UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP
Faculdade de Tecnologia - FT

Mestrado em Tecnologia e Inovação

Modelagem de Redes de Computadores por Métodos Estatísticos

Dissertação de Mestrado

Renata Lussier Spagnol

Limeira - SP

Dezembro/2011

11 2011

UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP
Faculdade de Tecnologia - FT

Renata Lussier Spagnol

Modelagem de Redes de Computadores por Métodos Estatísticos

Dissertação apresentada ao Curso de Mestrado da Faculdade de Tecnologia da Universidade Estadual de Campinas, como requisito para a obtenção do título de Mestre em Tecnologia.

Área de Concentração: Tecnologia e Inovação

Orientador: Dr. André F. de Angelis

Coorientadora: Dra. Laura L. R. Rifo.

Limeira - SP

Dezembro/2011

FICHA CATALOGRÁFICA ELABORADA POR SILVANA MOREIRA DA SILVA SOARES – CRB-8/3965
BIBLIOTECA UNIFICADA FT/CTL
UNICAMP

Sp13m

Spagnol, Renata Lussier, 1985-
Modelagem de redes de computadores por métodos estatísticos / Renata Lussier Spagnol. – Limeira, SP : [s.n.], 2011.

Orientador: André Franceschi de Angelis.
Coorientador: Laura Letícia Ramos Rifo.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Controle de processo – Métodos estatísticos. 2. Internet. 3. Redes de computadores. I. Angelis, André Franceschi de. II. Rifo, Laura Letícia Ramos. III. Universidade Estadual de Campinas. Faculdade de Tecnologia. IV. Título.

Informações para Biblioteca Digital

Título em inglês: Modeling of computer networks by statistical methods

Palavras-chave em inglês (Keywords):

1- Internet
2- Computer networks
3- Statistical process control
4- ARIMA

Área de concentração: Tecnologia e Inovação

Titulação: Mestre em Tecnologia

Banca examinadora: André Franceschi de Angelis, Carlos Antônio Ruggiero e Edson Luiz Ursini

Data da Defesa: 09-12-2011

Programa de Pós-Graduação em Tecnologia

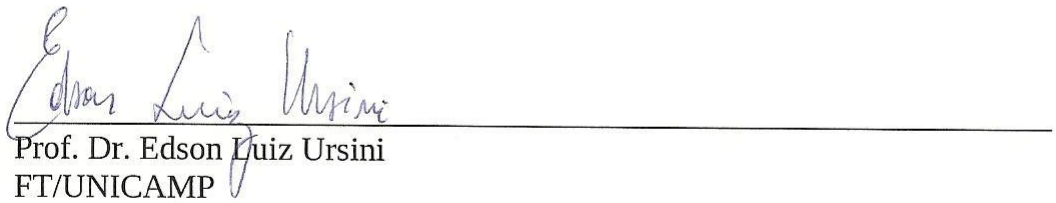
DISSERTAÇÃO DE MESTRADO ACADÊMICO

Modelagem de Rede de Computadores por Métodos Estatísticos

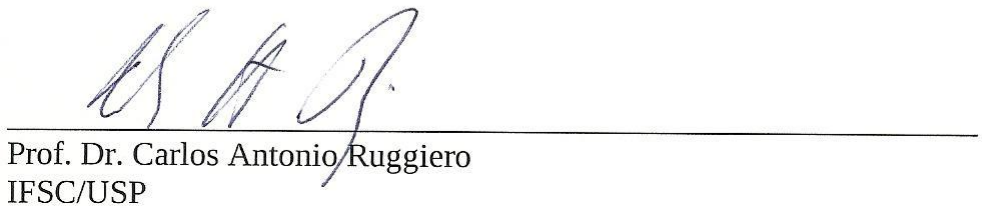
Autor: Renata Lussier Spagnol

A Banca Examinadora composta pelos membros abaixo aprovou esta Dissertação:

Prof. Dr. André Franceschi de Angelis
FT/UNICAMP



Prof. Dr. Edson Luiz Ursini
FT/UNICAMP



Prof. Dr. Carlos Antonio Ruggiero
IFSC/USP

Agradecimentos

Agradeço primeiramente a Deus por estar viva e com saúde e disposta a novos desafios.

Agradeço aos meus pais por todo apoio e dedicação, aos anos investidos com tanto amor e confiança. O fato de eu ser uma pessoa que gosta de estudar e dedicada a tudo que faço vem da educação passada por eles.

Agradeço ao meu orientador André, primeiramente pela oportunidade de acrescentar outras técnicas a sua tese. Agradeço também, por todo conhecimento passado, em especial em redes de computadores do qual não tinha antes contato, aos vários dias dedicados as orientações, sempre paciente e muito dedicado.

Agradeço a minha coorientadora Laura, pelos vários anos de trabalho dedicados ao meu desenvolvimento, desde a graduação, sempre ressaltando a importância de se ter dar continuidade aos estudos e de fazer uma pós-graduação independente da área. Neste ano, mesmo longe, ela conseguiu estar tão perto.

Agradeço também ao Prof. Dr. Emanuel Pimentel pela sua disposição em auxiliar nos gráficos de CEP, a sua ajuda foi muito válida.

Por fim agradecer as minhas amigas que entenderam todos aos meus não aos nossos encontros, ao meu namorado pelo apoio aos finais de semana em casa para estudar e a ajuda de alguns colegas que hoje se tornaram amigos.

A dúvida é o princípio da sabedoria.

Aristóteles

Se queres prever o futuro, estuda o passado.

Confucio

Abreviaturas

AIC - *Akaike Information Criterion* (Critério de Informação Akaike)

AR - Administrador de Rede de Computadores

AR(p) - Processo Autorregrassivo

ARFIMA (p,d,q) - Processo Autorregrassivo Integrado Fracinonário Média Móvel

ARIMA (p,d,q) - Processo Autorregrassivo Integrado Média Móvel

ARMA(p,q) - Processo Autregressivo Média Móvel

ATM - *Asynchronous Tranfer Mode* (Modo de Transferência Assíncrono)

BIC - *Bayesian Information Criterion* (Critério de Informação Baysiano)

CEP - Controle Estatístico do Processo

CUSUM - Soma Cumulativa

DoS - *Denial of Service*

IP - Protocolo de *Internet*

ISO - *International Organization for Standardization* (Organização Internacional para Padronização)

LAN - *Local Area Network* (Rede Local)

MA(q) - Processo Média Móvel

MMEP - Média Móvel Exponencialmente Ponderada

mrCEP - Modelo de Cartas de Controle em Angelis [1]

OSI - *Open Systems Interconnection* (Interconexão de Sistemas Abertos)

RC - Rede de Computadores

TCP - *Transmission Control Protocol* (Protocolo de Controle de Transmissão)

UDP - *User Datagram Protocol* (Protocolo de Transmissão de Dados)

WAN - *Wide Area Network* (Área Ampla de Rede)

Resumo

A sociedade atual é dependente das Redes de Computadores para seu cotidiano e, portanto, mantê-las em boas condições de operação é essencial. Reagir aos problemas é uma estratégia que implica em degradação ou interrupção da rede e incorre geralmente em altos custos. É preferível detectar antecipadamente os problemas e corrigí-los proativamente, o que implica no uso de técnicas preditivas para controle, tais como os métodos estatísticos. Este trabalho determinou a possibilidade de se avaliar a rede com um menor número de variáveis em relação a um modelo existente e apontou maneiras de aprimorar a qualidade do monitoramento com uso técnicas estatísticas mais recentes e menos usuais. Os experimentos realizados consistiram-se na análise de traços de uma rede real previamente armazenados em bases de dados, sobre os quais foram aplicados cálculos de coeficiente de correlação linear para redução de variáveis. Ajustou-se um modelo para a rede com métodos de análises de Séries Temporais e foram testadas as cartas de Soma Acumulativa (CUSUM) e de Média Móvel Exponencialmente Ponderada (MMEP) em substituição às de média e amplitude. Obteve-se uma redução inicial de 23 para 4 na quantidade de variáveis a monitorar estatisticamente, com possibilidade de se chegar a uma única medida, simplificando os processos de controle da rede. Foi possível ajustar um Modelo Autorregressivo Integrado Média Móvel (ARIMA) para a rede e monitorá-la através de cartas CUSUM e MMEP, demonstrando-se a última mais adequada ao problema.

Palavras-chaves: Internet, redes de computadores, controle estatístico do processo(CEP), ARIMA.

Abstract

The nowadays society depends on computer networks for its daily activities and, therefore, it is essential to keep them in good operation conditions. React to the problems is a strategy that implies the network degradation or its interruption and increases maintenance costs. It is preferable the early detection of the problems and its proactive correction. This approach implies in the use of control prediction techniques, as stochastic methods. The present work has showed that the use of recent and less common statistics techniques can enhance the monitoring quality of the networks with fewer variables than a previous model. The linear correlation coefficient method was employed for the reduction of the number of variables over previously data base stored network traces. It was performed a model adjustment for the network using the temporal series method. The Cumulative Sum control chart (CUSUM) and the Exponentially Weighted Moving Average (EWMA) were used in replacement of common charts of average and range. It was obtained an initial reduction from 23 to 4 in the statistical monitored variables and it is possible to reach only one measure in some conditions, simplifying the network control process. It was possible to adjust an Autoregressive Integrated Moving Average (ARIMA) to the network and monitor it through CUSUM and EWMA. The last one was demonstrated to be the most suitable to the problem.

Keywords: Internet, computer networks, statistical process control and ARIMA

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

As redes de computadores (RCs) têm sido utilizadas cada vez mais no cotidiano das pessoas e das organizações em todo o mundo. O crescimento rápido e sustentado da Internet, a grande rede mundial que interliga inúmeras redes, adiciona a cada instante novos usuários e novos serviços ao mundo virtual. Esse processo trouxe evidentes ganhos à sociedade, desde o acesso global a dados até as comodidades do comércio eletrônico e notícias em tempo real. Por outro lado, implica necessariamente em uma forte e crescente dependência tecnológica da sociedade em relação às RCs. Para ilustrar esta idéia de dependência, bastam exemplos como serviços bancários, de telecomunicações, de controle e informação, correio eletrônico, navegação na Internet, etc.

Verificada essa dependência, tem-se como consequência direta que a indisponibilidade ou mesmo a degradação das RCs pode trazer prejuízos de grande peso à sociedade. Prejuízos que podem variar desde nível global, como perda de conectividade e serviços da Internet, até o nível mais local, ilustrados pela dificuldade de acesso e manutenção de dados de organizações e usuários finais. Apesar disto, manter as boas condições operacionais das RCs é uma questão ainda em aberto, que continua a motivar desde as pesquisas acadêmicas até os esforços da indústria de tecnologia da informação.

As RCs estão sujeitas a uma série de problemas nas diversas fases do seu ciclo de vida. Neste trabalho, foram considerados aqueles problemas relativos à fase de operação da rede, deixando-se de abordar, por exemplo, questões advindas de projetos deficientes.

Ainda assim, interrupções, congestionamentos, lentidão e outros percalços podem ter causas naturais como desgastes dos equipamentos, acidentais, como rupturas de cabos, ou ações intencionais, como é o caso de ataques contra a rede e seus componentes. Este amplo escopo, aliado à multiplicidade de tecnologias de redes, torna muito complexo o trabalho do administrador de rede (AR), considerado como a figura responsável pelo projeto, implementação e operação da RC.

Se um problema ocorre com a rede, o AR deve resolvê-lo no menor tempo possível, dentro de custos compatíveis com o sistema. Conforme o caso, são necessárias paradas para reconfiguração da RC. Assim, a abordagem corretiva ou reativa apresenta-se como uma solução de alto custo, incapaz de eliminar completamente a degradação da RC. É mais interessante prever a ocorrência do problema, pois, assim, o AR tem condições melhores para, numa abordagem proativa, evitar a materialização do prejuízo, a um custo operacional reduzido. É evidente que as abordagens não se excluem e, em um cenário ideal, o AR deveria agir sempre proativamente, reservando ações reativas para as situações em que não houve sucesso na prevenção.

Em paralelo, é importante ressaltar que praticamente todas as RCs são projetadas e construídas em camadas ou níveis, cada qual envolvendo serviços e recursos específicos, tal como as soluções descritas nas arquiteturas ISO/OSI e TCP/IP. Este conceito de camadas permite que soluções sejam focadas em níveis específicos, reduzindo seu escopo e trazendo a questão de manutenção das RCs a patamares administráveis. Este trabalho tem seu interesse no tráfego das RCs, correspondente à Camada de Rede das arquiteturas citadas.

A abordagem proativa, comentada anteriormente, tem aplicação em outros contextos que não as RCs. Com efeito, as indústrias de manufatura utilizam rotineiramente ferramentas de Controle Estatístico do Processo (CEP) para monitorar suas linhas de produção. É justamente o CEP que permite às indústrias corrigir problemas da produção antes que seus efeitos se manifestem nas unidades manufaturadas. Logo, aplicar CEP a RCs dentro de um escopo adequadamente escolhido torna-se uma investigação interessante e com bom potencial de resultados, mesmo consideradas as diversidades conceituais e as características específicas de produtos industriais (físicos) e serviços de rede (vir-

tuais). Nesta direção, foi encontrado em Angelis [1] um modelo RCs para aplicação de CEP, referenciado como mrCEP por brevidade. O mrCEP usava uma base de dados composta de pouco menos de trinta variáveis que devem ser monitoradas pelo AR e utilizava cartas de controle (gráficos) semelhantes aos adotados pela indústria de manufatura. Estas cartas estavam devidamente adaptadas às características da distribuição estatística dos dados das RCs, onde foi verificada a não-normalidade dos dados. Estas adaptações, embora corretas do ponto de vista matemático, implicaram em alguma perda de sensibilidade no controle e no monitoramento das redes. Trabalhos posteriores foram capazes de, na mesma metodologia CEP, aumentar a sensibilidade do mrCEP, trazendo-a em ponto bastante satisfatório, com baixa ocorrência de falsos-positivos, Yokoyama[25]. Porém, nenhuma tentativa foi feita para aumentar a relevância das informações através, por exemplo, de uso de medidas e cartas diversas das utilizadas do ambiente industrial, que levassem em consideração as particularidades das RCs. Outro ponto a considerar é que o CEP, na indústria, costuma monitorar um pequeno conjunto de variáveis simultaneamente, por exemplo, diâmetro e altura de uma peça. Raramente se pede a um operador o acompanhamento de mais que três ou quatro indicadores. O mrCEP, por outro lado, sugere quase três dezenas de variáveis para monitorar as RCs e, ainda que se considere a completa automação do processo, esse número representa dificuldades operacionais para os ARs.

Esta pesquisa se propôs investigar e aperfeiçoar o mrCEP, dados os pontos anteriormente mencionados, e seguiu em duas linhas. Na primeira, foi analisada a possibilidade de uma redução do número de variáveis a monitorar, sem perda dos resultados do mrCEP. Para tanto, variáveis suspeitas de guardarem entre si alta correlação foram comparadas e, confirmada a suposição, foram tratadas como duplicadas, eliminando a necessidade do monitoramento de ambas. Na análise dos dados, a correlação teve bons resultados, diminuindo realmente o número de variáveis. A segunda linha investigativa seguiu pela melhoria da informação disponível. Buscou-se responder se cartas de controle diferentes das utilizadas originalmente poderiam prover um entendimento mais profundo das RCs e de seu estado de operações. Em particular, utilizou-se substituição das cartas de média e

amplitude, que têm baixa memória temporal, por medidas como média móvel exponencialmente ponderada e soma acumulativa, capazes de aproveitar melhor a informação das séries temporais e estudar a autocorrelação da série. As novas cartas se mostraram mais coerentes para o tipo de dados analisados.

Este trabalho utilizou o mesmo banco de dados de tráfego/traços de rede que validaram o mrCEP como fonte primária, não necessitando, portanto, de uma fase de coleta de dados ou monitoramento da rede. Assim, foi possível estabelecer comparações significativas entre os trabalhos.

A próxima seção apresenta a Revisão Bibliográfica pertinente ao tema, seguida por Bases Teóricas onde revisaram-se os métodos estatísticos utilizados. Devido à complexidade envolvida, este capítulo relaciona os métodos vistos em função das diferenças entre as áreas de redes (formação do AR) e estatística (foco desta dissertação). O capítulo de Materiais e Métodos mostra em detalhes os recursos de software utilizados e aspectos relevantes da base de dados. Em seguida, em Análise Estatítica da Rede, contempla-se desde a análise descritiva dos dados, os modelos de previsão e forma de monitoramento até os resultados obtidos. Em Discussão dos Resultados aponta-se a diferença entre os modelos ajustados. Na Conclusão são expostos os principais pontos abordados no estudo e seus desfechos. O texto é finalizado com as Referências Bibliográficas da pesquisa.

Capítulo 2

Revisão Bibliográfica

Em busca de oferecer um serviço de melhor qualidade para os usuários de serviços de RC e facilitar o monitoramento do AR, na década de 2000 intensificou-se o estudo do comportamento do tráfego. Os estudos procuram evitar interrupções das RCs, sejam elas causadas por quedas ou ataques de intrusos. A seguir estão exemplificados esses estudos, sendo que a maioria utiliza algum método estatístico como ferramenta principal para o entendimento da rede.

Um dos meios para melhorar a qualidade da rede é oferecer um serviço mais seguro, resistente a ataques. A prevenção e identificação de ataques são assuntos sempre em pauta pelas empresas relacionadas com a segurança de serviços da Internet. Existe uma preocupação com extravio de informações confidenciais dos usuários ou mesmo pelo desconforto de um usuário que ao acessar um *site* e baixar um arquivo ou abrir seu *email*, acabe contraindo algum vírus. Em Creeger [9], alguns membros de empresas especializadas e importantes no ramo de segurança de *software*, discutem a necessidade de, cada vez mais, entender como se comportam os ataques e como identificá-los antes que eles venham a prejudicar seus usuários. Os ataques estão sendo cada vez melhor elaborados e esse desenvolvimento dificulta a sua identificação.

Um ataque pode se manifestar de várias maneiras. Uma delas é mediante um congestionamento intencional na rede. Alguns métodos de defesa procuram detectar esse tipo

de ataque, procurando entender e controlar o volume de tráfego ou mesmo detectando se houve diferenciação dos padrões do fluxo. Para esses entendimentos e monitoramentos podem ser utilizados métodos mais simples ou sofisticados para oferecer um serviço com segurança. Em Xuan [24], é proposto um método sofisticado, onde se usa um teste que permite a detecção rápida e precisa de ataques por meio dos métodos clássicos de grupo de teste (GT). Para a construção do método, foi necessário regular as solicitações de serviço para corresponder à matriz do sistema e assim estabelecer limites adequados para o servidor fonte (indicador de uso), para gerar resultados exatos dos testes. Os algoritmos de detecção de decodificação do GT se mostraram uma boa opção para identificar ocorrência de ataques DoS (*Deny of Service*) e até em outros tipos de ataque na rede, gerando no final uma baixa taxa de falsos positivos e negativos.

Nos primeiros estudos sobre qual seria o modelo estatístico indicado para estimar o comportamento da rede, acreditava-se que o tráfego da rede seguia um modelo estatístico de Poisson. Em Paxson[17], nota-se que esse modelo não é o mais adequado. Algumas características dos dados de rede infringem pré-requisitos do modelo. Guiesi[12] diz que o modelo Poissoniano funciona bem para modelar chegadas em seções, como acontece com os tráfegos TELNET e o FTP. A Internet e as redes Ethernet são estruturas auto-similares, com alta variabilidade e bom ajuste de distribuições de cauda pesada.

Becchi [2] também relata que os primeiros modelos de tráfego foram baseados em processos de Poisson. E foram inicialmente utilizados no contexto dos serviços de telefonia, onde as chegadas de chamada podem ser consideradas independentes e identicamente distribuídas. O modelo de Poisson não se mostrou adequado para descrever os dados de tráfego em LANs e WANs modernos, reafirmando a teoria de Paxson [17]. Em Becchi [2] foi comprovado que existe correlação entre os pacotes de dados de um fluxo de rede, características de auto-similaridade e cauda pesada. O mesmo texto apresenta ainda as limitações do modelo de Poisson e explica como o modelo de autossimilaridade difere dos tradicionais.

Rutka[18] constatou que os dados de fluxo de rede possuem propriedades de longa dependência, são altamente correlacionadas e, devido a esta característica, a série deve

ser tratada com métodos específicos. Sugere como método estatístico a aplicação de modelos autorregressivos de média móvel (ARIMA) para prever futuros tráfegos.

Elagha [11], estudou o tráfego da rede ATM *(Asynchronous Tranfer Mode)* buscando encontrar algumas particularidades do tipo de série. A princípio, tinha-se a suspeita que os dados possuíam características de longa dependência, auto-similaridade e/ou estrutura fractal ou multifractal. O trabalho analisou os dados de rede ATM da Universidade do Mediterrâneo Oriental (UEM), para descobrir o grau de autossimilaridade. Não foram encontrados autossimilaridades em alguns pontos, mas as observações apresentaram características de estruturas de fractais ou multifractais. Para a estrutura dos dados, foi proposto a aplicação de um modelo de previsão ARIMA.

