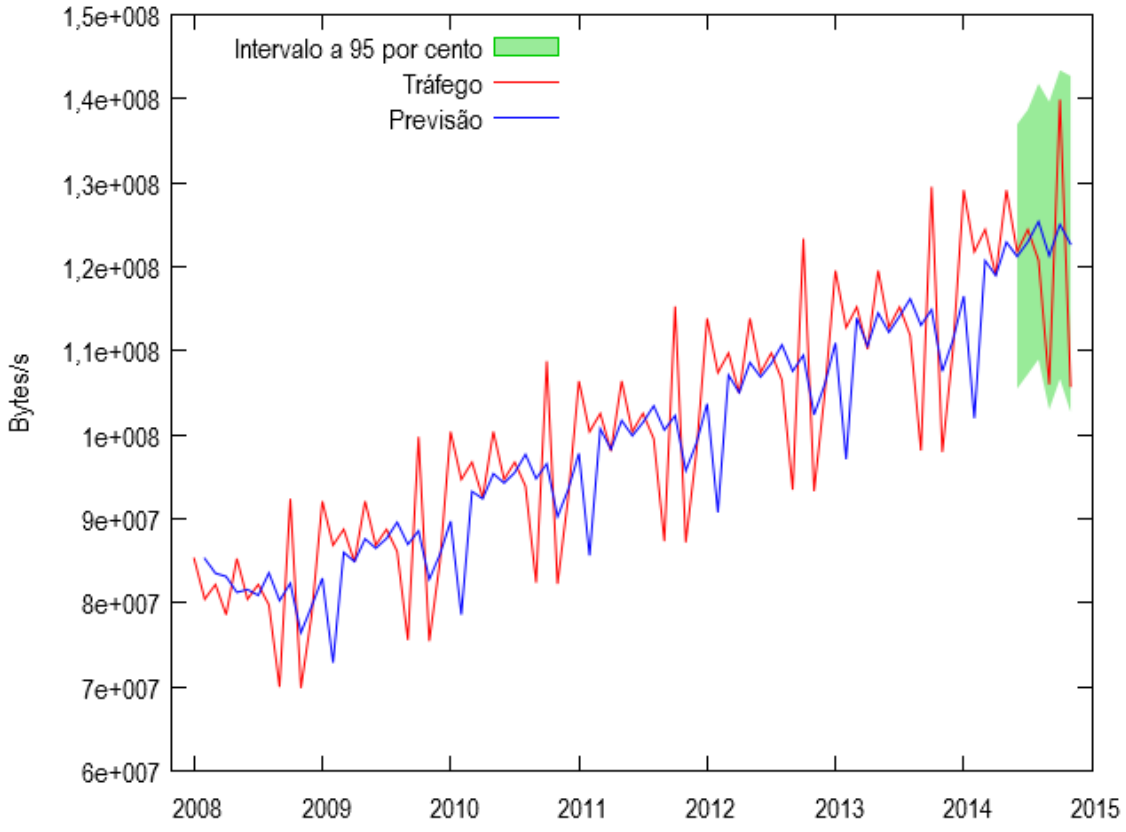


Figura 4.20: Previsão de 6 meses utilizando ARIMA(2,1,2)

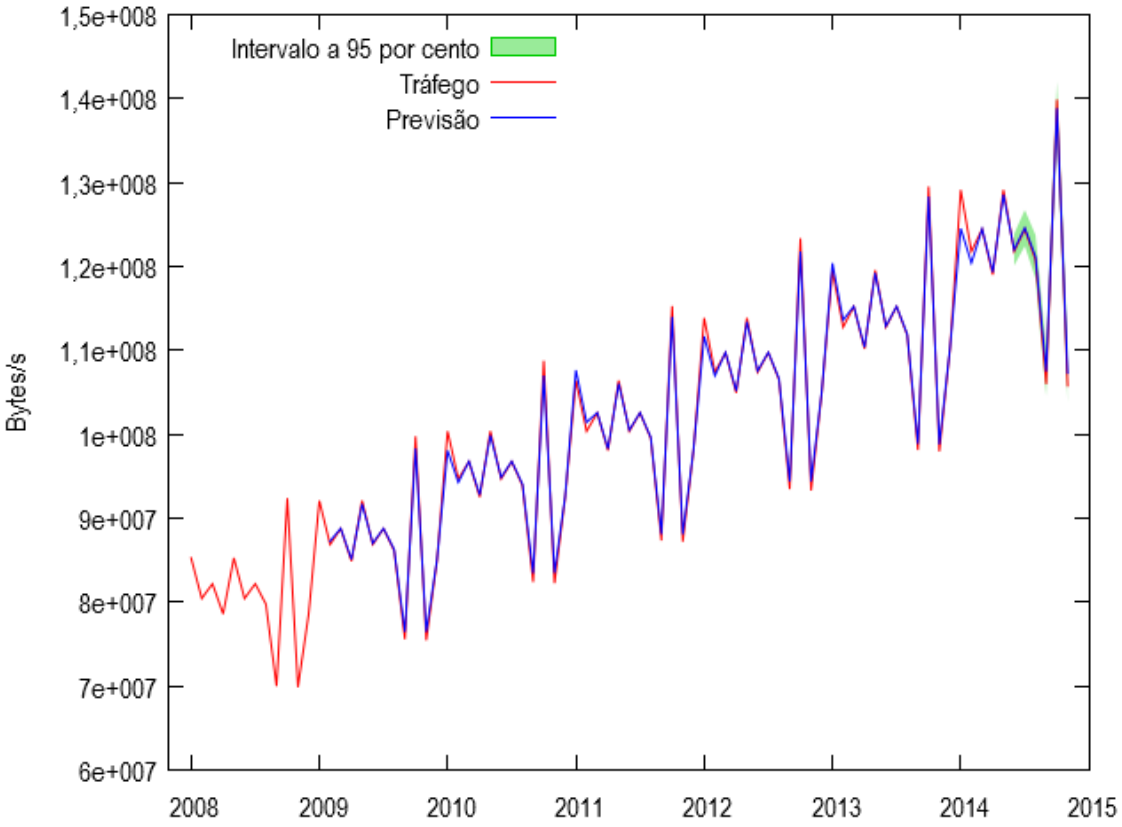


Figura 4.21: Previsão de 6 meses utilizando SARIMA(1,1,1)(0,1,0)

4.5 Considerações Finais

As previsões realizadas sobre os dados gerados a partir dos dados coletados de 2014 mostram que a previsão pelo modelo SARIMA é adequada, caso hajam dados suficientes para estimar o modelo.

Ainda em tempo, considerou-se também diferentes formas de interpretação dos dados, trabalhando não só com granularidades diferentes, mas também com perspectivas diferentes sobre os dados, como a utilização de valores de pico, em vez de apenas valores de média.

Utilizar valores de pico pode proporcionar uma previsão mais "pessimista" no sentido de que irá apresentar resultados em relação às maiores taxas observadas. Ou seja, os valores de pico caracterizam o pior caso do tráfego ter um comportamento o mais próximo possível da capacidade do enlace, como discutido na Seção 3.2.2.

Contudo, a condição de pico não necessariamente é percebida todo tempo, pois nem sempre o tráfego de um determinado enlace estará em condição de pico. Ainda assim, os valores de pico podem trazer informações tendenciosas devido a fatos isolados, como falhas de segurança, ataques, medições erradas, dentre outros fatores, que acabam por tornar os valores de pico não tão representativos quanto os valores de média. Neste último caso, a previsão realizada fica mais ajustada ao comportamento médio do tráfego no enlace, o que entrega um resultado mais próximo da realidade.