Um ataque pode ser classificado também como uma anomalia e esta pode ser uma indicação de uma situação perigosa na rede ou em outros sistemas. A mineração de dados é usada como método em Isakssont [13], para identificar as tentativas ou ataques sobre um sistema de computadores ou na rede. Tem-se como objetivo identificar as anomalias da rede, por meio da mudança de comportamento dos dados. O modelo proposto por ele detecta os problemas baseando-se na frequência de ocorrência e, em seguida, mede o desvio da distância de cada anomalia no espaço dos dados. O nível de risco com o qual a anomalia está associada é avaliado pelo desvio entre os dados históricos sem avaria e o previsto pelo modelo. O modelo estatístico foi construído com base no Modelo *Markov Extensible* (EMM), que é baseado em uma técnica de modelagem espaço-temporal, e trabalha com base nas ocorrências das transições de estados na cadeia de *Markov* gerada. Foram incluídas no estudo técnicas de *clustering*, associadas às cadeias de Markov. O trabalho inicialmente separa o que é um valor que possui um comportamento sem problema na rede de um valor que não pertence a nenhum grupo existente. Faz isso verificando a cardinalidade do *cluster* associado: se é pequena, entende-se como uma anomalia.

Os estudos acima são importantes para o entendimento do comportamento da rede. E este entendimento ajuda na previsão de congestionamentos, no delineamento da tendência dos dados e auxilia o AR nas tarefas de planejamento. Além da dificuldade de compreender o comportamento da rede, que é bem variado, outro ponto que dificulta seu

monitoramento é que geralmente têm-se diversas variáveis que influenciam o tráfego e, para criar um modelo que o represente bem, é necessário escolher as mais significativas. Em Silva [19], é aplicada estatística multivariada na análise e previsão do tráfego da rede. O estudo sugere a utilização de análise de componentes principais para modelar o fluxo da rede e sua tendência, usando um número mínimo de variáveis. Para a redução desse número existem alguns métodos estatísticos, por exemplo, retirando do modelo as variáveis correlacionadas.

Em Angelis [1], estudou-se um método para monitorar as redes de computadores e assim prever futuras interrupções por meio de cartas de Controle Estatístico de Processo (CEP). O estudo trata de uma nova visão sobre a questão de gerenciamento da rede, unindo o conhecimento a priori dos fluxos às técnicas de CEP, com a intenção de antecipar as ações corretivas de maneira que os problemas sejam detectados e sanados antes de sua efetiva ocorrência. No estudo, observou-se a rede local do Instituto de Física de São Carlos (IFSC/USP). O banco de dados contempla 23 variáveis que descrevem grandezas associadas ao tráfego presente na rede. As variáveis escolhidas para o estudo são: número de *bytes*, considerando as combinações possíveis entre origem e destino em relação à rede observada (interna ou externa), de pacotes e dos fluxos do protocolo IP e outros e de protocolos de transporte TCP e UDP. As cartas foram construídas para as variáveis julgadas como significativas pelo autor. Uma maneira mais indicada para a escolha das variáveis é conciliar a informação inicial do autor com a análise de correlação linear entre as variáveis, pois com essa análise, ele conseguiria otimizar seu estudo, evitando monitorar ao mesmo tempo variáveis que estejam correlacionadas, visto que, se ocorrer uma mudança de comportamento em uma, este se refletirá na outra. O ideal seria reduzir o número de variáveis em pelo menos 50 por cento, usando no máximo uma dezena.

Outro ponto observado em Angelis [1] é que os dados não seguem uma distribuição normal. Sabendo que o modelo Gaussiano não se ajusta adequadamente aos dados, é aconselhável não utilizar a construção dos limites dos gráficos $\bar{X}$ e R do modo convencional. Devido a isso, o autor construiu os limites dos gráficos $\bar{X}$ e R por meio da desigualdade de Chebyshev.

Um método indicado para permitir a determinação dos limites é a utilização de gráficos de Média Móvel Exponencialmente Ponderada e de Soma Cumulativa (Montgomery [14]), já que estes gráficos suportam dados livres de distribuição, ou seja, não precisam ter o comportamento de uma distribuição conhecida, como por exemplo a normal.

Baseado no estudo de Angelis [1], Yokoyama[25] estudou a fundo os métodos de CEP para gerar cartas mais sensíveis às variações do tráfego da rede, sem considerar a possível redução da quantidade das variáveis que deveriam ser monitoradas. Inicialmente, utilizou limites variáveis para os gráficos de atributos, porém isso não foi satisfatório, já que esses gráficos são utilizados para a construção de itens com a classificação 'conformes' ou 'não conforme'. Essa classificação, segundo seus estudos, é impraticável para um tráfego de uma RC. Analisando outros tipos de gráficos de CEP, Yokoyama[25], optou pelos gráficos $\bar{X}$ e R, sendo que $\bar{X}$ estima a média e R a amplitude dos dados. Para estimar o desvio entre observações tem-se o Gráfico S, porém o cálculo para esse gráfico é mais trabalhoso que o R.

Um problema encontrado por Angelis [1] foi que os dados não seguiam uma distribuição normal. A amostragem dos dados poderia ser feita de duas maneiras: tomar-se amostras pequenas e frequentes ou tomar-se amostras grandes e pouco freqüentes. Yokoyama[25] optou pelo primeira maneira. Nesse caso, por meio do Teorema do Limite Central, os dados deveriam convergir para uma distribuição se fossem variáveis independentes com mesma distribuição e variância finita, mas não se verificou essa convergência. A maneira adotada para minimizar o efeito de não-normalidade foi a construção de limites variáveis. Assim, conseguiu que os limites acompanhassem as flutuações do tráfego com boa sensibilidade, sendo cortados somente em situações aparentemente atípicas do comportamento da rede. Para construir limites variáveis, foi necessário alterar a maneira como eram calculados os limites em Angelis [1].

Como observado anteriormente, para a utilização do CEP no monitoramento de variáveis características de um processo, existem algumas condições necessárias para a sua implementação. Moreira [15], diz que uma delas é que as observações devam ser independentes.

Alguns processos possuem observações que apresentam autocorrelação. Para esse caso,

as ferramentas do CEP como $\bar{X}$ e R não funcionam corretamente, podendo resultar em 'falsos alarmes'. Moreira [15] sugere que, no caso de informações autocorrelacionadas, é necessário tratá-las antes e depois controlá-las estatisticamente. Como forma de tratamento, sugere primeiramente fazer um ajuste de um modelo de série temporal ARIMA, que remova a autocorrelação dos dados. Daqui, se obtem-se resíduos independentes e pode-se aplicar um gráfico de controle

Uma sugestão para monitorar os resíduos é usar os gráficos de controle de Soma Acumulativa (CUSUM) ou a Média Móvel Exponencialmente Ponderada (MMEP), já que estes detectam pequenas mudanças na média do processo. Segundo Moreira [15], existem alguns estudos que propõem utilizar os gráficos de MMEP e CUSUM para dados autocorrelacionados sem ajuste prévio de um modelo e estes obtiveram resultados satisfatórios.

Anteriormente a esta pesquisa, realizou-se no ano de 2009 um estudo prévio sobre o comportamento da rede Spagnol[20]. Utilizaram-se dados de uma rede (LAN) disponíveis de acordo com Cowperwait [7], e obtiveram-se resultados satisfatórios quanto ao monitoramento ARFIMA. Foram observadas evidências de que os dados de um fluxo de rede são autocorrelacionados e possuem uma grande variabilidade, sugerindo um modelo. Aplicou-se primeiramente o modelo autorregressivo fracionário integrado de média móvel (ARFIMA) para as observações e depois as mesmas foram monitoradas pelos gráficos de CEP (CUSUM e MMEP). Os resultados obtidos foram satisfatórios, sendo possível ajustar um modelo que absorvesse as necessidades dos dados e gráficos de CEP capazes de captar as interrupções. Um ponto importante é que os dados apresentavam apenas uma variável. Portanto, nesse caso, não é preciso escolher variáveis, como é a questão de parte dos estudos mostrados anteriormente.

Este trabalho propôs diminuir o número de variáveis a serem monitoradas, ajustar um modelo estatístico e utilizar gráficos de CEP para controlar a rede. Para atingir a proposta, utilizou-se a análise de correlação linear para determinar o grau de dependência entre as variáveis significativas observadas no monitoramento da rede, eliminando as variáveis altamente correlacionadas. Para ajustar um modelo estudou-se a metodologia de séries temporais. O modelo mais adequado para os dados de rede foi um ARIMA.

Utilizaram-se gráficos de controle, CUSUM e MMEP, para monitorar a série da rede analisada.

O capítulo seguinte descreve resumidamente as Bases Teóricas mais relevantes para esta pesquisa.

Capítulo 3

Bases Teóricas

Este capítulo revisa brevemente as bases conceituais importantes para o trabalho, agrupando-as em Análise de Correlação Linear, Modelos de Séries Temporais e gráficos de CEP, visando a oferecer ao leitor uma referência consistente e focada na teoria que suporta esta pesquisa.

3.1 Análise de Correlação Linear

Considerando duas variáveis aleatórias X e Y, com distribuição conjunta $F(X,Y)$, define-se a esperança de X e denota-se E(X) ou EX, como $E(X) = \int dF(X,Y)$, e a variância de X denota-se $Var(X)$ ou simplesmente $\sigma_X^2 = E((X - EX)^2)$. Define-se a covariância entre X e Y como $Cov(X,Y) = E[(X - EX))(Y - EY))]$.

Assume-se então que $E(X)$, $E(Y)$, $E(X,Y)$, $E(X^2)$ e $E(Y^2)$ existem e são finitos.

Definição 3.1.1 *O coeficiente de correlação é um índice de correlação linear entre duas variáveis, de X e Y, define-se como:*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{3.1.1}$$

são variáveis de X e de Y, respectivamente, Bussab [5].

Propriedades:

- *$\rho(X, Y)$ possui valores entre -1 e 1;*
- *Esse coeficiente é igual a 1 ou -1 se e somente se X e Y estiverem relacionados linearmente, $Y = aX + b$, para $a \neq 0$.*
- *Quando $\rho(X, Y) = 0$, entende-se que X e Y são não correlacionados linearmente, ou simplesmente, não correlacionados.*

Contextualizando com o presente estudo, tem-se que as redes de computadores possuem algumas variáveis que influenciam no seu comportamento e monitorar todas, como já dito, é inviável. Para reduzir o número inicial de variáveis foi avaliada a correlação linear entre elas.

Em Spagnol [21], realizou-se um estudo sobre a correlação entre variáveis do fluxo da RC usada em Angelis [1]. Foi possível reduzir um número satisfatório de variáveis. Porém, ainda era ainda necessário um estudo mais detalhado para certificar quais variáveis poderiam ser retiradas sem perda de efetividade no monitoramento. Esse cuidado foi tomado neste trabalho.

3.2 Modelos de Séries Temporais

Intuitivamente, uma série temporal pode ser qualquer conjunto de observações ordenadas no tempo e que apresenta dependência entre instantes de tempo. A dependência serial é um fator importante para gerar previsões de valores futuros.

Além de prever valores futuros, pode-se também ajustar um modelo probabilístico adequado para a série, analisando a existência de periodicidade.

Uma questão importante para a construção do modelo é selecionar o mais simples, ou seja, busca um número mínimo de parâmetros, de modo que o modelo ainda seja útil para os fins desejados.

Para a escolha do modelo, uma das condições importantes é a estacionariedade da série, isto é, se ela apresenta certas características constantes no tempo. A seguir, tem-se estes conceitos formalizados.

Definição 3.2.1 *Seja T um conjunto arbitrário. Um processo estocástico é uma família Z = Z(t), t ∈ T, Z(t), tal que, para cada t ∈ T, Z(t) é uma variável aleatória, Morettin [16].*

Definição 3.2.2 *Uma classe dos processos estocásticos, chamada de processos estacionários, é baseada na suposição de que o processo está em um determinado estado de equilíbrio estatístico. Um processo estocástico é dito estritamente estacionário se suas propriedades não são afetadas por uma mudança na origem do tempo, ou seja, se a distribuição de probabilidade conjunta associada com as observações $Z_{t1}, Z_{t2}, ..., Z_{tm}$, feitas em qualquer período de tempo $t_1, t_2, ..., tm$ é o mesmo que aquela associada com m observações $Z_{t1+k}, Z_{t2+k}, ..., Z_{tm+k}$ e com $t_1 + k, t_2 + k, ..., t_m + k$. Assim, para um processo discreto para ser estritamente estacionário, a distribuição conjunta de qualquer conjunto de observações deve não ser afetado pela mudança de todos os tempos de observação para a frente (forward) ou para trás (backward) e por qualquer quantidade k, Box & Jenkins [3].*

A média e a variância de um processo estacionário :

$$E(Z_t) = E(Z_{t+k})$$
$$Var(Z_t) = Var(Z_{t+k}) \text{ para todos } t \text{ e } k.$$

A maioria dos procedimentos de análise estatística de séries temporais supõe estacionariedade do processo em questão. Porém, o mais comum é encontrar séries não-estacionárias. Quando se tem casos de não-estacionariedade, é possível algumas vezes aplicar transformações na série, obtendo uma série estacionária. Uma transformação muito usada é a aplicação de diferenças sucessivas na série. Muita vezes se obtém uma estacionariedade aproximada com, no máximo, duas diferenças. Outro ponto é que basta que o processo seja fracamente estacionário e ergótico, não precisa ser estritamente estacionário.

Contextualizando com a RC, para modelar o fluxo de rede necessitou-se de um aprofundamento na metodologia de séries temporais, a fim de encontrar o modelo, mais adequado. Levou-se em consideração a escolha de um modelo aderente e com um número de parâmetros mínimo.

Existem outros modelos estatísticos, além dos de séries temporais, para modelar e analisar dados como, por exemplo, modelo de regressão linear. Porém, para as características que os dados do fluxo apresentaram, o modelo autorregressivo integrado e de média móvel foi o mais indicado. Apesar do modelo escolhido ser um ARIMA, também foi estudado o ARFIMA, já que algumas referências indicavam esse modelo como uma opção para ajustar dados de redes de computadores devido à alta variabilidade e à longa dependência.

Para a análise inicial para a escolha do modelo, observou-se o comportamento da média e verificou-se que havia necessidade de transformação nos dados. Por fim, o modelo ARIMA foi uma parte essencial para a análise da série do estudo.

3.2.1 Processos de Média Móvel

Definição 3.2.3 *Z_t é um processo média móvel de ordem q se*

$$Z_t = a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q} = \theta(B) a_t \tag{3.2.1}$$

onde θ_1, θ_2, ..., θ_q são constantes, Brockwell [4] e a_t é o ruído branco.

Z_t é uma série temporal estacionária. A condição de invertibilidade para um $MA(q)$ é se as raízes da equação característica, $\theta(B) = 0$, estejam fora do círculo unitario, Morettin [16]. Uma explicação mais detalhada sobre a estacionáriedade e invertibilidade pode ser encontrada em Cowperwait [7] ou Wei[23].

Contextualizando com a série do fluxo da rede, ajustou-se um modelo média móvel puro MA(q), onde q indica o número de parâmetros no modelo. Porém o modelo ajustado não é o mais adequado, analisando a média constatou-se que ela não era estacionária no tempo.

3.2.2 Processos Autorregressivos

Definição 3.2.4 *Um processo auto-regressivo, AR(p) é dado por:*

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_p Z_{t-p} + a_t \tag{3.2.2}$$

ou

$$\phi_p(B) Z_t = a_t \tag{3.2.3}$$

onde a_t é o ruído branco e ϕ_1, ϕ_2, ..., ϕ_p são constantes, Wei [23].

A Equação 3.2.2 deve satisfazer certas condições para que o processo seja estacionário, Box & Jenkins [3].

Neste trabalho, ajustou-se um modelo autorregressivo puro AR(p) onde p é o número de parâmetros utilizado no modelo. Porém, o modelo AR(p) não se mostrou adequado para representar a série de rede, pois necessitou-se de modelo misto.

3.2.3 Modelos de Box-Jenkins - ARMA

O modelo auto-regressivo puro AR (q) e o modelo média móvel puro MA (p) já eram conhecidos quando Box & Jenkins[3] desenvolveram uma metodologia para estimar e identificar modelos que incorporam as duas abordagens. Incorporando-se os dois tipos, tem-se o modelo ARMA (modelo autorregressivo de média-móvel). Depois, foi agregado o modelo ARIMA, que incorpora o parâmetro que indica o número de diferenças necessárias para a série se tornar estacionária.

Em algumas séries, a utilização de ambos os termos pode significar um modelo com um menor número de parâmetros do que seria necessário para descrevê-lo com a mesma significância utilizando apenas um de média móvel ou um autorregressivo.

Definição 3.2.5 *Seja Z_t um processo estocástico estacionário*

Z_t é um processo ARMA(p,q) se Z_t é estacionário e se para todo t,

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi p Z_{t-p} - a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \qquad (3.2.4)$$

ou

$$\phi(B) Z_t = \theta(B) a_t \qquad (3.2.5)$$

onde a_t tem distribuição N(0, σ^2) e os polinômios (1 - $\phi_1 a$- ... -$\phi_p a^p$) e (1 -$\theta_1 a - ... - \theta_q a^q$), não têm fatores comuns, Brockwell[4].

Existência e Unicidade:

A solução estacionária Z_t da equação 3.2.4 existe, se e somente se,

$\phi(a) = 1 - \phi_1 a - ... - \phi_p a^p \neq 0$ para todo $|a| = 1$, Brockwell [4].

Causalidade:

Um processo ARMA(p, q) , Z_t tem causalidade se existir constantes ϕ_j tal que

$\sum_{j=0}^{\infty} |\phi_j| < \infty$ e

$Z_t = \sum_{j=0}^{\infty} \phi_j a_{t-j}$ para todo t.

Causalidade é equivalente a condição de :

$\phi(a) = 1$ - $\phi_1 a - ... - \phi_1 a^p \neq 0$ para todo $|a| \leq 1$, Brockwell [4].

3.2.4 Modelos de Box-Jenkins - ARIMA

Para séries que apresentam comportamento não-estacionário, o modelo indicado para representá-la é o modelo auto-regressivo integrado e de média móvel, ARIMA (p, d, q), onde o 'd' é um parâmetro que indica o número de diferenças que a série sofreu para se tornar estacionária.

Definição 3.2.6 *Se d é inteiro e não negativo, um processo ARIMA(p, d, q) pode ser expressado por:*

$$\phi(B)(1 - B)^d Z_t = \theta(B) a_t \qquad (3.2.6)$$

onde $(1-B)^d Z_t$ pode ser escrito como $\Delta^d Z_t$

No modelo ARMA, todas as raízes devem estar fora do círculo unitário para que o processo seja estacionário. Se 'd' raízes da equação característica $\phi(B) = 0$ estão na fronteira e o restante fora do círculo, a série é chamada de não-estacionária homogênea.

Na prática, no caso de processos não-estacionários homogêneos, o usual é tomar de uma a duas diferenças para obter uma série aproximadamente estacionária, Morettin [16]:

- Uma diferença: quando a série é não-estacionária quanto ao nível. A série oscila em torno de um nível médio por um determinado período e depois passa para outro nível temporário.

- Duas diferenças: quando a série é não-estacionária quanto à inclinação. A série oscila em uma direção por um período de tempo e muda para outra direção.

Neste estudo, foram utilizados os procedimentos descritos anteriormente. Em especial, verificou-se a necessidade de tomar uma diferença na série. Ensaiou-se um modelo misto ARMA e um ARIMA, mas constatou-se que a série necessitava de tomar uma diferença, sendo então o ARIMA mais adequado.

3.2.5 Modelos ARFIMA

Os modelos ARIMA (p,d,q) são adequados para a modelagem do comportamento de séries temporais em curto prazo, ou seja, processos de 'memória curta', onde o parâmetro d é inteiro, que estabelece o nível de diferenças necessárias para tornar uma série temporal estacionária. Estudos mais recentes, Trevisan [22], propõem uma generalização dessa modelagem em relação ao parâmetro d, podendo este assumir não só valores inteiros, mas também representar graus de diferenças fracionárias. Modelos com essa propriedade permitem estudar séries caracterizadas por longas dependências temporais, chamados processos de 'memória longa'.

Um processo de memória longa é um processo estacionário em que a função de auto-correlação decresce hiperbolicamente (suavemente) para zero ao contrário de um modelo

ARIMA com d inteiro que decresce exponencialmente (uma vez que a função de autocorrelação decresce rapidamente para zero). Esse modelo é chamado de processo autorregressivo fracionário integrado de média móvel, ARFIMA (p,d,q) ou FARIMA, onde F significa "fractional".

Definição 3.2.7 *Um processo Z_t ARFIMA (p,d,q) com -1/2<d <1/2, se Z_t for estacionário e satisfizer a equação*

$$\phi(B)(1-B)^d Z_t = \theta(B)a_t \tag{3.2.7}$$

onde a_t é ruído branco e $\phi(B)$ e $\theta(B)$ são polinômios em B de graus p e q respectivamente, Morettin [16].