Se os valores de pico são muito frequentes, isto é, se o enlace passa uma boa porcentagem do tempo em condições de alta utilização, então, ainda assim, o valor de média será uma boa estimativa a ser utilizada para previsão, pois irá acompanhar essa característica de alta utilização, visto que a média também será elevada. Portanto, é relevante o uso de valores de média para os propósitos de previsão de tráfego.

Outras fontes de dados também foram consultadas, como a base dados CAIDA (*Center for Applied Internet Data Analysis* [62]), que é um centro de esforço colaborativo entre organizações comerciais, governamentais e de pesquisa, para manutenção da infraestrutura de Internet. A CAIDA forneceu dados sobre registros (*traces*) anônimos de tráfego, coletados no período de 2008 a 2015. A princípio, essa massa de dados seria fundamental para alcançar mais resultados. Contudo, quando se obteve acesso à base, percebeu-se que os dados não apresentavam continuidade, por conta de sumarização devido à limitação de capacidade de armazenamento, o que não permitiu chegar a resultados conclusivos.

Assim, tem-se uma metodologia, válida a ser aplicada na forma de uma ferramenta de previsão de tráfego, que atenda às necessidades dos provedores de acesso, em relação ao provisionamento de recursos de forma antecipada, a fim de evitar que um enlace alcance condições de congestionamento, afetando negativamente na

qualidade do serviço prestado.

Além da previsão de tráfego, a primeira parte da metodologia, que tratou da caracterização dos dados, agrega ainda mais valor a proposta, combinando a caracterização, o ajuste de distribuição e a previsão de tráfego, possibilitando a construção de uma ferramenta robusta de tomada de decisão. Robusta no sentido de uma expansão da solução, em termos de ferramenta de prospecção de novas redes, na qual é factível a simulação de novas topologias. Assim, é possível experimentar diversas configurações para a rede em estudo, com o diferencial de utilizar os resultados da caracterização de redes já existentes, como entrada para a simulação da nova rede.

Com isso, vislumbra-se um ambiente de simulação de redes que permita aos administradores de rede ajustar parâmetros de configuração, tais como o número de elementos na rede e velocidade dos enlaces, tendo o tráfego gerado com base em informações e características de tráfego real. Além disso, poder também acompanhar o comportamento dessa rede ao longo do tempo, através dos resultados de previsão.

Vale lembrar que um modelo é uma ferramenta de tomada de decisão. E se o modelo permite que sejam tomadas decisões corretas, ainda que não contemple todas as características do mundo real, o modelo é aceitável, como mencionado por Grossglauser e Bolot em [63].

Capítulo 5

Conclusão e Trabalhos Futuros

Na linha de Redes de Computadores, diversos desafios são realidade. Questões como roteamento, previsão de tráfego, detecção de anomalias, qualidade de serviço, entre outras, são de grande importância nesta área. E ainda, como estudos mostram, há um aumento na demanda por recursos devido ao crescimento no volume de tráfego global [2], o que torna essas questões ainda mais desafiadoras.

O problema abordado neste trabalho tratou da previsão de séries temporais, no qual, através de modelos matemáticos, foi possível estimar o comportamento futuro do tráfego de um determinado enlace de dados. Para isto, este trabalho apresentou uma metodologia de previsão de tráfego tendo como base os trabalhos de [11], [29] e [30]. Também foi levada em consideração a caracterização e ajuste dos dados coletados, afim de aproximar o comportamento estatístico do tráfego no enlace estudado.

Para a previsão, foi utilizado o modelo de séries temporais ARIMA e sua vertente sazonal SARIMA. Os resultados dos testes realizados confirmam a viabilidade da metodologia, apresentando previsões para 1 mês com os dados de 1 ano fornecidos pela RedeRio de Computadores/FAPERJ, sendo factível de ser aplicada na forma de uma ferramenta de previsão de tráfego.

As contribuições deste trabalho são os modelos e distribuições identificados para o estudo de caso utilizado, obtidos através da caracterização dos dados coletados. E também, a descrição da metodologia apresentada no Capítulo 3, que foi detalhada e que pode ser utilizada como uma ferramenta de auxílio para tomada de decisão.

As principais dificuldades encontradas ao longo do desenvolvimento do trabalho se deram pela necessidade de um repositório de dados, no qual um histórico de informações de rede fosse registrado e armazenado, preferencialmente em uma granularidade o maior possível e por um longo período de tempo. Através deste repositório, seria possível desenvolver diversos trabalhos, não só em relação à previsão de tráfego, mas também em termos de outros trabalhos, tais como detecção de anomalias, inferências sobre qualidade de serviço, alocação dinâmica de recursos,

entre outros, que podem se beneficiar dos dados armazenados.

Apesar de haver bases de dados sobre redes, como é o caso do CAIDA [62] e do crescimento da tabela BGP (*Border Gateway Protocol*) [64], na maioria das vezes, essas bases não mantêm os dados sobre tráfego por muito tempo, o que acaba por limitar determinados estudos.

Sugere-se como um trabalho futuro a elaboração de uma base de dados para armazenamento de informações de rede, que possa servir de fonte de dados a serem utilizados em trabalhos acadêmicos e até em ferramentas de negócio. Propõe-se ainda que essa plataforma seja disponibilizada preferencialmente de forma pública, salvas as condições de segurança e privacidade dos dados que forem coletados.

O estudo apresentado nesta dissertação cria diversas oportunidades de trabalhos futuros, sendo a principal delas a composição de uma ferramenta de monitoramento e planejamento de Redes de Computadores, que disponibilize aos gerentes e administradores de rede uma plataforma na qual seja possível analisar o comportamento dos enlaces da rede ao longo de uma janela de tempo futura predeterminada. Com isso, permite-se simular situações futuras da rede, fazendo ajustes e adaptações nas métricas utilizadas (vazão, atraso e taxa de perda, por exemplo), observando como a rede se comporta sob as novas condições, dentro de uma janela de previsão, o que pode contribuir para a tomada de decisão.

Outro trabalho futuro que pode ser pontuado é a implementação de um sistema de decisão de ações a serem executadas através de mecanismos inteligentes, como Redes Neurais Artificiais e Aprendizado de Máquina. Esses mecanismos auxiliariam nas decisões a serem tomadas de forma a automatizá-las, tornando o sistema mais atuante na rede.

Também vislumbra-se aprimorar as previsões, tendo em vista a utilização de informações de roteamento, a fim de prover mais certeza em uma decisão a ser tomada e para o ambiente de simulação comentado na Seção 4.5.

Outro aspecto que pode ser explorado é o estudo de outras abordagens no que tange aos modelos matemáticos utilizados. Propõe-se o estudo de técnicas de Aprendizado de Máquina para a realização das previsões. É interessante também buscar um modelo que considere uma série temporal não estacionária como objeto de estudo. Em outras palavras, isso significa buscar um modelo que seja capaz de tratar uma série temporal de maneira mais geral, incluindo as componentes de tendência e sazonalidade, que usualmente são removidas através das operações de diferença. E também, como pode ser observado ao longo do trabalho, a componente de sazonalidade aparece em diversas escalas de tempo, como a sazonalidade diária, semanal e mensal. Propõe-se assim um estudo que contemple mais de um tipo de sazonalidade de forma simultânea. Com isso, os modelos manteriam características reais da séries temporal, o que poderia levar a resultados mais precisos.

Para concluir, outra linha de pesquisa que pode ser incluída nas continuações deste trabalho é a linha de Redes Definidas por *Software* (*Software-Defined Networking - SDN*) [65]. Este campo de estudo vem tendo forte atenção pela comunidade científica, sendo um campo de pesquisa com diversos trabalhos em aberto, conforme visto em [65, 66]. O conceito fundamental de SDN é a separação do plano de controle (no qual os algoritmos de roteamento executam) do plano de encaminhamento dos dados (correspondente aos equipamentos de rede) [65]. A adoção desta abordagem proporciona uma abertura para um vasto campo de estudos, permitindo a atuação imediata na rede após as previsões serem feitas. Desta forma, pode-se atuar na melhoria do desempenho das redes, tendo como base uma estimava sobre a demanda futura do tráfego, com previsões de curto a médio prazo.