Se $d = 0$, Z_t é um modelo autorregressivo e de médias móveis, ARMA (p,q). Quando d $\neq 0$ e é não inteiro, a função de autocorrelação $\rho(k)$ tem decaimento hiperbólico, $\rho(k) \sim$ exp $|k|^{2d-1}$ com $|k| \to \infty$. As autocorrelações originadas de um modelo ARMA(p, q) têm um decaimento exponencial $p(k) \sim a_t$, 0 < a < 1 (Box & Jenkins [3]). Outro ponto é que o parâmetro d fracionário está ligado ao parâmetro de Hurst, que pode ser encontrado em Crato [8].

No caso do ARFIMA, um processo de longa dependência, se $0 < d < 0{,}5$. Um estudo mais detalhado pode ser encontrado em Cowperwait [7].

Na literatura, alguns estudos indicam que os dados do fluxo da RC possuem longa dependência e são altamente correlacionados, indicando que os modelos de séries temporais ARFIMA são uma boa forma de descrever o seu comportamento da rede.

Ajustou-se um ARFIMA para o fluxo de rede, mas o ARIMA se mostrou mais significativo para os dados analisados e mais fácil de operar e replicar.

3.2.6 Critérios de seleção de Modelos

Às vezes mais de um modelo se ajusta bem aos dados e quando isso ocorre é preciso usar um critério de desempate. Existem diversos métodos para verificar qual o modelo mais

significativo, sendo os mais utilizados o *Akaike Information Criterion* (AIC) e o *Bayesian information criterion* (BIC).

O AIC é um teste entre modelos onde, dado um conjunto de dados, vários modelos são classificados de acordo com o índice de AIC. O melhor modelo é aquele que tiver o menor valor de AIC calculado. Esse critério também se aplica ao BIC.

Definição 3.2.8 *Um modelo estatístico de k parâmetros é ajustado aos dados; para avaliar a qualidade do ajuste é utilizado o critério de AIC, Wei [23].*

O AIC é dado por:

$$AIC = -2ln\hat{L} + 2k \tag{3.2.8}$$

onde $\hat{L}$ é o valor máximo da função de verossimilhança, k o número de parâmetros do modelo para o processo ARMA.

O valor máximo da função de verossimilhança é dado por:

$$lnL = -\frac{n}{2}ln2\pi\sigma_a^2 - \frac{1}{2\sigma_a^2}S(\phi, \mu, \theta). \tag{3.2.9}$$

sendo n = número de observações. A soma dos quadrados é dada por:

$$S(\phi, \mu, \theta) = \sum_{t=-\infty}^{n} [E(a_t|\phi, \mu, \theta, Z)]^2 \tag{3.2.10}$$

e a $E(a_t|\phi, \mu, \theta, Z)$ é a esperança condicional de a_t dado ϕ, μ, θ e Z.

Maximizando ϕ, μ, θ e σ_a^2 e tendo $\sigma_a^2 = \frac{S(\phi,\mu,\theta)}{n}$,

$$ln(\hat{L}) = -\frac{n}{2}ln\sigma_a^2 - \frac{n}{2}(1 + ln2\pi) \tag{3.2.11}$$

sendo o segundo termo uma constante, o AIC se reduz para:

$$AIC(k) = nln\sigma_a^2 + 2k \tag{3.2.12}$$

Portando, escolhe-se o valor de k ($p + q$ do processo ARMA) tal que o AIC seja mínimo.

O BIC também é um teste entre modelos, onde o melhor modelo é dado por aquele que possui o menor valor calculado.

O BIC é menos provável de superestimar a ordem da parte autoregressiva, Wei [23].

Definição 3.2.9 *O BIC tem os mesmos critérios do AIC e é dado por:*

$$BIC(k) = -2ln\sigma_a^2 + kln(n). \tag{3.2.13}$$

Nesta pesquisa os conceitos anteriores foram essenciais para ajustar modelos de séries temporais. Para encontrar um modelo adequado aplicou-se a diversidade de modelos vistos (AR(p), MA(q), ARMA(p, q), ARIMA (p, d, q) e ARFIMA (p, d, q)),para diferentes valores dos parâmetros. Os critérios de AIC e BIC auxiliaram na escolha do modelo do fluxo de rede, que foi, um ARIMA(3,1,1).

¨

3.3 Gráficos de Controle Estatístico do Processo

Além de ajustar um modelo para o fluxo da RC em questão, buscou-se um método que pudesse monitorá-lo. A técnica utilizada foram os gráficos de Controle Estatístico do Processo (CEP). Neste trabalho, buscou-se encontrar o gráfico de CEP que melhor representasse os dados das RCs.

Em CEP, um dos gráficos mais conhecidos e utilizados é o de controle de Shewhart ($\bar{X}$ e R), tendo como vantagem a sua fácil aplicação. Porém, uma grande desvantagem dessas cartas é que elas utilizam apenas as informações contidas no último ponto plotado, descartando qualquer informação fornecida pela sequência inteira de pontos. Devido a essa característica, esse tipo de gráfico é insensível a pequenas mudanças no processo (da ordem de 1,5 σ ou menos) conforme Montgomery [14]. E o mais importante para os dados em questão é que as observações devem ser independentes e normalmente distribuídas, Moreira [15]. Nota-se que os dados da rede possuem autocorrelação, ou seja, eles não são

independentes e a informação contida nos pontos anteriores é importante para entender o comportamento tráfego da RC.

O Gráfico de Controle da Soma Acumulativa (CUSUM) e o de Média Móvel Exponencialmente Ponderada (MMEP ou 'EWMA') também são cartas de CEP, porém possuem características e propriedades diferentes dos gráficos de Shewhart. Por exemplo, o MMEP não necessita que os dados sejam independentes e normalmente distribuídos, admitindo observações correlacionadas.

Uma desvantagem para esses dois métodos é que a metodologia não é tão simples como os gráficos de $\bar{X}$ e R, exigindo mais conhecimento específico.

Contextualizando com o estudo da rede analisada, optou-se por monitorar os dados da RC com os gráficos MMEP e CUSUM, já que eles são mais sensíveis a pequenas variações.

3.3.1 Gráficos de Controle da Soma Acumulativa - CUSUM

Na Soma Acumulativa, as informações das amostras de um processo são acumuladas e ponderadas igualmente, ou seja, as amostras têm o mesmo peso. O CUSUM, além de detectar com mais eficiência pequenas mudanças, é muito eficaz com amostras de tamanho $n = 1$, Cunha [10],(n= 1 signfica que houve apenas um único valor observado várias vezes). Neste trabalho, foi considerada apenas uma amostra com n = 4700 observações.

A seguir detalham-se alguns conceitos de CUSUM baseados em Montgomery [14].

Definição 3.3.1 *O gráfico de CUSUM incorpora diretamente toda informação na sequência dos valores da amostra em relação a um valor-alvo. Suponha-se que tem-se diversas amostras de tamanho $n\geq 1$, e que $\bar{x}_j$ seja a média da j-ésima amostra. Então, se μ_0 é o alvo para a média do processo, o Gráfico de Controle da Soma Acumulativa é formado plotando-se a amostra i versus quantidade de C_i. E C_i é a soma acumulativa até a i-ésima amostra.*

$$C_i = \sum_{j=1}^{i} (\bar{x}_j - \mu_0) \tag{3.3.1}$$

Avalia-se o processo da seguinte maneira: se o processo permanece sob controle no valor-alvo μ_0, a soma cumulativa é um passeio aleatório com média zero. Porém, se a média é deslocada para um valor superior $\mu_1 > \mu_0$, pode-se dizer que tem uma tendência positiva que se desenvolverá na soma cumulativa C_i. Inversamente, se a média se desloca para baixo para um valor $\mu_1 < \mu_0$, então uma tendência negativa se desenvolverá em C_i. Quando um desses dois casos ocorre, pode-se afirmar que a média do processo mudou e deve ser feito um estudo para identificar as possíveis causas.

Há duas maneiras de representar o CUSUM, o tabular e a forma máscara V. Das duas, a tabular é preferível por ser mais fácil de ser aplicada. Pelo fato de os dois métodos serem indicados para o fluxo da RC, optou-se pelo método tabular para facilitar a reprodução do estudo para outros conjuntos de dados.

Definição 3.3.2 *O CUSUM tabular acumula desvios de μ_0 que estão acima do alvo, na estatística C^+, e acumula desvios de μ_0 que estão abaixo do alvo, na estatística C^-. As estatísticas C^+ e C^- são chamadas CUSUM uniteralmente superior e inferior e são calculadas da seguinte maneira:*

$$C_i^+ = max[0, x_i - (\mu_0 - K) + C_{i-1}^+] \tag{3.3.2}$$

$$C_i^- = max[0, (\mu_0 - K) - x_i + C_{i-1}^-] \tag{3.3.3}$$

onde K é usualmente chamado o valor de referência (ou valor de tolerância ou de folga), e é sempre escolhido a meio caminho entre o valor-alvo μ_0 e o valor da média fora de controle μ_1. Se a mudança é expressa em unidades de desvio padrão como $\mu_1 = \mu_0 + \delta\ \sigma$ (ou $\delta = \mu_1$ - μ_0/σ) então K é a metade da magnitude da mudança ou pode-se denotar por (Montgomery [14]):

$$K = \frac{\delta}{2}\sigma = \frac{\mu_1 - \mu_0}{2} \tag{3.3.4}$$

E escolhe-se K em relação ao tamanho da mudança que deseja detectar, sendo $K = k\sigma$, um valor frequente usado é de $k = 1/2$ e este foi o utilizado, Montgomery [14].

Nota-se que C_i^+ e C_i^- acumulam desvios a partir do valor-alvo μ_0 que são maiores do que K, com ambas as quantidades recolocadas em zero ao se tornarem negativas. Se tanto C_i^+ e C_i^- excederem o intervalo de decisão H, o processo é considerado fora de controle. O intervalo de decisão, pode ser expresssado por : $H = h\sigma$ onde geralmente aplica-se $h = 4$ ou 5.

3.3.2 Gráficos de Controle da Média Móvel Exponencialmente Ponderada - MMEP

O desempenho do MMEP é parecido com o CUSUM, porém tem a vantagem de ser mais fácil de operar e de estabelecer parâmetros. Para o monitoramento do fluxo de rede o MMEP possui alguns pontos muito relevantes, Montgomery [14]:

- não necessita de observações que sejam independentes entre si,
- é muito usado na literatura para modelagem e previsão de séries temporais,
- é tipicamente usado para observações individuais, pois os gráficos são robustos para distribuições não-normais, sendo quase um teste não paramétrico (livre de distribuição).

Definição 3.3.3 *O gráfico é construído da seguinte maneira:*

$$z_i = \lambda x_i + (1 - \lambda) z_{i-1} \tag{3.3.5}$$

ou

$$z_i = \lambda \sum_{j=0}^{i-1} (1 - \lambda)^j x_{i-j} + (1 - \lambda)^i z_0 \tag{3.3.6}$$

onde $0 < \lambda \leq 1$ *é uma constante e o valor inicial é o alvo do processo, de modo que* $z_0 = \mu_0$. *E* z_i *é a média ponderada de todas as médias de amostras anteriores e* x_i *é i-ésima amostra, Montgomery [14].*

Os pesos λ(1 - λ) decrescem geometricamente com a idade da média amostral. Além disso, os pesos têm soma igual a 1.

$$\lambda \sum_{j=0}^{i-1} (1-\lambda)^j = 1 \tag{3.3.7}$$

Os limites do MMEP são determinado por:

$$LSC = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]} \tag{3.3.8}$$

$$LC = \mu_0 \tag{3.3.9}$$

$$LSI = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]} \tag{3.3.10}$$

onde o fator L é a largura dos limites de controle.

Em geral, utilizam-se valores de λ no intervalo $0{,}05 \leq \lambda \leq 0.25$, pois fornecem um bom resultado na prática. Com valores menores de λ, podem-se detectar menores mudanças. Usualmente se usa $L = 3$ que corresponde aos limites três sigmas usuais das cartas de controle, com este L tem-se um bom desempenho com valores maiores de λ. Quando λ é relativamente bem pequeno, o ideal é reduzir a largura dos limites, adotando L entre 2,6 e 2,8, segundo Montgomery [14].

Capítulo 4

Materiais e Métodos

Este capítulo descreve os recursos e os procedimentos usados na pesquisa.

4.1 Materiais

O tratamento dos dados foi inteiramente feito com o auxílio de recursos computacionais, conforme segue:

- Computador pessoal Dell Inspirion 1525, com processador Intel Pentium Dual-Core T2350 - 1,86GHz, 2 GB de memória RAM e disco de 120 GB.
- Sistema Operacional MS - Windows Vista, versão 32 bits, edição Home Basic e Service Pack 1.
- Programas estatísticos: Minitab 15, JUMP (SAS) 7.0 e R 2.9.2 Project.

4.2 Bases de dados

As análises deste trabalho foram feitas a partir da observação do tráfego de uma rede real, em operação contínua. Os dados foram obtidos da base de traços de rede utilizadas em Angelis [1] e correspondem à observação da Ethernet do IFSC/USP entre as 10:00 h de 15 de outubro de 2001 e às 23:58 h de 21 de outubro de 2001, perfazendo 4.710 registros, disponibilizados para o trabalho na forma de arquivos de texto puro ASCII. Os

registros acumularam dados a cada 2 minutos de observação e agruparam as informações por fluxos, cuja definição é o conjunto de informações trocadas entre duas máquinas via rede, abrangendo toda a comunicação entre elas.

Dos traços *(trace)* existentes, escolheu-se aquele denominado 'Amostra 3', que capturou todo o tráfego da rede com o emprego de um *sniffer* (Netramet). Deve-se notar que essa amostra correspondente a um período de operação normal da rede, sem ocorrência de ataques ou problemas detectáveis pelos usuários e administradores.

4.3 Redução de Variáveis

Para reduzir o número de variáveis do mrCep, procedeu-se da seguinte maneira:

- as variáveis relativas ao controle da coleta de dados, introduzidas pela operação do *sniffer*, como número do fluxo e regra de coleta foram descartadas;
- as variáveis correspondentes à identificação de interfaces de rede Ethernet (*MAC Address*), por serem irrelevantes no contexto, foram descartadas;
- as invariantes em cada fluxo foram também desprezadas, pois não contribuem para análises, como, por exemplo, máscara de rede, porta TCP/UDP, protocolo de transporte e de rede e tipo de endereçamento das camadas 2 e 3 da arquitetura TCP/IP;
- os endereços IP individualmente e a máscara de rede foram usados unicamente para determinar a localização interna ou externa da máquina em relação à rede;
- para as variáveis restantes, Pacotes fonte - destino, Bytes fonte - destino, Pacotes destino - fonte e Bytes destino - fontes foram calculados o coeficientes de correlação linear e aquelas com coeficiente maior que 0,7 foram consideradas duplicadas e, portanto, removidas da lista de variáveis a serem monitoradas.

Ao final, a variável que melhor representa a rede é o número de pacotes fonte - destino.

4.4 Modelagem da rede

As análises desta pesquisa tomaram unicamente os valores de números de pacotes IP enviados da máquina fonte (iniciadora da comunicação) para o computador destino (respondente). Por brevidade, estes dados serão doravantes referenciados como pacotes fonte-destino.

Na amostra considerada, havia 6 pontos discrepantes (*outliers*), que foram retirados do conjunto de 4710 registros para viabilizar a análise estatística.

Foram definidos 4 cenários para auxiliar a escolha do modelo mais adequado para os dados, conforme Tabela 4.1

Tabela 4.1: Cenários do trabalho

Cenário	Descrição	Série
1	Série Pacotes fonte-destino pontual	Série 1
2	Série Pacotes fonte-destino Período Fluxo alto (6:00 às 20:00)	Série 2
2	Série Pacotes fonte-destino Período Fluxo Baixo (20:02 às 5:58)	Série 3
3	Média dos dias da Série 2	Série 4
3	Média dos dias da Série 3	Série 5
4	Mediana dos dias da Série 2	Série 6
4	Mediana dos dias da Série 3	Série 7

Para cada cenário, buscou-se ajustar um modelo estatístico conforme a sequência de passos seguintes:

- Considerou-se os modelos AR(p), MA(q), ARMA(p, q), ARIMA(p, d, q) e ARFIMA(p, d, q).
- procurou-se o ajuste com menor número de parâmetros possível em cada modelo, observando-se as características das funções de ACF e a PACF da série na Tabela 4.2, Wei [23]. Em paralelo, buscou-se os menores valores alcançáveis de AIC e BIC.

Escolheu-se o ARIMA para modelar todos os cenários, vista a sua adequação aos dados.

O Cenário 4 foi o que melhor representou a rede e a pesquisa utilizou-se unicamente desse cenário para as cartas de controle.

Tabela 4.2: Características das funções ACF e PACF para processos estacionários

Processo	ACF	PACF
AR(p)	Decaimento exponencial	Corte depois do lag p
MA(q)	Corte depois do lag q	Decaimento exponencial
ARMA(p, q)	Corte depois do lag (q-p)	Corte depois do lag (p-q)

4.5 Monitoramento da Rede

As cartas CUSUM e MMEP foram usadas para investigar o comportamento das séries de dados do Cenário 4 e o processo de construção dos gráficos foi feito como segue:

- o valor alvo adotado foi a média de cada série, do fluxo alto e o outro do baixo.
- no CUSUM - utilizou-se o valor de referência, $(K) = 1/2\ \sigma$ e o intervalo de decisão, $H = 4\sigma$ e $H = 5\sigma$;
- no MMEP - Utilizou-se a largura dos limites de controle $(L) = 2$ e 3 e o peso: $\lambda =$ 0,05 , $\lambda = 0{,}2$ e $\lambda = 0{,}4$.

Capítulo 5

Análise da Rede

O desenvolvimento desta pesquisa envolveu 3 aspectos: a redução de variáveis do mrCEP, a modelagem e o monitoramento da rede, conforme descrito a seguir.

5.1 Redução do Número de Variáveis

Como descrito em materiais e métodos, as variáveis consideradas como relevantes para a representação da rede foram: Pacotes fonte - destino, Bytes fonte - destino, Pacotes destino - fonte e Bytes destino - fonte.

Tabela 5.1: Coeficiente de correlação

	Pacotes ft.dt	Bytes ft.dt	Pacotes dt.ft
Bytes ft.dt.	0,85		
Pacotes dt.ft	0,85	0,46	
Bytes dt.ft	0,64	0,15	0,94

onde fonte - destino = ft.dt e destino - fonte = dt.ft

Calculou-se o coeficiente de correlação linear entre elas, Tabela 5.1. Nota-se que algumas variáveis são altamente correlacionadas com $\rho(X, Y) > 0.7$, sendo elas: Pacotes ft.dt vs Bytes ft.dt, Pacotes ft.dt vs Pacotes dt.ft e Pacotes dt.ft vs Bytes dt.ft.

Para a escolha das variavéis a serem monitoradas, considerou-se o seguinte racional:

sendo Pacotes dt.ft vs Bytes dt.ft correlacionados, descartou-se a variável Bytes dt.ft. Pacotes ft.dt vs Pacotes dt.ft, também estão correlacionadas, eliminou-se a variável Pacotes dt.ft. E por fim observada a correlação entre Pacotes ft.dt vs Bytes ft.dt com correlação maior que 0,7, manteve-se a variável Pacotes ft.dt. Portanto foi indicada apenas uma variável a ser monitorada.

Portanto para monitorar e estimar um modelo foi usado o número de Pacotes fonte - destino. O mesmo tipo de análise pode ser replicado para as outras variáveis.

5.2 Modelagem da rede

Observando-se a série Pacotes fonte - destino, notou-se que alguns pontos variaram significativamente em relação à média, Figura 5.1 a-).

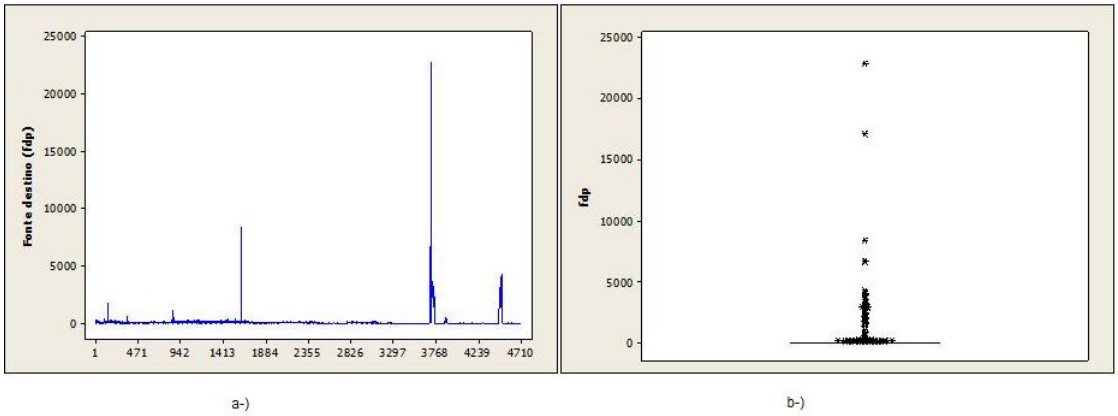


Figura 5.1: a-)Gráfico da Série ao longo do tempo. b-)Boxplot da Série.