Referências Bibliográficas

[1] CETIC.BR. "Pesquisa TIC Domicílios". 2015. Disponível em: <`http://cetic.br/pesquisa/domicilios/`>. Acessado em Agosto de 2015.

[2] CISCO SYSTEMS. "Cisco Visual Networking Index: Forecast and Methodology 2012 – 2017", *White Paper da Cisco*, 2013.

[3] KUROSE, J. F., ROSS, K. W. *Computer Networking: A Top-Down Approach*. 5ª ed. USA, Addison-Wesley Publishing Company, 2009. ISBN: 0136079679, 9780136079675.

[4] SOULE, A., LAKHINA, A., TAFT, N., et al. "Traffic Matrices: Balancing Measurements, Inference and Modeling". In: *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '05, pp. 362–373, New York, NY, USA, 2005. ACM. ISBN: 1-59593-022-1. doi: 10.1145/1064212.1064259.

[5] ROUGHAN, M., THORUP, M., ZHANG, Y. "Traffic Engineering with Estimated Traffic Matrices". In: *Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement*, IMC '03, pp. 248–258, New York, NY, USA, 2003. ACM. ISBN: 1-58113-773-7. doi: 10.1145/948205.948237.

[6] KLEINROCK, L. *Communication Nets: Stochastic Message Flow and Delay*. New York, McGraw-Hill Book Company, 1964.

[7] MEDINA, A., TAFT, N., SALAMATIAN, K., et al. "Traffic Matrix Estimation: Existing Techniques and New Directions". In: *Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '02, pp. 161–174, New York, NY, USA, 2002. ACM. ISBN: 1-58113-570-X. doi: 10.1145/633025.633041.

[8] BOX, G. E. P., JENKINS, G. M. *Time Series Analysis: Forecasting and Control*. Prentice Hall PTR, 1994. ISBN: 0130607746.

[9] MORETTIN, P. A., TOLOI, C. M. C. *Análise de séries temporais*. ABE - Projeto Fisher. Edgard Blucher, 2006. ISBN: 9788521203896.

[10] LEON-GARCIA, A. *Probability, Statistics and Random Processes for Electrical Engineering*. Addison-Wesley Publishing Company, 1994.

[11] PAPAGIANNAKI, K., TAFT, N., ZHANG, Z.-L., et al. "Long-term forecasting of Internet backbone traffic: observations and initial models". In: *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, v. 2, pp. 1178–1188, Março 2003. doi: 10.1109/INFCOM.2003.1208954.

[12] PAPAGIANNAKI, K., TAFT, N., ZHANG, Z., et al. "Long-term forecasting of Internet backbone traffic", *IEEE Transactions on Neural Networks*, v. 16, n. 5, pp. 1110–1124, Setembro 2005. ISSN: 1045-9227. doi: 10.1109/TNN.2005.853437.

[13] ZHANI, M., ELBIAZE, H., KAMOUN, F. "Analysis and prediction of real network traffic", *Journal of networks*, v. 4, n. 9, pp. 855–865, 2009. doi: 10.4304/jnw.4.9.

[14] KUAN HOONG, P., TAN, I. K. T., YIK KEONG, C. "BitTorrent Network Traffic Forecasting With ARMA", *International journal of Computer Networks & Communications*, v. 4, n. 4, pp. 143–156, 2012. ISSN: 09752293. doi: 10.5121/ijcnc.2012.4409.

[15] REDERIO. "RedeRio de Computadores/FAPERJ". 2015. Disponível em: <`http://www.rederio.br`>. Acessado em Agosto de 2015.

[16] REDECOMEP. "Redes Comunitárias de Educação e Pesquisa - Redecomep". 2015. Disponível em: <`http://www.redecomep.rnp.br`>. Acessado em Agosto de 2015.

[17] LEVEL 3. "Level 3 Communications". 2015. Disponível em: <`http://www.level3.com/pt`>. Acessado em Agosto de 2015.

[18] BROWNLEE, N., MILLS, C., RUTH, G. "RFC-2722 Traffic Flow Measurement: Architecture". Outubro 1999. Disponível em: <`https://tools.ietf.org/html/rfc2722#section-2.1`>.

[19] QUITTEK, J., ZSEBY, T., CLAISE, B., et al. "RFC-3917 Requirements for IP Flow Information Export". Outubro 2004. Disponível em: <`https://tools.ietf.org/html/rfc3917`>.

[20] MUUSS, M. "Ping - Networking Utility". 1983. Disponível em: <http://linux.die.net/man/8/ping>.

[21] DELGADILLO, K. "NetFlow Services and Applications - Whitepaper". 1996. Disponível em: <http://ehealth-spectrum.ca.com/download/netflow.pdf>.

[22] TRIVEDI, K. *Probability and statistics with reliability, queuing, and computer science applications*. Prentice-Hall, 1982.

[23] MAKRIDAKIS, S., WHEELWRIGHT, S., HYNDMAN, R. *Forecasting: Methods and Applications*. 3 ed. New York, USA, John Willey & Sons, Inc., 1998. ISBN: 0-471-53233-9.

[24] ABRAHAM, B., LEDOLTER, J. *Statistical methods for forecasting*. Wiley, 1983. ISBN: 9780471867647.

[25] WILLINGER, W., TAQQU, M. S., SHERMAN, R., et al. "Self-similarity Through High-variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level". In: *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '95, pp. 100–113, New York, NY, USA, 1995. ACM. ISBN: 0-89791-711-1. doi: 10.1145/217382.217418.

[26] CROVELLA, M. E., BESTAVROS, A. "Self-similarity in World Wide Web traffic: evidence and possible causes", *IEEE/ACM Transactions on Networking*, v. 5, n. 6, pp. 835–846, 1997.

[27] A.J. VAN ELBURG, R., VAN OOYEN, A. "A new measure for bursting", *Neurocomputing*, v. 58-60, pp. 497–502, jun. 2004. ISSN: 09252312. doi: 10.1016/j.neucom.2004.01.086.

[28] AKAIKE, H. "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, v. 19, 1974. ISSN: 0018-9286. doi: 10.1109/TAC.1974.1100705.

[29] BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. *Time series analysis: forecasting and control*. 2008. ISBN: 9780470272848.

[30] GROSCHWITZ, N. K., POLYZOS, C. "A Time Series Model of Long-Term NSFNET Backbone Traffic", *IEEE International Conference on Communications*, v. 3, pp. 1400–1404, 1994.

[31] MORI, T., KAWAHARA, R., NAITO, S., et al. "On the characteristics of Internet traffic variability: spikes and elephants", *CCECE 2003 - Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436)*, pp. 99–106, 2004. doi: 10.1109/SAINT.2004.1266104.

[32] LELAND, W. E., TAQQU, M. S., WILLINGER, W., et al. "On the Self-similar Nature of Ethernet Traffic". In: *Conference Proceedings on Communications Architectures, Protocols and Applications*, SIGCOMM '93, pp. 183–193, New York, NY, USA, 1993. ACM. ISBN: 0-89791-619-0. doi: 10.1145/166237.166255.

[33] WISITPONGPHAN, N., PEHA, J. M. "Effect of TCP on self-similarity of network traffic", *Proceedings. 12th International Conference on Computer Communications and Networks (IEEE Cat. No.03EX712)*, pp. 370–373, 2003. doi: 10.1109/ICCCN.2003.1284196.