Na Figura 5.1 b-) é possível notar essa variação por meio do boxplot. Esse é um gráfico que possibilita representar a distribuição de um conjunto de dados com base em algumas medidas descritivas, mediana ($q2$), o quartil inferior ($q1$), o quartil superior ($q3$). A variação dos dados é tão significativa que a caixa do boxplot contendo os quartis está comprimida devido à grande dispersão dos pontos atípicos. Em especial tem-se 6 pontos que são os que mais estão discrepantes em relação ao conjunto de dados, os chamados *outliers* são pontos que não se enquadram no padrão do restante dos dados; eles podem ser erros de medida ou um comportamento diferente do padrão.

.

Baseado na Figura 5.1 b-) e analisando a variação em relação à média, optou-se por retirar os 6 pontos (Tabela 5.2), sendo um deles porque é a primeira observação da coleta que não acumula dados dos 2 minutos anteriores e, portanto, está incompleto e os restantes devido à grande variação pontual. Esses pontos foram retirados levando em consideração que em nenhum momento a rede analisada sofreu queda/interrupção. Devido a este fato, entendeu-se que essa variação foi atípica e como não comprometeu a qualidade da rede naquele momento, optou-se por retirar tais observações a fim de ajustar um modelo mais próximo do comportamento normal de uma rede em perfeitas condições de funcionamento. Portanto, todas as análises seguintes foram sem tais observações.

Tabela 5.2: Pontos discrepantes da amostra.

Dia	Hora	Pacotes
15	10:00	6
17	15:38	8456
20	13:24	6768
20	13:26	6695
20	13:36	22845
20	13:48	17154

Tabela 5.3: Medidas de localização e dispersão.

	Nº de observações	Média	Desvio padrão	Mínimo	Máximo
Total	4710	150,1	545,1	6	22845
Sem *outliers*	4704	137,2	305,4	43	4377

Analisando novamente a série de dados, Figura 5.2 a-), é possível notar que existe variação do fluxo de pacotes ao longo do tempo, como os 'picos' no final da série. Porém, entendeu-se que essa variação é característica da rede. Na Figura 5.2 b-) pode-se notar que o boxplot ainda apresenta pontos atípicos, porém antes o ponto máximo era 22845 e agora é de 4377, (Tabela 5.3).

Na Tabela 5.3 analisando-se o desvio padrão, o mínimo e o máximo dos valores observados, nota-se que a variabilidade das observações diminuia com a retirada dos pontos discrepantes.

Para ajustar um modelo mais aderente, optou-se por analisar a série Pacotes fonte - destino distribuindo os dados em 4 cenários, como citados na seção de métodos.

5.2.1 Cenário 1

Constatou-se que a média da Série 1 era constante ao longo do tempo. Na Figura 5.2 a-), nota-se que existem variações, que indicavam que a série deveria tomar uma diferença. Antes de passar por uma diferença a Série 1, foram analisadas a função de autocorrelação(ACF) e a de autocorrelação parcial (PACF).

Na Figura 5.3, observa-se o comportamento delas: na ACF os *lags* decaem lentamente e na PACF o número de *lags* significativos indicaram a ordem do modelo, um ARMA(5,1), porém esse decaimento pode indicar a necessidade de tomar uma difereça na série.

Na Figura 5.4, tem-se a ACF e PACF após uma diferença, e na Tabela 5.4 nota-se que o ARIMA(5,1,3) é o modelo mais indicado considerando-se: ACF e PACF e pelo AIC, ou seja, era necessária uma transformação na série.

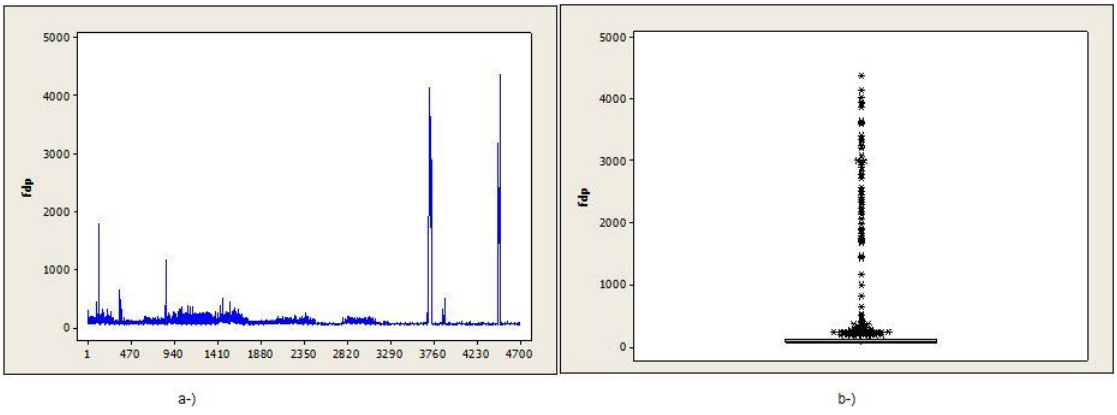


Figura 5.2: a-)Gráfico da Série ao longo do tempo (sem *outliers*). b-)Boxplot da Série (sem *outliers*).

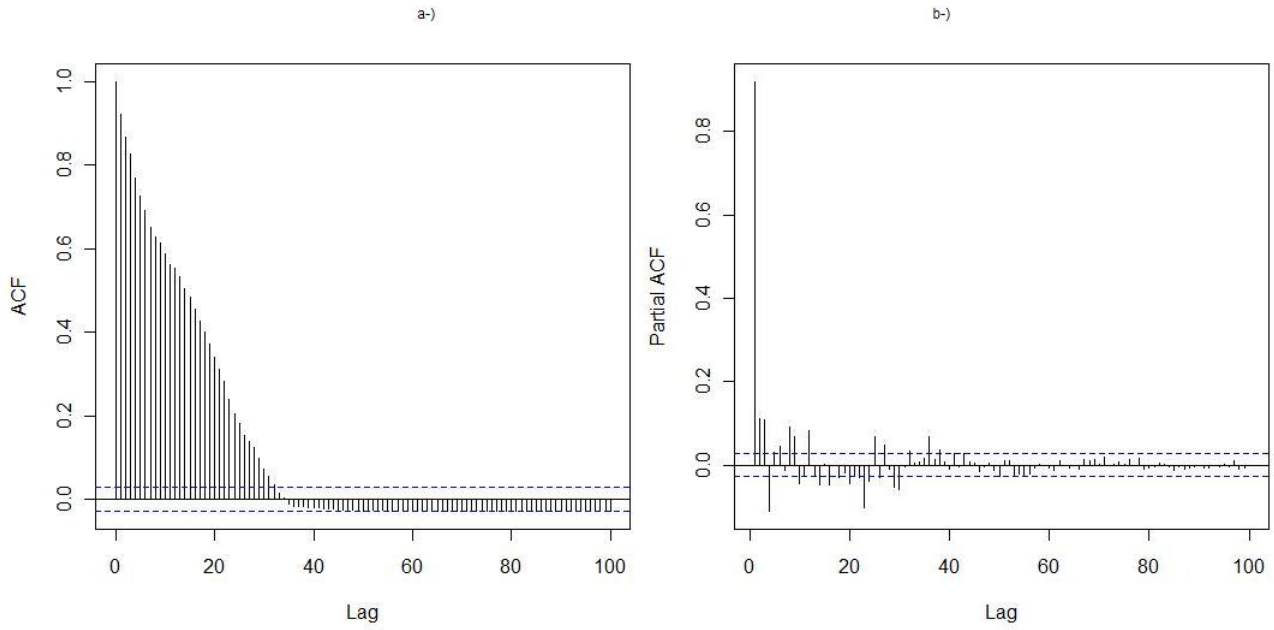


Figura 5.3: a-)Gráfico da Função de Autocorrelação da Série 1. b-) Gráfico da Função de Autocorrelação Parcial da Série 1.

5.2.2 Cenário 2

Para o dimensionamento da rede, divide-se em períodos de maior e menor movimento, no Cenário 2, a série foi separada em 2 períodos: um de maior fluxo (das 6:00 às 20:00) e o outro contendo o menor fluxo. Chegou-se a essa divisão analisando-se o fluxo no decorrer de cada dia. Na Figura 5.5 a-), tem-se cada dia expresso no instante de cada coleta. Nota-se que, ao longo do fluxo, o horário entre 6:00 às 20:00 tem maior quantidade de pacotes do que o restante.

Na Figura 5.5 b-), observa-se a dispersão do fluxo de pacotes por dia e nota-se que alguns dias tiveram uma variação superior aos restantes, sendo eles: 20 e 21.

A análise desse cenário iniciou-se pela Série 2, considerando-se o período diário do tráfego mais intenso. Em consequência, para cada dia amostrado, o fluxo de pacotes foi agrupado em dois períodos: das 6:00 às 20:00 (maior fluxo) e das 0:00 às 5:58 e 20:02 às 23:58 (menor fluxo).

Na Figura 5.6 a-) tem-se o 'boxplot' da Série 2, onde há uma grande dispersão dos

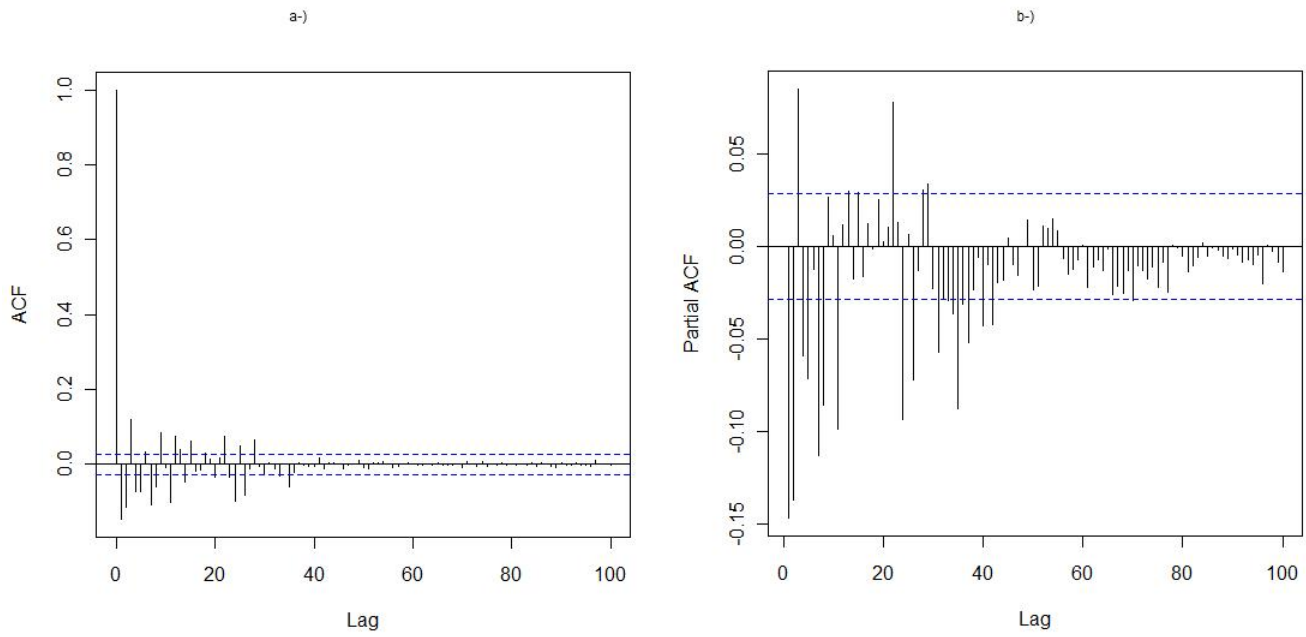


Figura 5.4: a-)Gráfico da ACF da Série 1 com uma diferença. b-) Gráfico da PACF da Série 1 com uma diferença.

dados depois do $q3$. Tal dispersão se deve à grande variação contida nos dias 20 e 21.

A quantidade de pacotes enviados do inicio até por volta da observação 2000, da Série 2, sofre pouca variação e a média é praticamente constante. Porém, após este período, tem-se dois momentos com uma quantidade superior ao início, correspondentes aos dias 20 e 21, Figura 5.6.

Observando a ACF e a PACF da Série 2, Figura 5.7, nota-se que a ACF possui um decaimento significativo e que, na PACF, os 3 primeiros *lags* são significativos, indicando que a parte autoregressiva do modelo é provavelmente de ordem 3 (AR(3)).

Tabela 5.4: Tabela de critério AIC e BIC para modelo Série 1.

Modelo	AIC	BIC
(1, 0, 0)	58321,58	58324,03
(1, 1, 0)	58393,48	58397,93
(1, 1, 1)	58328,41	58339,31
(1, 0, 1)	58251,23	58260,13
(2, 0, 2)	58196,25	58218,06
(3, 0, 1)	58171,53	58193,36
(4, 0, 1)	58152,14	58180,42
(4, 0, 0)	58152,45	58174,26
(5, 0, 1)	58134,17	58168,89
(6, 0, 1)	58138,74	58179,93
(5,0.49,1	58149,94	58186,67
(5, 1, 3)	**58069,34**	**58118,98**
(5,1,1)	58194,58	58231,32

Na Figura 5.8 tem-se a ACF e PACF após uma diferença e na Tabela 5.5 nota-se que o modelo diferenciado é o que se ajusta melhor aos dados. O ARIMA (5,1,3) é o mais indicado, sendo o que possui o menor AIC.

Para o restante das observações, deu-se o nome de Pacotes fonte - destino baixo fluxo (Série 3). A mesma análise da Série 2 foi realizada para a Série 3, que tem um menor desvio padrão. Na Figura 5.9 a-), nota-se que a 'caixa' do boxplot expandiu-se e os pontos atípicos diminuíram em relação à Série 2, indicando uma maior concentração dos dados. Apesar da variação não ser tão expressiva quanto a Série 2, na Figura 5.9 b-) é possível notar que a Série 3 tem um maior fluxo e variação apenas no começo da amostra até por volta da observação 1130, indicando que a média não é constante em grande parte do tempo.

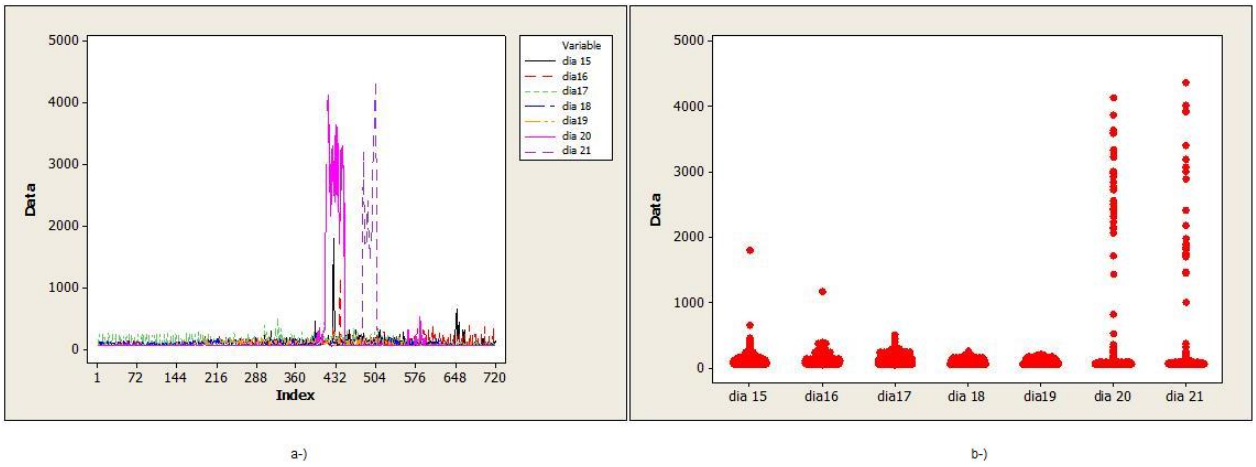


Figura 5.5: a-)Gráfico dispersão do fluxo por dia(início às 0:00 e termino às 23:58). b-) 'Dotplot' - Fluxo de pacotes por dia.

Tabela 5.5: Tabela de critério AIC e BIC para modelo da Série 2 .

Modelo	AIC	BIC
(3, 0, 0)	36190,65	36204,5
(3, 1, 1)	36210,94	36232,7
(3, 0, 1)	36163,53	36183,3
(2, 0, 1)	36245,49	36259,3
(2, 1, 1)	36217,51	36233,4
(2, 0, 0)	36220,26	36228,1
(4, 0, 1)	36144,67	36170,4
(9, 0, 1)	36104,64	36160,1
(3, 0, 2)	36104,03	36129,8
(3, 1, 2)	36173,2	36200,9
(4, 0, 2)	36105,88	36137,5
(5,1,3)	**36091,1**	**36129,8**

A ACF da Série 3 não apresentou nenhum decaimento exponencial, mas todos os *lags*,

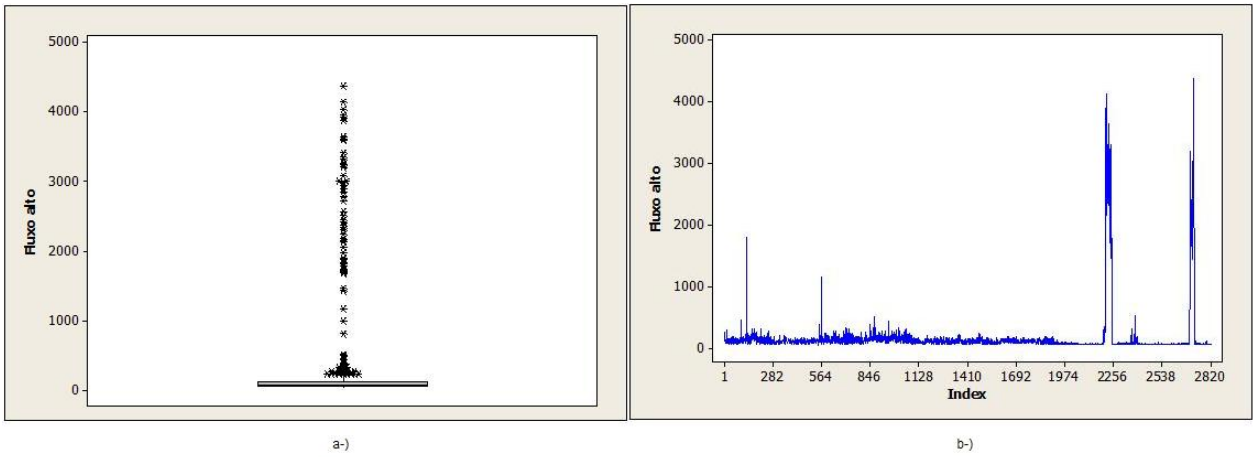


Figura 5.6: a-)Boxplot da base Série 2. b-)Gráfico da Série 2 ao longo do tempo.

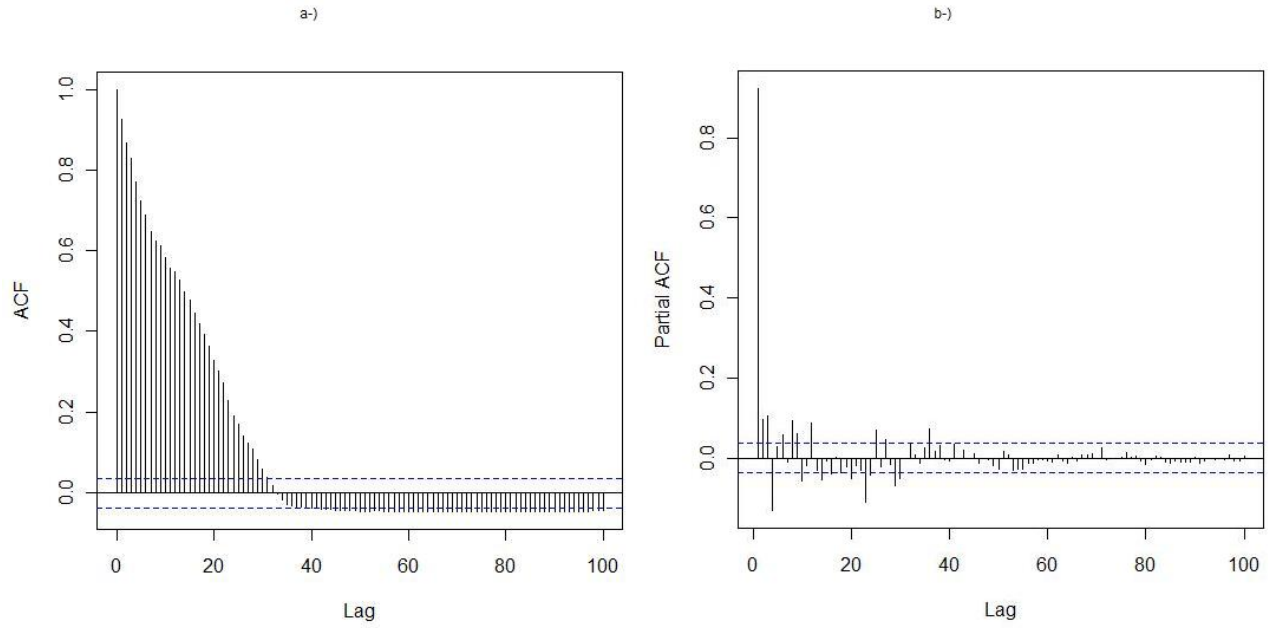


Figura 5.7: a-)ACF da Série 2. b-)PACF da Série 2

até 100º significativos, estão acima dos limites, Figura 5.10 a-). Na PACF, Figura 5.10 b-), não se tem um decaimento expressivo e há alguns *lags* significativos. Analisando a série e as funções, optou-se por tomar uma diferença, sendo que em grande parte do fluxo, a média não era constante.