[34] FENG, H., SHU, Y. "Study on network traffic prediction techniques", *Proceedings. 2005 International Conference on Wireless Communications, Networking and Mobile Computing, 2005.*, v. 2, n. 3, pp. 995–998, 2005. doi: 10.1109/WCNM.2005.1544219.

[35] HOLANDA FILHO, R., MAIA, J. E. B. "Network traffic prediction using PCA and K-means". In: *Network Operations and Management Symposium (NOMS), 2010 IEEE*, pp. 938–941, April 2010. doi: 10.1109/NOMS.2010.5488338.

[36] BRANCH, P. A., CRICENTI, A. L., ARMITAGE, G. J. "An ARMA(1,1) prediction model of first person shooter game traffic", *Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing, MMSP 2008*, pp. 736–741, 2008. doi: 10.1109/MMSP.2008.4665172.

[37] VILELA, G. S. *Caracterização de Tráfego Utilizando Classificação de Fluxos de Comunicação*. Dissertação de Mestrado, COPPE/UFRJ, 2006.

[38] BROWNLEE, N., CLAFFY, K. "Understanding Internet Traffic Streams : Dragonflies and Tortoises", *IEEE Communications Magazine*, v. 40, n. 10, pp. 110–117, Outubro 2002. ISSN: 0163-6804. doi: 10.1109/MCOM.2002.1039865.

[39] MATTA, I. "The war between mice and elephants". In: *Ninth International Conference on Network Protocols, 2001*, pp. 180–188. IEEE Comput. Soc, 2001. ISBN: 0-7695-1429-4. doi: 10.1109/ICNP.2001.992898.

[40] CORTEZ, P., RIO, M., ROCHA, M., et al. "Multi-scale Internet traffic forecasting using neural networks and time series methods", *Expert Systems*, v. 29, n. 2, 2012. ISSN: 02664720. doi: 10.1111/j.1468-0394.2010.00568.x.

[41] BABIARZ, R., BEDO, J.-S. "Internet Traffic Mid-term Forecasting : A Pragmatic Approach Using Statistical Analysis Tools". In: *Proceedings of the 5th International IFIP-TC6 Conference on Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, pp. 110–122, 2006.

[42] HU, K., SIM, A., ANTONIADES, D., et al. "Estimating and Forecasting Network Traffic Performance Based on Statistical Patterns Observed in SNMP Data". In: *Proceedings of the 9th International Conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM'13, pp. 601–615, Berlin, Heidelberg, 2013. Springer-Verlag. ISBN: 978-3-642-39711-0. doi: 10.1007/978-3-642-39712-7_46.

[43] TOPKE, C. R. *Uma Metodologia para Caracterização de Tráfego e Medidas de Desempenho em Backbones IP*. Dissertação de Mestrado, COPPE/UFRJ, 2001.

[44] OETIKER, T. "MRTG: The Multi Router Traffic Grapher". In: *Proceedings of the 12th Conference on Systems Administration (LISA-98), Boston, MA, USA, December 6-11, 1998*, pp. 141–148, 1998. Disponível em: <`http://www.usenix.org/publications/library/proceedings/lisa98/oetiker.html`>.

[45] HYNDMAN, R., KOSTENKO, A. V. "Minimum Sample Size requirements for Seasonal Forecasting Models", *Foresight: The International Journal of Applied Forecasting*, , n. 6, pp. 12–15, 2007.

[46] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. Disponível em: <`http://www.R-project.org/`>. ISBN 3-900051-07-0.

[47] VENABLES, W. N., RIPLEY, B. D. *Modern Applied Statistics with S*. Statistics and Computing. Springer, 2002. ISBN: 9780387954578.

[48] DUNN, P. F. *Measurement, Data Analysis, and Sensor Fundamentals for Engineering and Science*. Taylor & Francis Group, 2011. ISBN: 9781439875292.

[49] ADAS, A. "Traffic models in broadband networks", *IEEE Communications Magazine*, v. 35, n. 7, pp. 82–89, 1997. ISSN: 0163-6804. doi: 10.1109/35.601746.

[50] ZHANG, C., SUN, S., YU, G. "A Bayesian network approach to time series forecasting of short-term traffic flows", *International IEEE Conference on Intelligent Transportation Systems Conference*, pp. 216–221, 2004. doi: 10.1109/ITSC.2004.1398900.

[51] DA SILVA, V. L. P. *Identificação de Anomalias em Fluxos de Rede Utilizando Previsões em Séries Temporais pelo Método de Holt-Winters*. Dissertação de Mestrado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2015.

[52] DA SILVA FILHO, J. B. *Detecção de Anomalias em Fluxos de Redes de Computadores Utilizando Técnicas de Redes Neurais e Estimadores Lineares*. Dissertação de Mestrado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2015.

[53] DE ABREU SILVA, C. *Proposta e Implementação de uma Ferramenta para Gerência de Segurança em Redes Baseada numa Nova Metodologia Usando Análise de Tráfego em Backbones IP*. Dissertação de Mestrado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2006.

[54] KLEINROCK, L. *Queueing Systems*, v. I: Theory. Wiley Interscience, 1975.

[55] HYNDMAN, R. J., ATHANASOPOULOS, G. *Forecasting : Principles & Practice*. N. September. 2013. ISBN: 9780987507105.

[56] COTTRELL, A., LUCCHETTI, R. "Gretl Software - Gnu Regression, Econometrics and Time-series Library". Disponível em: <`http://gretl.sourceforge.net/`>. Acessado em 17/10/2014.

[57] LJUNG, G. M., BOX, G. E. P. "On a Measure of Lack of Fit in Time Series Models", *Biometrika*, v. 65, n. 2, pp. pp. 297–303, 1978. ISSN: 00063444.

[58] DING, X., CANU, S., DENOEUX, T., et al. "Neural Network Based Models For Forecasting". In: *Proceedings of ADT'95*, pp. 243–252. Wiley and Sons, 1995.

[59] MATHWAVE TECHNOLOGIES. "EasyFit". 2004-2012. Disponível em: <`http://www.mathwave.com/easyfit-distribution-fitting.html`>.

[60] FLYNN, M. R. "Fitting human exposure data with the Johnson SB distribution", *Journal of Exposure Science and Environmental Epidemiology*, v. 16, n. 1, pp. 56–62, 2006. doi: 10.1038/sj.jea.7500437.

[61] BROCKWELL, P. J., DAVIS, R. A. *Introduction to time series and forecasting*. 2002. ISBN: 0387953515.

[62] CAIDA. "CAIDA: Center for Applied Internet Data Analysis". 2015. Disponível em: <`http://www.caida.org`>. Acessado em Agosto de 2015.

[63] GROSSGLAUSER, M., BOLOT, J.-C. "On the Relevance of Long-range Dependence in Network Traffic", *IEEE/ACM Transactions on Networking*, v. 7, n. 5, pp. 629–640, Outubro 1999. ISSN: 1063-6692. doi: 10.1109/90.803379.

[64] Geoff Huston. "BGP Routing Table Reports". `http://bgp.potaroo.net/`, 2013. Acessado em Agosto de 2015.

[65] NUNES, B. A. A., MENDONCA, M., NGUYEN, X.-N., et al. "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks", *Communications Surveys Tutorials, IEEE*, v. 16, n. 3, pp. 1617–1634, Third 2014. ISSN: 1553-877X. doi: 10.1109/SURV.2014.012214.00180.

[66] YASSINE, A., RAHIMI, H., SHIRMOHAMMADI, S. "Software-Defined Network Traffic Measurement: Current Trends and Challenges", *Instrumentation Measurement Magazine, IEEE*, v. 18, n. 2, pp. 42–50, April 2015. ISSN: 1094-6969.