Estudando-se a ACF da série após uma diferença, Figura 5.11 a-), nota-se que esta

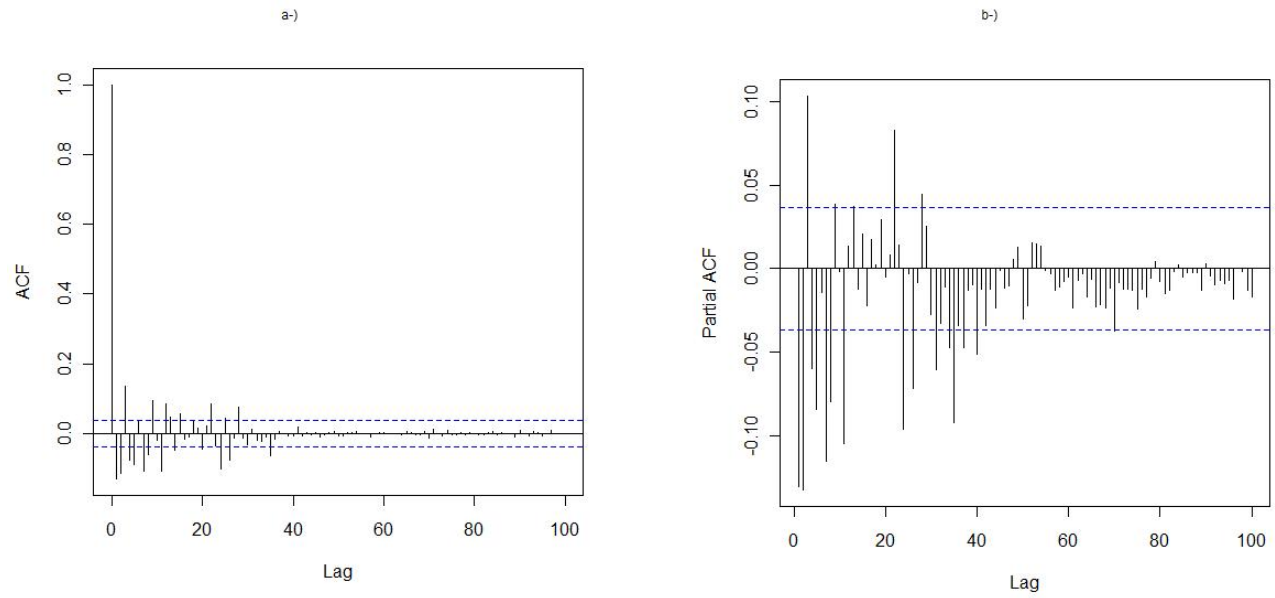


Figura 5.8: a-)ACF da Série 2 após uma diferença. b-)PACF da Série 2 após uma diferença

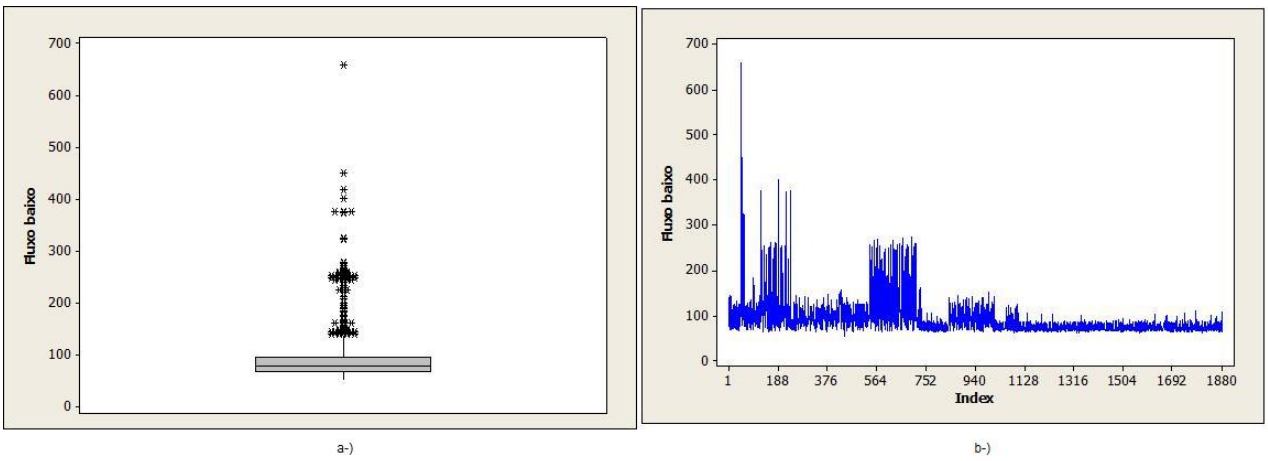


Figura 5.9: a-)Boxplot da Série 3. b-)Gráfico da série Série 3 ao longo do tempo.

possui um decaimento exponencial depois do primeiro *lag*. Na PACF, Figura 5.11 b-), tem-se indicação que a parte autorregressiva do modelo possui ordem de 4 a 6.

Observando-se a Tabela 5.6, conclui-se que o modelo mais indicado é um ARIMA(5,1,1), já que este possui os menores AIC e BIC. Realizou-se um teste para identificar se a série

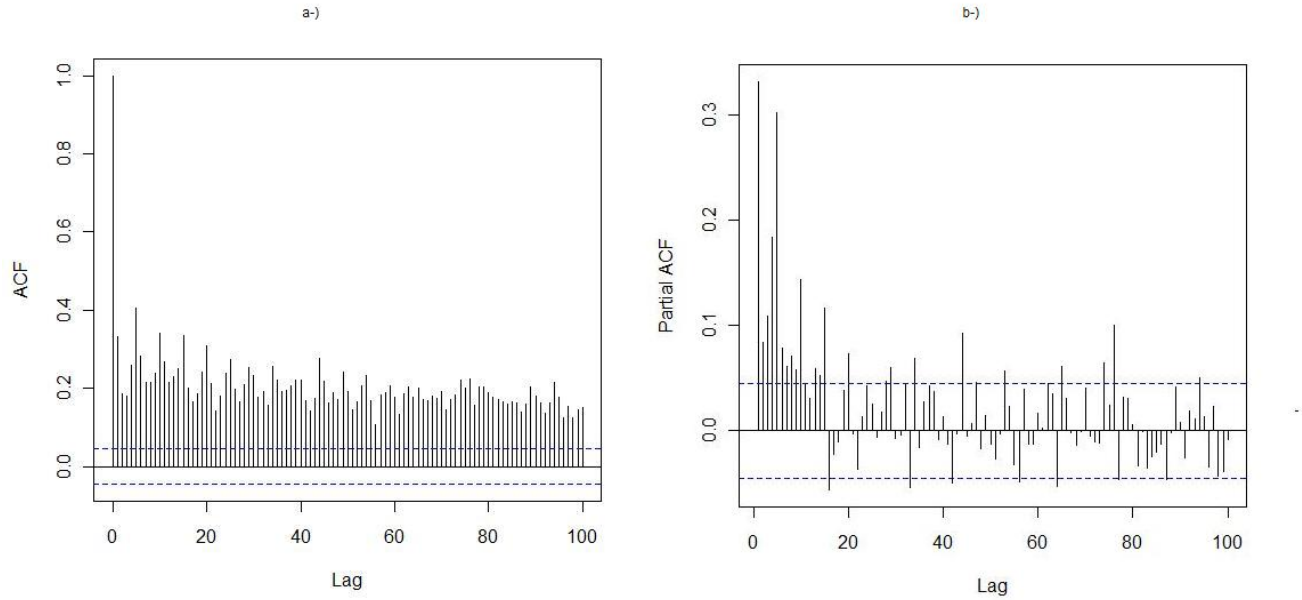


Figura 5.10: a-)ACF da Série 3. b-) PACF da Série 3.

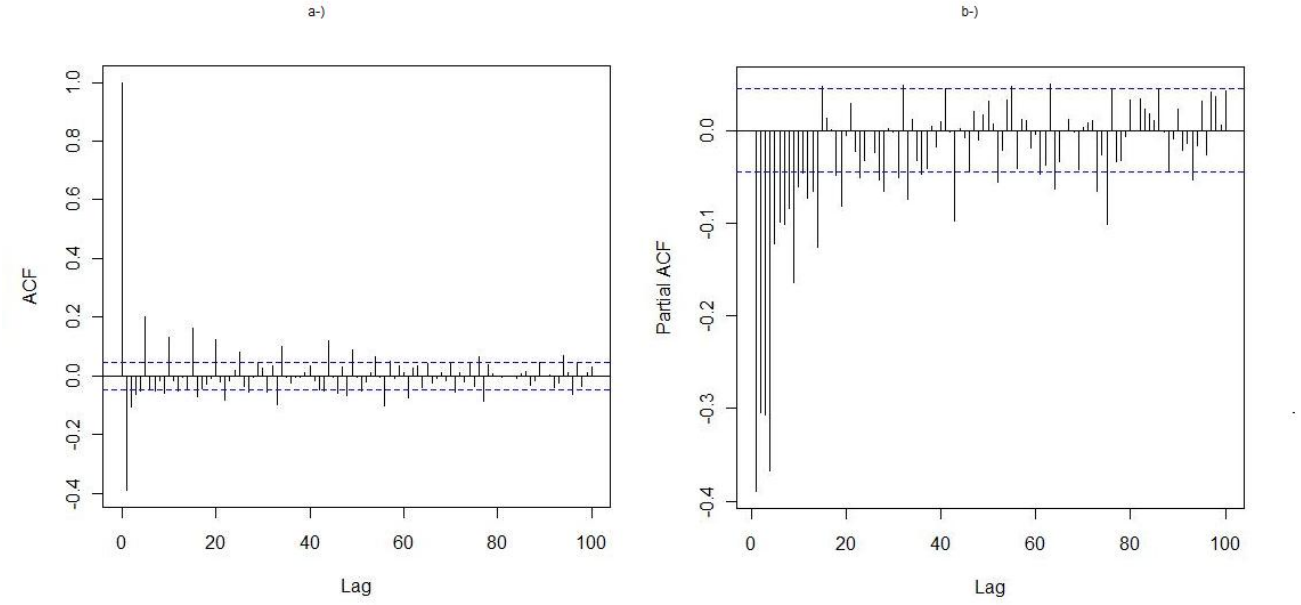


Figura 5.11: a-)Gráfico da Série 3 diferenciada uma vez. b-)ACF da Série 3 diferenciada

possuía longa dependência, porém com o 'd' sugerido pela a análise, isso não se verificou, sendo que o AIC com d=1 foi menor do que o com $d = 0{,}22$.

Em resumo, para o Cenário 2, tem-se dois tipos de modelo para cada período do fluxo: um ARIMA(5,1,3) e um ARIMA (5,1,1). Isto mostra a necessidade de se dividir os dados, já que cada série necessitou de um modelo diferente.

Tabela 5.6: Tabela de critério AIC e BIC para modelos da Série 3.

Modelo	AIC	BIC
(1, 1, 1)	18885,12	18894,2
(1,0,1)	18908,64	18915,7
(2,1,1)	18864,9	18879,5
(2,1,2)	18855,85	18876,0
(4,1,2)	18827,26	18858,5
(5,1,2)	18790,75	18827,5
(5,2,2)	18798,45	18835,2
(5,2,3)	18799,67	18842,0
(6,1,3)	18793,43	18841,3
(5, 1, 1)	**18788,79**	**18820,0**
(5, 1, 0)	18933,06	18958,8
(5,0.22,0)	18966,23	18991,9

5.2.3 Cenário 3

Para o Cenário 3, utilizou-se a mesma separação de períodos do cenário anterior. Entretanto em vez de valores pontuais, utilizou-se a média obtida nos 5 dias de coleta no mesmo instante.

Para a média do fluxo com maior quantidade de pacotes transmitidos, Série 4, observou-se a dispersão dos dados, Figura 5.12 a-), que apresentou uma maior variação entre $q2$ e $q3$ do que em $q1$ e $q2$. Ainda se tem alguns pontos atípicos e mesmo utilizando a média dos dados, a variação ainda é significativa.

Analisando-se a Figura 5.12 b-), é nítida a influência da variação dos dias 20 e 21.

Mesmo utilizando-a a média dos dias do mesmo instante, a variação ainda é significativa. Deve-se notar que há pouca variação dos pontos fora dos 'picos' indicados na 5.12 b-).

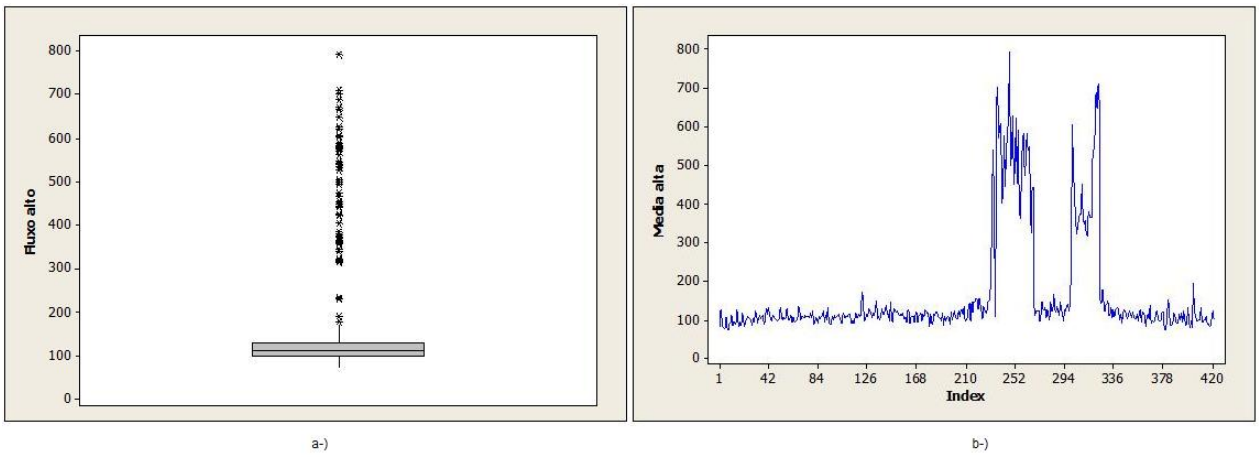


Figura 5.12: a-)Boxplot da Série 4. b-)Gráfico da Série 4 ao longo do tempo

A ACF da Série 4, Figura 5.13 a-), não possui decaimento exponencial e a PACF, Figura 5.13 b-), possui 3 *lags* significativos, indicando um modelo autorregressivo de ordem 3. Porém, como observado na Figura 5.12a-), a série não tem média constante em alguns momentos e portanto, analisaram-se modelos após uma diferença também.

A série foi diferenciada uma vez, Figura 5.14, e as ACF e PACF foram refeitas, Figura 5.15, respectivamente. Continuaram indicando um autorregressivo ordem 3, porém com uma diferença e a ACF com 3 *lags* significativos.

Para a escolha do modelo, realizaram-se simulações com ordens diferentes. Na Tabela 5.7, tem-se o AIC e BIC para cada simulação e observa-se que a indicação de um ARIMA(3,1,3) pela ACF e PACF também é indicado pelo menor AIC (não pelo BIC).

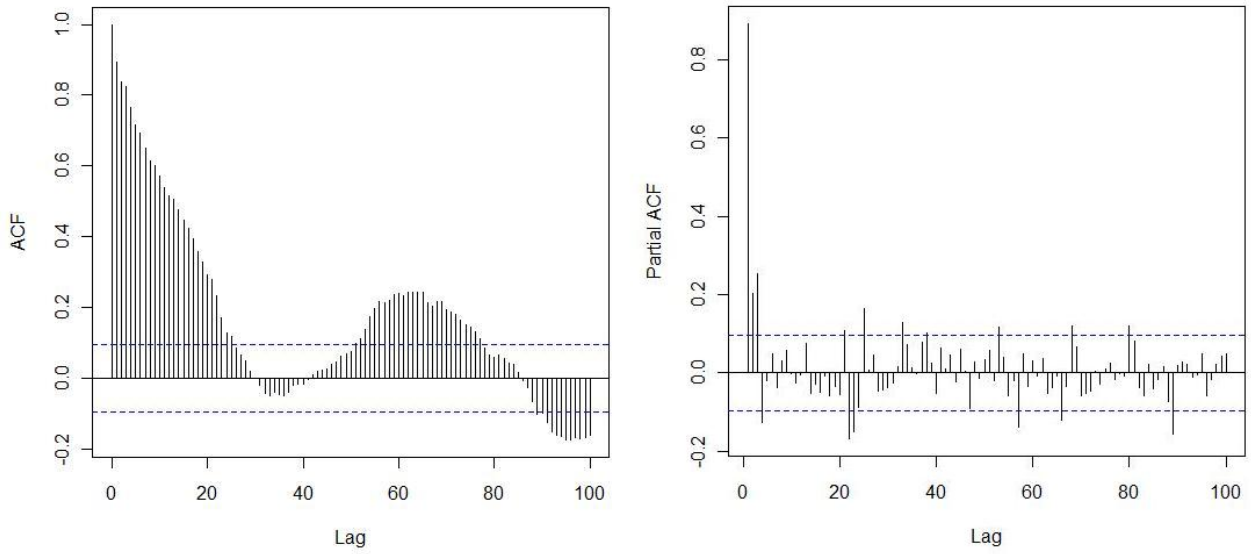


Figura 5.13: a-)ACF de Série 4. b-) PACF de Série 4.

Tabela 5.7: Tabela de critério AIC e BIC para modelo da Série 4.

Modelo	AIC	BIC
(1, 0, 1)	4679,97	4684,07
(3, 0, 1)	4661,13	4673,31
(3, 1, 1)	4655,96	4676,17
(3, 1, 0)	4653,97	4670,15
(3, 0, 0)	4663,89	4672,01
(2, 1, 0)	4656,43	4668,55
(2, 1, 1)	4654,66	4670,83
(3,0.49,0)	4667,8	4683,97
(3, 1, 2)	4655,94	4680,20
(3, 1, 3)	**4650,8**	**4679,10**

Para o restante da amostra, que possui o período com a média do fluxo com a baixa quantidade de pacotes enviados, Série 5, realizou-se a mesma.

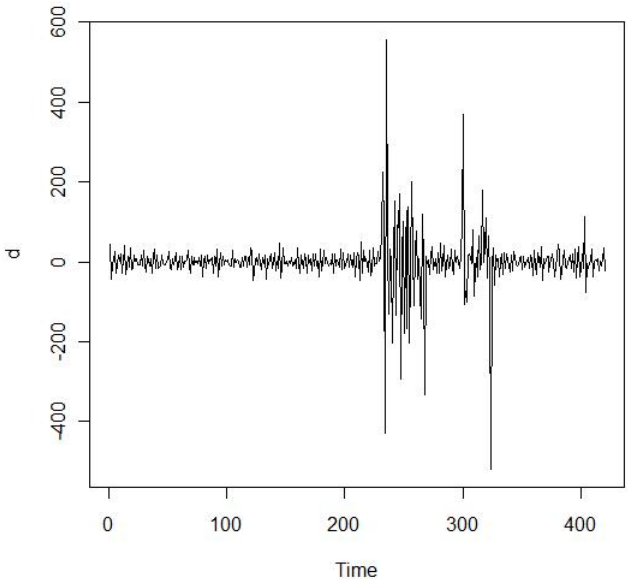


Figura 5.14: Gráfico dá Série 4 diferenciada.

Observando a Série 5, Figura 5.16 a-), nota-se que é bem mais estável que a Série 3, já que a Série 5 usa a média dos dias. Na Figura 5.16 b-) nota-se que a quantidade de valores acima de q3 é bem menor em relação às séries já analisadas.

Analisando-se as ACF e PACF, Figura 5.17, não se vê indicação clara de um modelo. Observando-se a Tabela 5.8, o modelo que apresentou o menor AIC foi um ARIMA (5,1,1), ou seja, para a Série 5 se tornar estacionária, foi necessário uma diferença, d =1. Na Figura 5.18, tem-se a série tomada uma diferença, e nela é possível notar que os dados estão ao redor do eixo $x = 0$.

Observando-se a ACF e PACF da série diferenciada, Figura 5.19, confirma-se o modelo proposto pelo AIC. Na ACF, tem-se um *lag* significativo indicando, um MA (1), e a PACF

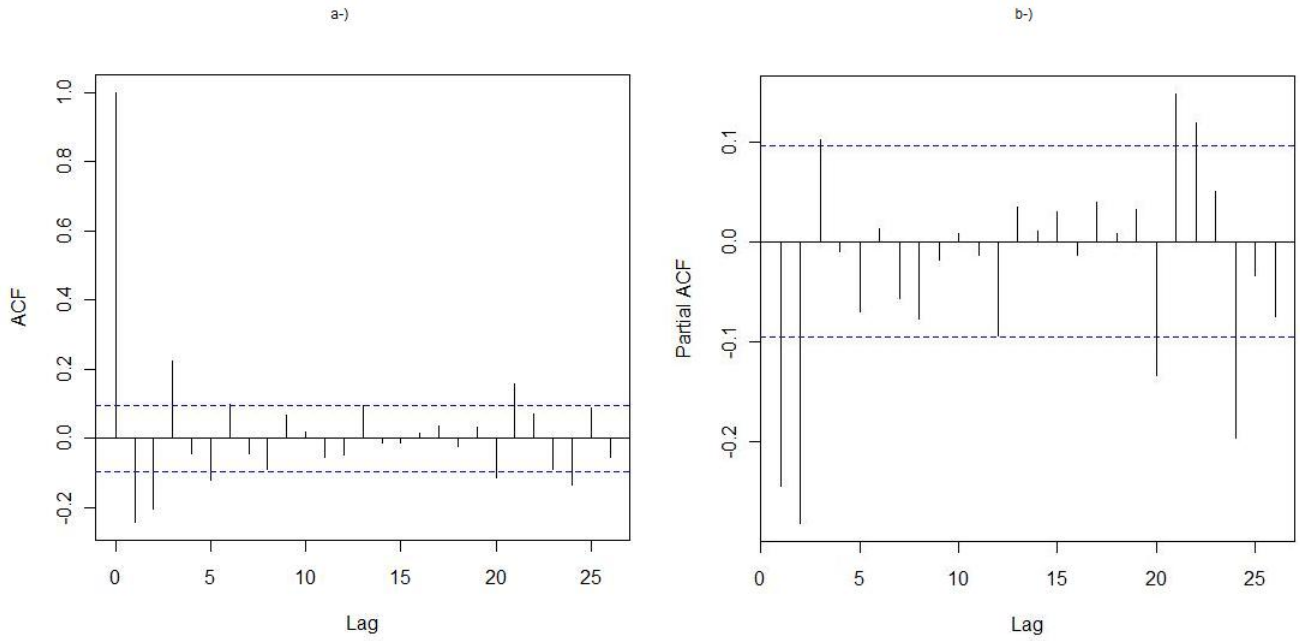


Figura 5.15: a-)ACF de Série 4 diferenciada. b-) PACF de Série 4 diferenciada

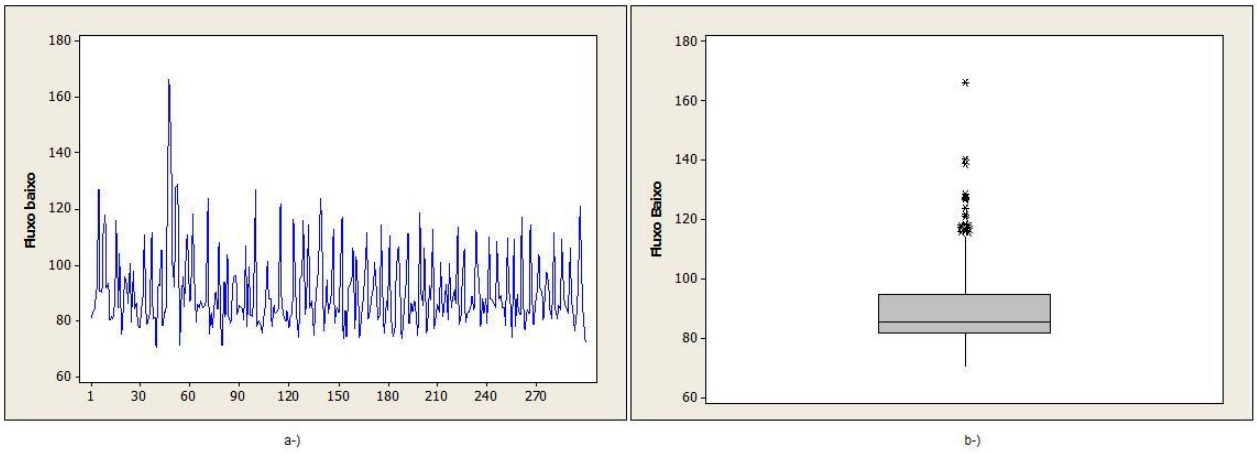


Figura 5.16: a-)Gráfico da Série 5. b-)Boxplot da Série 5.

tem 5 *lags* significativos, indicando que parte autorregressiva é de ordem 5. Então, por fim, tem-se ARIMA (5,1,1).

Para o Cenário 3, tem como melhor ajuste um modelo ARIMA(3,1,3) para a Série 4

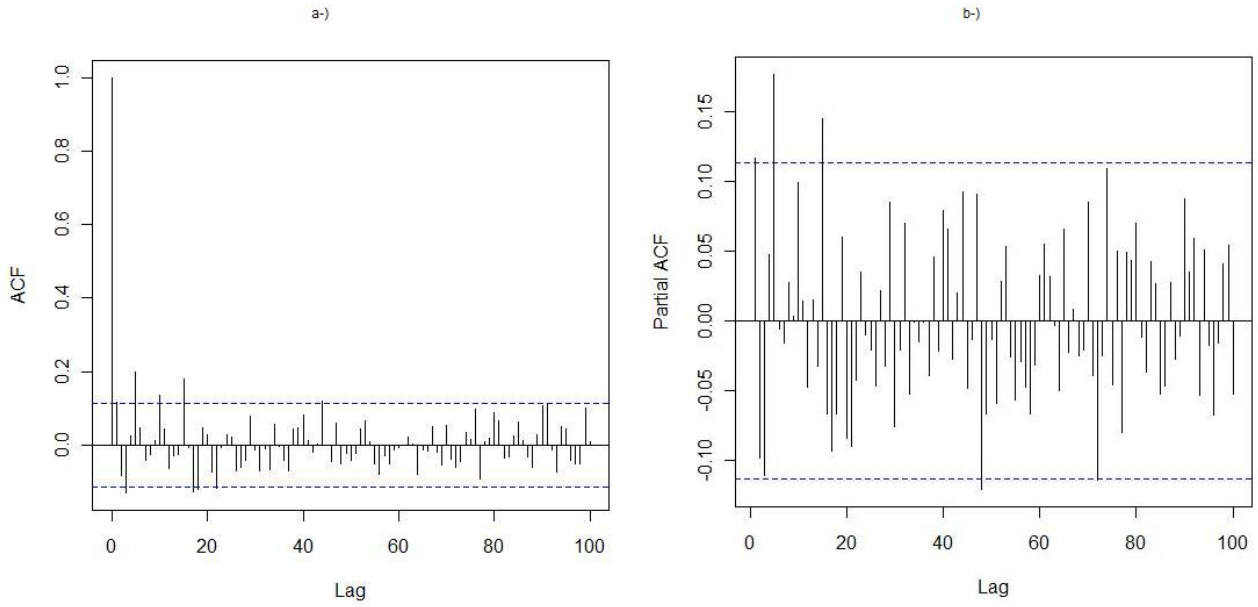


Figura 5.17: a-)ACF da Série 5. b-)PACF da Série 5.

Tabela 5.8: Tabela de critério AIC e BIC para modelos da Série 5.

Modelo	AIC	BIC
(1, 0, 1)	2417,01	2420,4
(0, 0, 1)	2415,29	2415,0
(0, 1, 1)	2416,39	2423,8
1, 1, 1)	2414,3	2425,4
(1, 0, 0)	2416	2415,7
(4, 1, 2)	2413,24	2439,1
(5, 1, 2)	2407,09	2436,7
(6, 1, 2)	2408,99	2442,3
(5, 1, 3)	2409,01	2442,3
(5,1,1)	**2405,09**	**2431,0**
(5, 0.048,1)	2426,83	2452,7

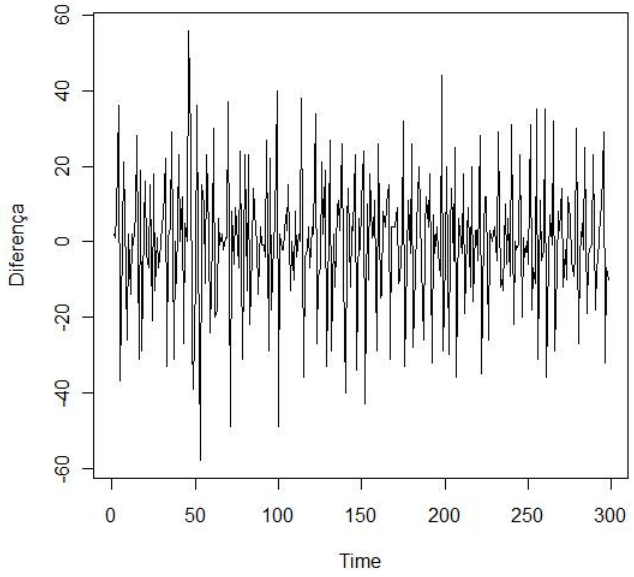


Figura 5.18: Gráfico da Série 5 após uma diferença.

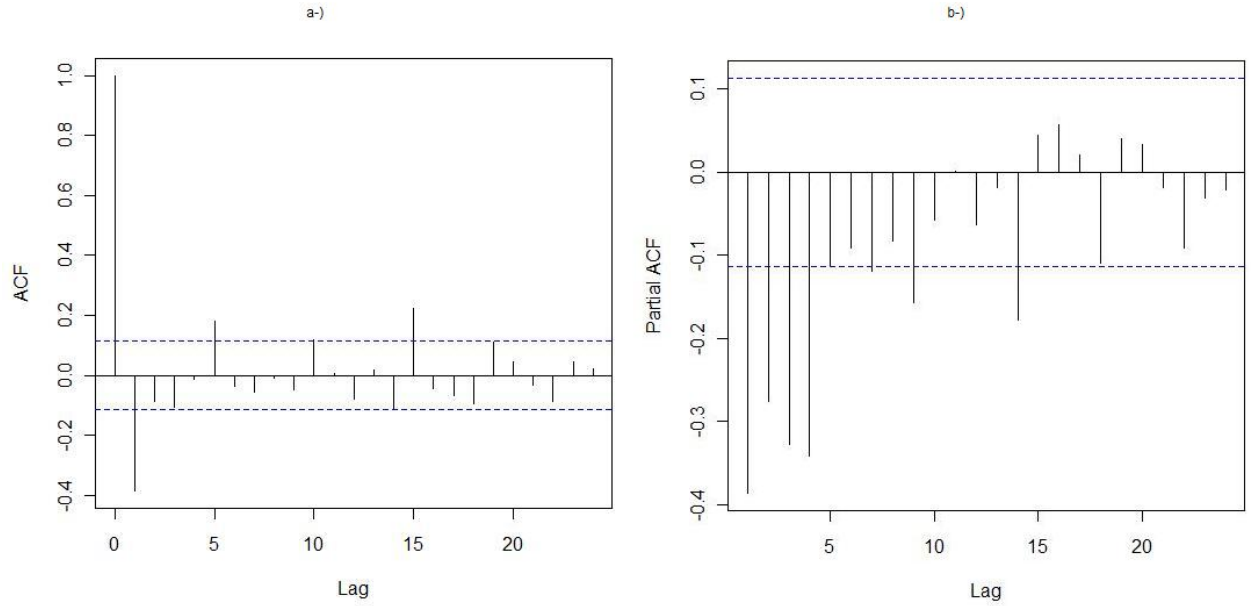


Figura 5.19: a-)ACF da Série 5 após uma diferença. b-)PACF da Série 5 após uma diferença.

e um ARIMA(5,1,1) para a Série 5.

5.2.4 Cenário 4

No Cenário 4, utilizou-se a mesma divisão do fluxo dos dados Cenário 2. Mas, em vez de todos os dados estarem representados unitariamente, a mediana dos dias de cada instante foi usada como medida representativa. A mediana não é tão sensível como a média às observações que são muito maiores ou muito menores do que as restantes (*outliers* ou valores atípicos). A mediana é indicada quando se tem valores 'muito grandes' ou 'muito pequenos', mesmo em pequena quantidade, pois, ela é uma medida mais robusta do que a média e não se carrega tanto esta variação que pode prejudicar a construção do modelo e/ou levar a decisões errôneas.

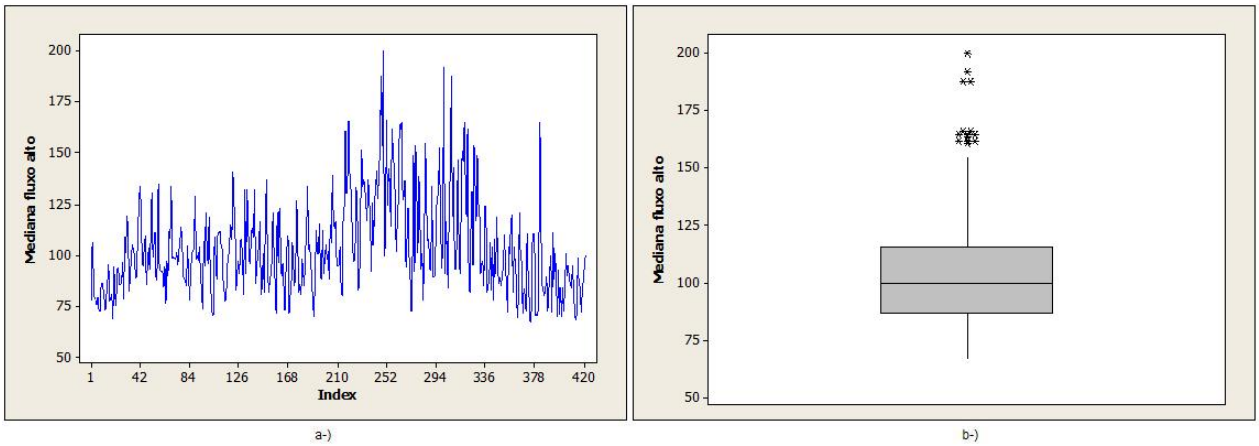


Figura 5.20: a-)A Série 6 ao longo do tempo. b-)Boxplot da Série 6.

Na Figura 5.20 a-) é possível observar a Série 6, que é composta pela mediana do período de maior intensidade de fluxo da rede (6:00 às 20:00). Pode-se notar que a mediana das observações é bem mais robusta que a média utilizada na Série 4, Figura 5.12 b-). Na Figura 5.20 b-), a variação do $q3$ até limite superior é menor comparado com a Figura5.12 a-). Com estas duas comparações, é nítido que a mediana não absorve tanto as variações dos dados como a média, levando a um modelo mais robusto.

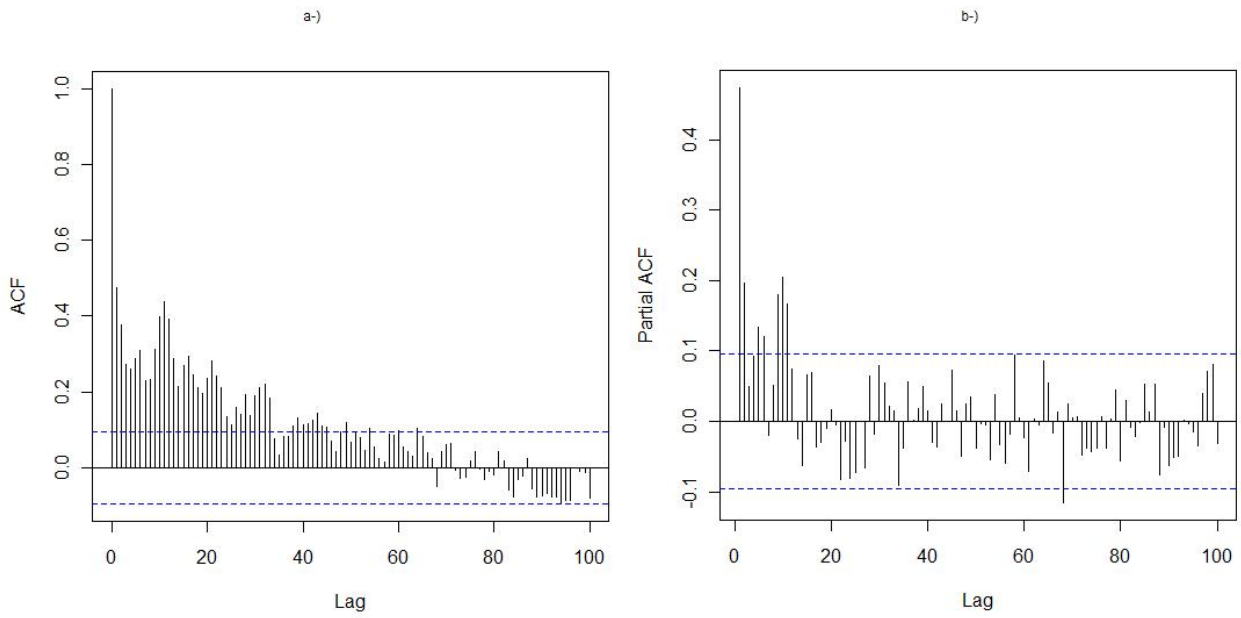


Figura 5.21: a-)ACF da Série 6. b-)PACF da Série 6.

A Figura 5.21 mostra as ACF e PACF da Série 6, indicando que os dados não são estacionários, mesmo quando utilizada a mediana. Foi necessário tomar uma diferença para que a série se tornasse estacionária, como apresentado na Figura 5.22.

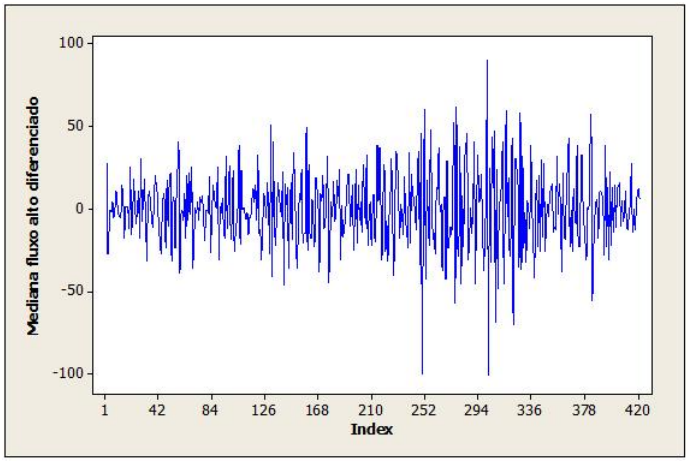


Figura 5.22: a-) Série 6 após uma diferença.

A ACF indica um *lag* significativo sugerindo um MA de ordem 1, e a PACF, um AR de ordem 3 ou 4, Figura 5.23.

Na Tabela 5.9, tem-se os AIC e BIC dos modelos ajustados e nota-se que os modelos ARIMA(4,1,1) e ARIMA (3,1,1) possuem mesmo AIC. Optou-se pelo segundo, uma vez que seu valor de BIC é menor.

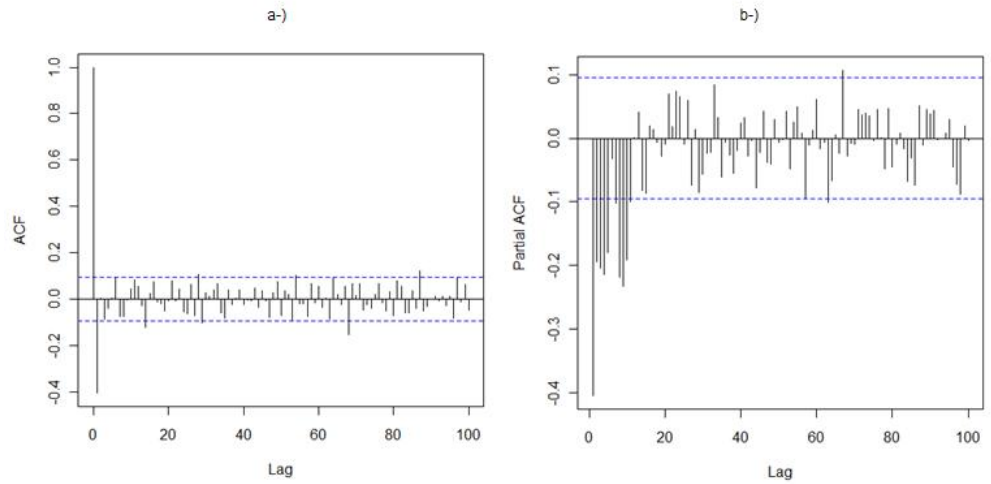


Figura 5.23: a-)ACF da Série 6 após uma diferença b-)PACF da Série 6 após uma diferença.

Para os dados com a quantidade de fluxo de pacotes menor(20:02 às 5:58), Série 7, realizou-se o mesmo estudo da Série 6. Na Figura 5.24 a-), nota-se que a série em sua maioria tem pouca variação ao longo do fluxo. Na Figura 5.24 b-) há apenas um ponto discrepante. Na Série 5, que utiliza a média das observações, tem-se mais pontos fora do $q3$ que usando a mediana.

Para ajustar um modelo, observaram-se as ACF e PACF da Série 7, Figura 5.25, indicando que a série necessita de uma diferença, como mostrado na Figura 5.26. Na

Tabela 5.9: Tabela de critério AIC e BIC para modelos da Série 6.

Modelo	AIC	BIC
(3, 0, 1)	3705,6	3717,7
(3, 1, 1)	**3693,2**	**3707,4**
(1, 1, 1)	3694,1	3700,2
(0, 1, 1)	3708,4	3710,4
(4, 1, 1)	3693,2	3711,4
(2,1,1)	3695,7	3705,9
(3, 0.32, 1)	3709,4	3723,6
(1, 0, 1)	3717,5	3721,6

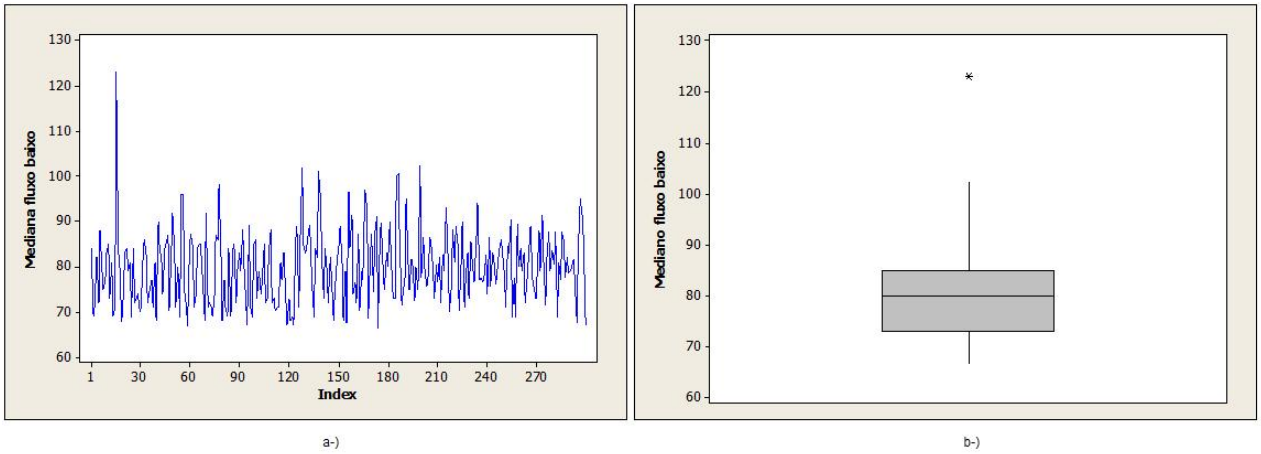


Figura 5.24: a-)Série 7 ao longo do tempo b-) Boxplot da Série 7.

sequência, observou-se a ACF e PACF da série após uma diferença, Figura 5.27, indicando um ARIMA(4,1,1), sendo que o na ACF tem-se um *lag* significativo indicando ordem um na parte média móvel e a PACF indica ordem 4 na parte autorregressiva.

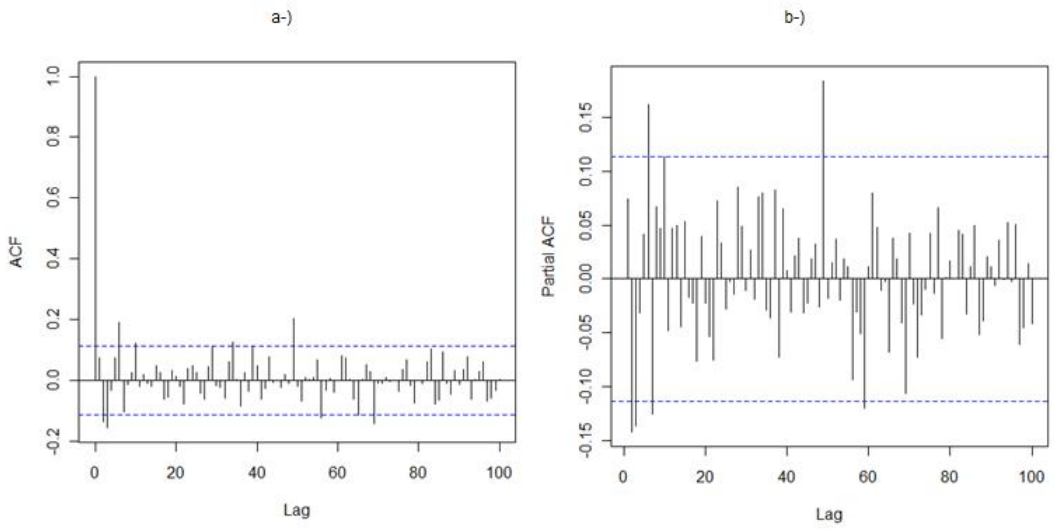


Figura 5.25: a-)ACF Série 7 b-) PACF Série 7.

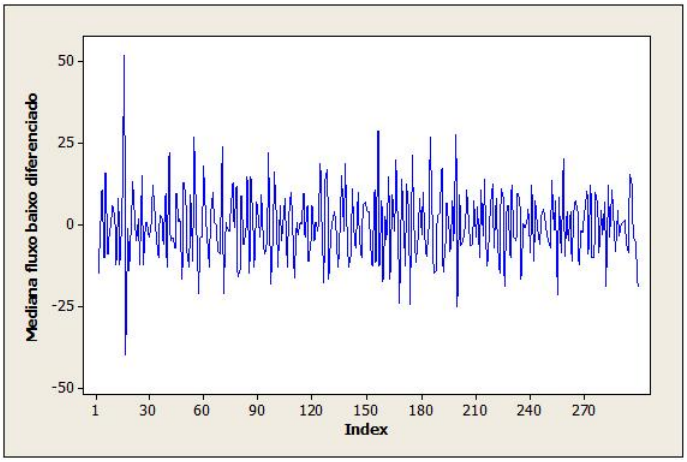


Figura 5.26: a-) Série 7 após uma diferença

Para confirmar o modelo mais adequado, calcularam-se o AIC e BIC de alguns modelos propostos. Na Tabela 5.10, notam-se os modelos que possuem maior ordem tem menor AIC. Conciliando-se as análises da ACF e PACF, indica-se um modelo de ordem menor, um ARIMA(4,1,1). Mas o ARIMA (3,1,1) pode também ser considerado, já que a diferença dos AIC e BIC é baixa e além do mais reduziria um parâmetro do modelo.

Após os exames descritos, escolheu-se o modelo ARIMA(3, 1, 1) para as Séries 6 e 7 com valores ϕ e θ distintos para cada caso.

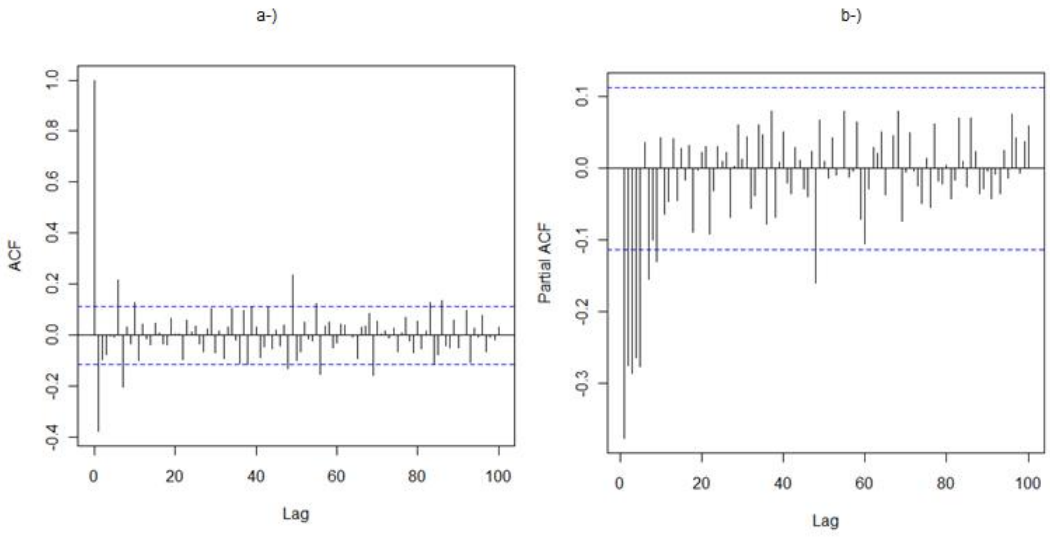


Figura 5.27: a-)ACF Série 7 após uma diferença b-) PACF Série 7 após uma diferença

Tabela 5.10: Tabela de critério AIC e BIC para modelos da Série 7.

Modelo	AIC	BIC
(0,1,1)	2112,5	2108,5
(5, 1, 1)	2106,0	2092,0
(6, 1, 1)	2100,8	2084,8
(7, 1, 1)	2097,0	2079,0
(4, 1, 1)	**2104,1**	**2092,1**
(3,1,1)	**2103,0**	**2093,0**
(3,0,1)	2107,6	2095,6
(1, 0, 1)	2114,5	2106,5

5.2.5 Caso Especial - Tráfego Atípico

Observando toda a amostra nos dias 20 e 21 tem-se dois períodos durante esses dias que possuem um fluxo intenso e com alta variabilidade, Figura 5.28. Optou-se por ajustar um modelo para esse fluxo caso o usuário possua um tráfego mais intenso que usual em um determinado período e necessite estimá-lo.

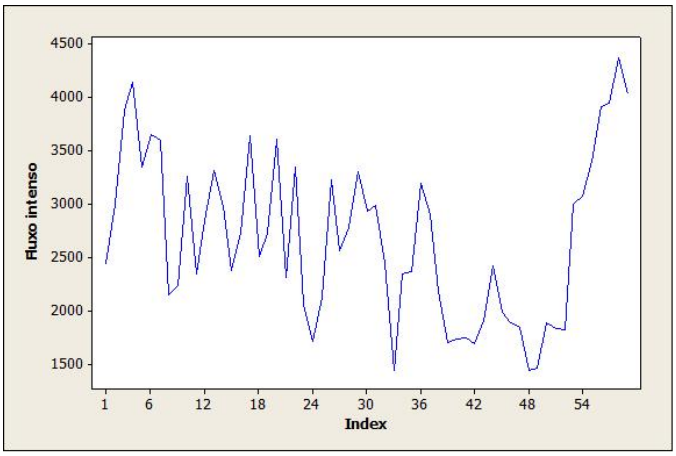


Figura 5.28: a-) Fluxo intenso de pacotes.

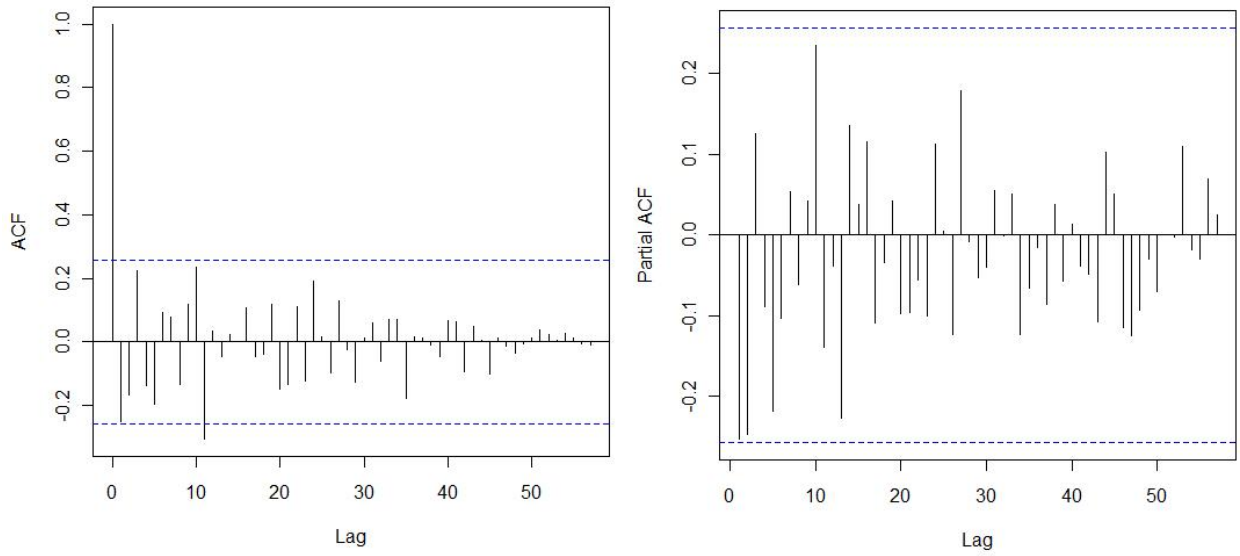


Figura 5.29: a-) ACF fluxo intenso diferenciado. b-)PACF fluxo intenso diferenciado.

Mesmo com alta variabilidade e autocorrelação dos dados, não foi constatado longa dependência. Observando a ACF e PACF da série após uma diferença, Figura 5.29,

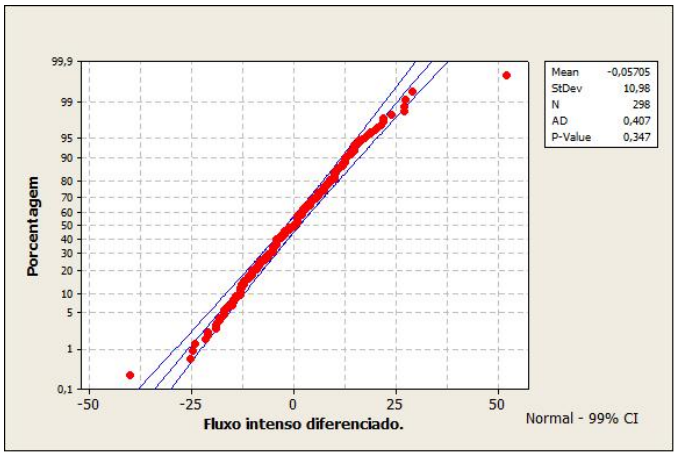


Figura 5.30: a-) Gráfico de probabilidade do fluxo intenso após uma diferença.

indica-se um modelo ARIMA(0,1,1) e analisando AIC de alguns modelos confirma-se este modelo é o mais indicado (possui o menor AIC dentre os outros).

Outro ponto é que os dados após uma diferença tem distribuição normal, Figura 5.30.

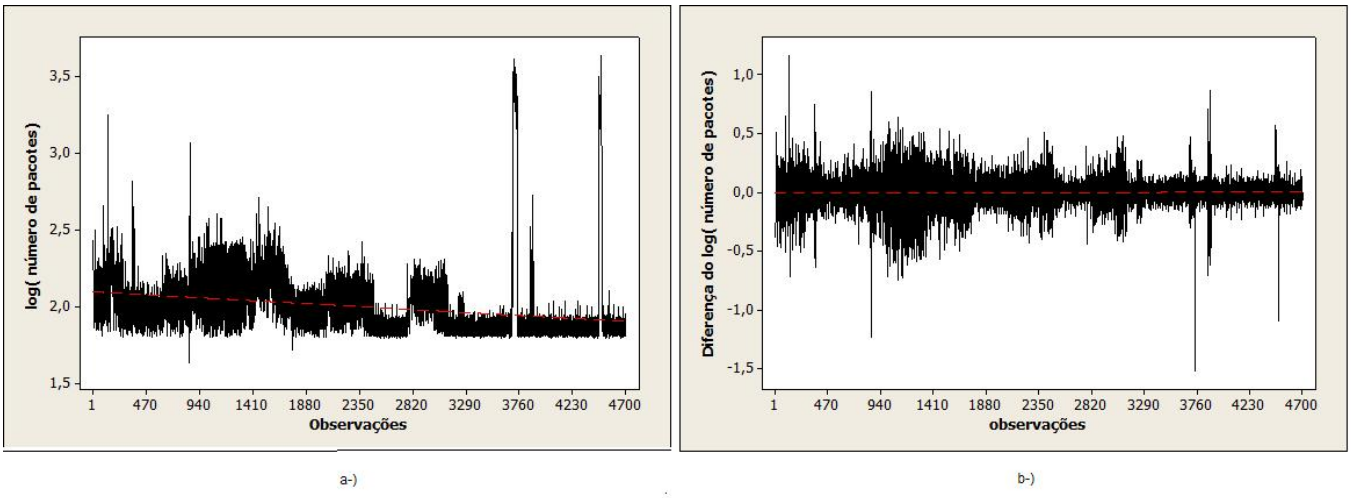
5.2.6 Pontos relevantes observados

Além dos 4 cenários descritos, foram consideradas outras formas de análise da rede, conforme segue

Uma maneira de minimizar a variabilidade dos dados é aplicar logaritmo nas observações. Na Figura 5.31 nota-se que a alta variabilidade dos dados ainda persiste mesmo após a série passar por uma diferença.

Como comentado e visto na Figura 5.5, os dias 20 e 21 tiveram variabilidade superior comparada com o restante dos dias analisados.

Para os Cenários 1 e 2 separou-se as séries por: dias típicos (dia 15 ao 19) e em dias de tráfego intenso (20 e 21). Mesmo com as divisões as variações do período que continha os dias 20 e 21, possuíam uma alta variação prejudicando a acuracidade do ajuste do modelo.



a-) b-)

Figura 5.31: a-) Logaritmo na base 10 das observações da Série 1. b-) Logaritmo na base 10 das observações da Série 1 - após uma diferença.

Por fim, nenhum dos dois métodos apresentou um resultado mais satisfatório do que o Cenário 4.

5.3 Monitoramento da Rede

As Séries 6 e 7 do Cenário 4, foram escolhidas para ilustrar a aplicação das cartas de controle CUSUM e MMEP, como descrito nas seções seguintes.

5.3.1 Gráfico CUSUM

Na Figura 5.32, tem-se a Série 6 monitorada por meio do CUSUM com valor alvo = 103,9, que é a média das observações. Nota-se que há alguns pontos que deslocam do valor - alvo, em sua maioria ultrapassa o limite superior. Com h =5 os limites são mais largos que com h = 4, ou seja, um pouco mais robustos. Mesmo com h =5 os limites estão muito comprimidos, levando vários pontos a estarem fora dos limites. Considerando que não houve nenhuma interrupção na rede, esse número expressivo de pontos indicam, que o gráfico de CUSUM não monitora bem esses dados que são bem autocorrelacionados como pode ser observado na função ACF da Figura 5.21. Nota-se que há vários *lags*

significativos (acima dos limites) indicando uma expressiva autocorrelação.

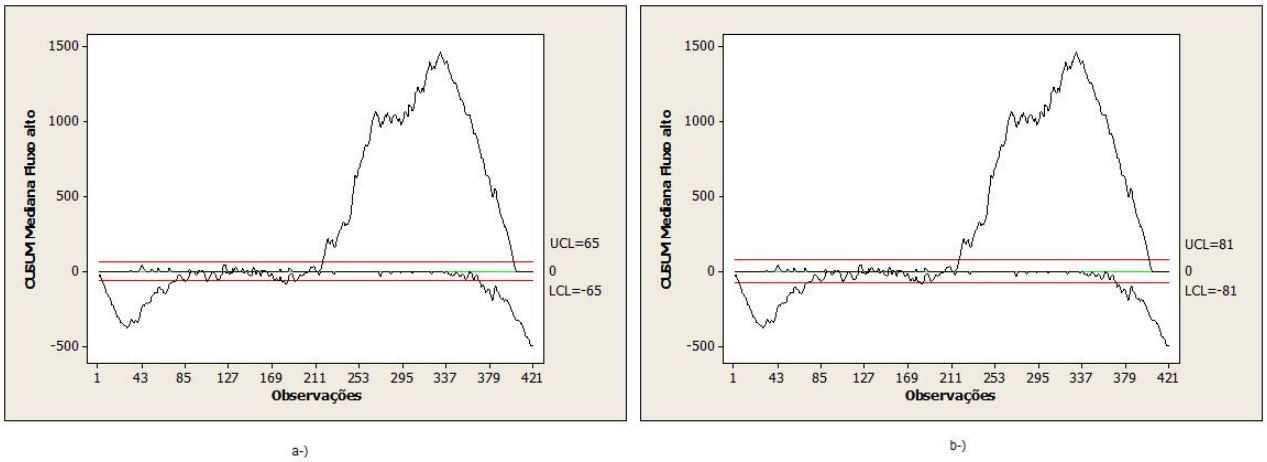


Figura 5.32: a-)CUSUM Série 6 com k =1/2 e h=4 b-)CUSUM Série 6 com k =1/2 e h=5

Nesse caso uma sugestão seria monitorar o resíduo do modelo ARIMA (3,1,1). Na Figura 5.33, analisando o resíduo, nota-se que poucos pontos ultrapassam os limites, satisfazendo o papel da ferramenta, porém para o administrador da rede esse processo não se torna trivial.

Na Série 7, utilizou-se o valor alvo de 79,9 (média do fluxo) para os gráficos da Figura 5.34. Observam-se poucos pontos fora dos limites, como era esperado sendo que essa série tem um fluxo mais bem comportado e também não houve queda na rede. Analisando-se os dois gráficos, com $h = 5$ detectou 4 pontos fora dos limites enquanto com $h = 4$, detectou-se 9. Para se ter menos falso-positivos o indicado é usar $h = 5$.

Neste caso não foi necessário utilizar o resíduo do modelo da Série 7, já que a ACF (Figura 5.25), não possui tanta autocorrelação (poucos *lags* acima dos limites) como a da Série 6.

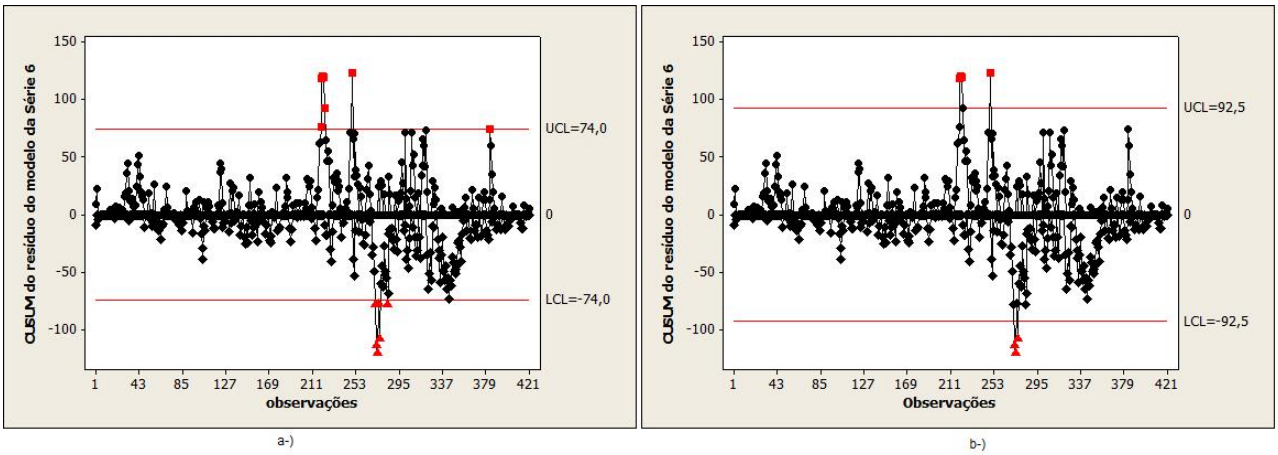


Figura 5.33: a-)CUSUM do residuo - Série 6 com k =1/2 e h=4 b-)CUSUM residuo Série 6 com k =1/2 e h=5

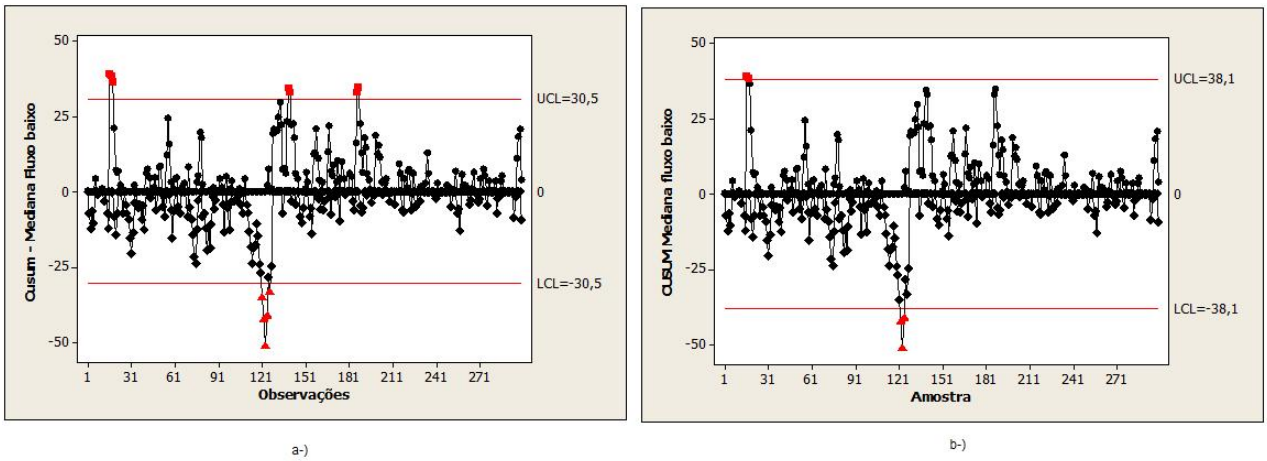


Figura 5.34: a-)CUSUM Série 7 com k =1/2 e h=4 b-)CUSUM Série 7 com k =1/2 e h=5

Realizou-se a mesma análise usando como valor alvo igual à mediana para as Séries 6 e 7 e não notou-se diferença significativa.

5.3.2 Gráfico MMEP.

Para as cartas de MMEP utilizou-se a média de cada série como valor alvo e variou-se o λ de 0,05 à 0,4 e o $L = 2$ e 3. Sendo λ o peso das observações e L a largura dos limites.

As cartas da Série 6 estão nas Figura 5.35 e 5.36, todas com valor alvo de 103,9. Nota-

se que com um λ menor o gráfico fica mais sensível a pequenas variações, comprimindo os limites, por exemplo, na Figura 5.35 b-) quando comparada com a 5.36 b-), outro ponto é que quando maior o valor de λ mais peso para as informações mais recentes. Quando fixado o λ e se altera o L, não há variação significativa, por exemplo , na Figura 5.35 b-) com a Figura 5.36 a-) fixou-se $\lambda = 0{,}05$ e variou-se L= 3 e L= 2 respectivamente.

Na Figura 5.36 b-), nota-se que com $\lambda = 0{,}4$, os limites são maiores do que com $\lambda =$ 0,2 e 0,05, sendo menos sensível a pequenas variações. Com o peso igual a 0,4 obteve-se um resultado satisfatório, sendo que os pontos que a carta alerta são observações que possuem uma variação significativa em relação às restantes.

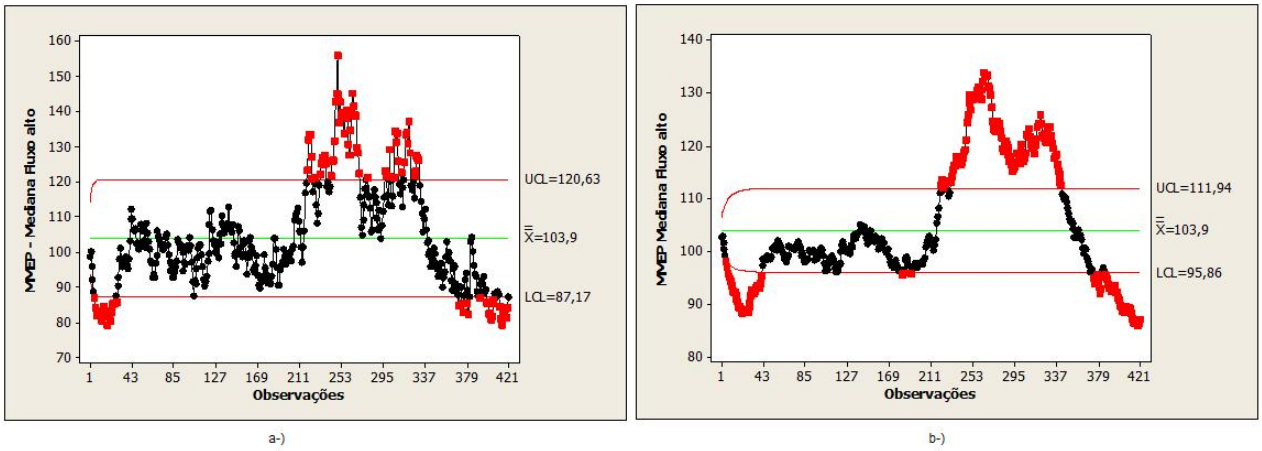


Figura 5.35: a-)MMEP Série 6 com L = 3 e $\lambda = 0{,}2$ b-)MMEP Série 6 com L = 3 e $\lambda =$ 0,05

A Série 7 possui observações com um grau menor de variação comparada com a Série 6, e esta menor variação é melhor para o monitoramento dos dados.

Na Figura 5.37 e 5.38 adotou-se o valor alvo de 79,9 que é a média das observações da Série 7. Nota-se que quanto menor o λ mais sensível o gráfico se torna e alterando o tamanho do L não se tem tanto impacto nos limites. Com $\lambda = 0{,}4$ os limites de controle

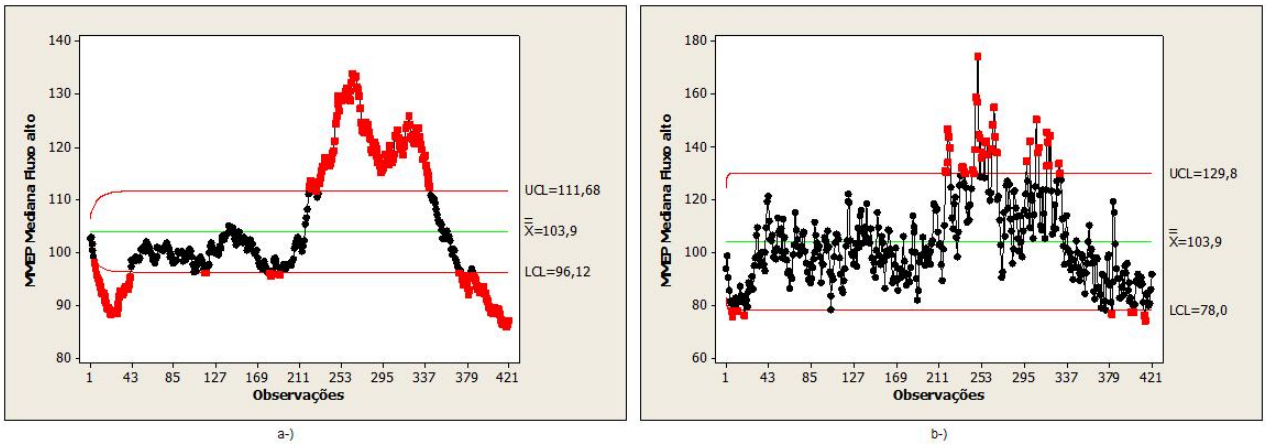


Figura 5.36: a-)MMEP Série 6 com L = 2 e λ = 0,05 b-)MMEP Série 6 com L = 3 e λ = 0,4

são maiores que com λ = 0,2 e 0,05, tornando a ferramenta mais robusta a pequenas variações, na Figura 5.38 b-) tem-se 2 pontos fora dos limites enquanto as cartas com λ menor tem um número maior de pontos fora dos limites. Por fim com L = 3 e o peso igual 0,4 o gráfico teve um desempenho satisfatório.

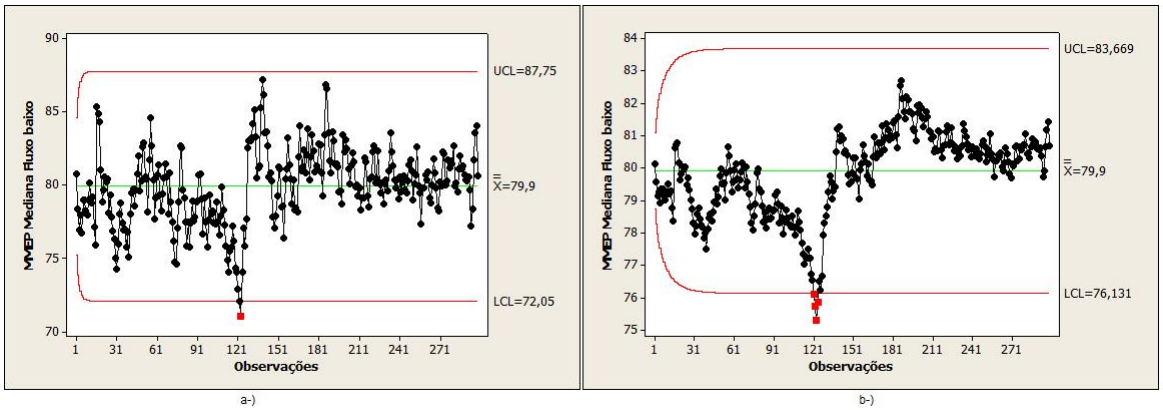


Figura 5.37: a-)MMEP Série 7 com L = 3 e λ = 0,2 b-)MMEP Série 7 com L = 3 e λ = 0,05

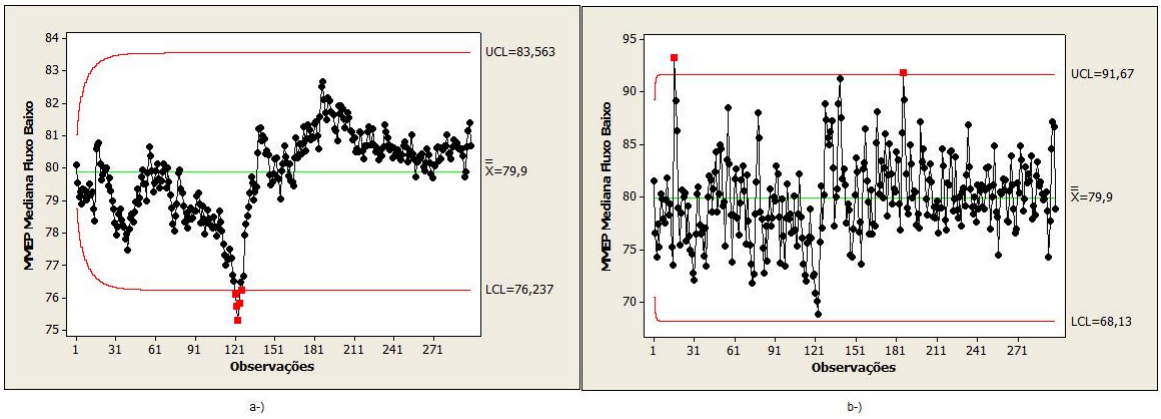


Figura 5.38: a-)MMEP Série 7 com L = 2 e λ = 0,05 b-)MMEP Série 7 com L = 3 e λ = 0,4

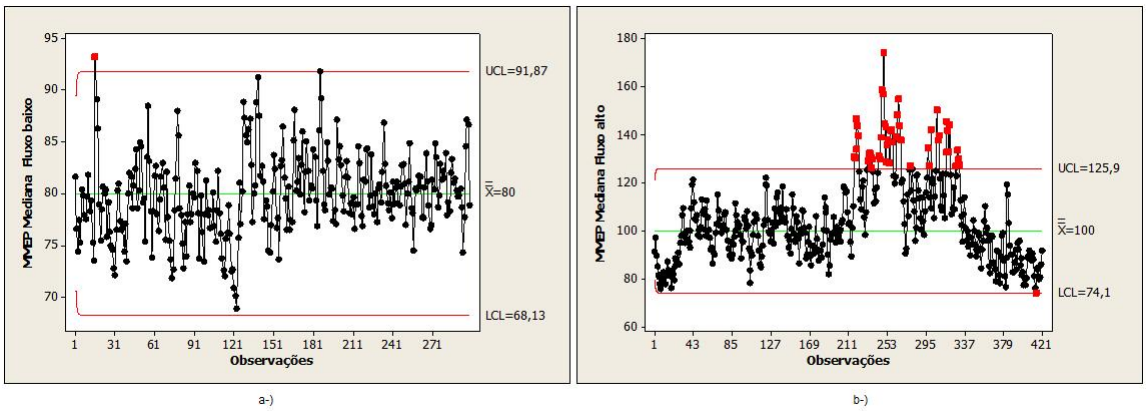


Figura 5.39: a-)MMEP Série 7 com L = 3 e λ = 0,4, com valor alvo = mediana b-)MMEP Série 6 com L = 3 e λ = 0,4, com valor alvo = mediana

Na Figura 5.39, tem -se a Série 6 e a Série 7 usando como valor alvo a mediana de cada série, não se nota diferença significativa em relação ao uso da média.

Em resumo para o caso das séries analisadas existem períodos instáveis e deseja-se maior peso nas informações mais recentes, por isso optou-se pelos seguintes parâmetros dos gráficos: $L = 3$ e $\lambda = 0{,}4$. O λ maior é indicado para modelos onde a série é mais instável gerando modelos mais robustos, Figura 5.38 b-) e 5.36 b-).

Os pontos que o modelo evidenciou fora do limite são os pontos que mais variaram

em relação à média, indicando que esses são pontos de atenção, mostrando que o método é satisfatório para detectar possíveis interrupções.

Capítulo 6

Discussão

Neste capítulo, é feita a discussão de aspectos específicos da pesquisa cujo entendimento depende do desenvolvimento descrito anteriormente. Por exemplo, foi necessário analisar diversos cenários e várias séries de dados para que a modelagem pudesse ser concluída e o resultado pudesse ser trazido à discussão. Considerou-se que a consolidação destas informações num capítulo próprio permite melhor entendimento do trabalho, além da comparação entre métodos.

Escolheu-se a análise de correlação linear, para reduzir o número de variáveis do modelo mrCEP pois é um método apropriado a este fim. Para o caso estudado, sua utilização reduziu de 4 variáveis para uma.

A utilização de uma diferença na série foi essencial para o ajuste de modelos mais significativos para os cenários estudados. Observando-se as ACF associadas aos AIC, os modelos de longa dependência foram insatisfatórios.

A construção de um modelo de séries temporais e monitoramento da rede foi conduzida como segue: a) os dados foram separados em períodos do dia com maior e menor fluxo de pacotes; b) foi usada a mediana, uma medida mais robusta a variações que a média aritmética, para representar a série. Estes passos auxiliaram o ajuste de modelos aderentes e a estimativa de limites de controle eficientes. No Cenário 4, conciliou-se a divisão por períodos e a mediana. Com esses pontos, pode-se ajustar um modelo que representa bem a rede analisada. Outro ponto levantado é que foi a única técnica em que, após uma

diferença, as séries apresentaram normalidade ao um nível de confiança de 1%, Figura 6.1 e 6.2.

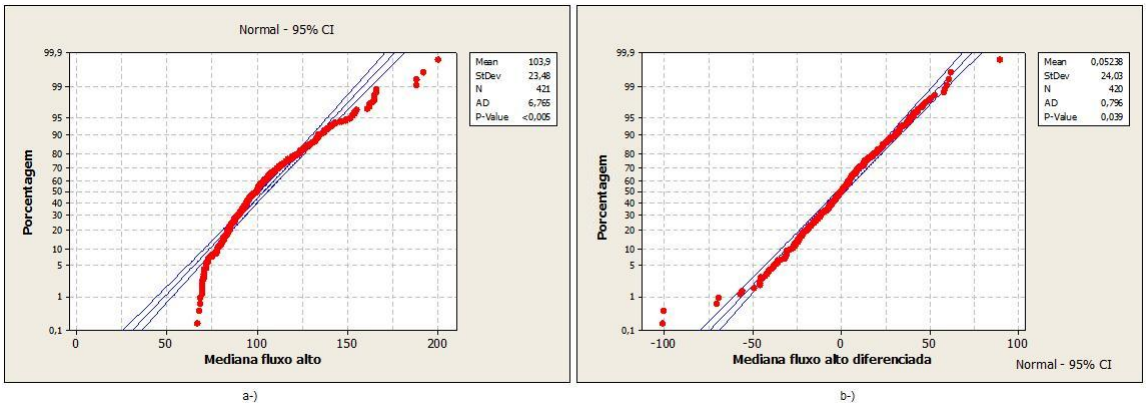


Figura 6.1: a-)Gráfico de probabilidade da Série 6 b-) Gráfico de probabilidade da Série 6 após uma diferença

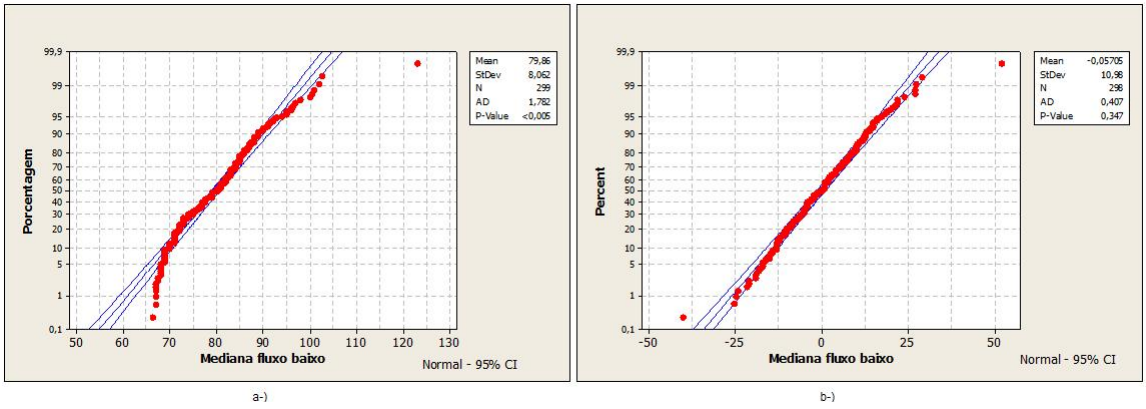


Figura 6.2: a-)Gráfico de probabilidade da Série 7 b-) Gráfico de probabilidade da Série 7 após uma diferença

Por fim o modelo que melhor se ajustou aos dados foi um ARIMA (3, 1, 1). Para a escolha do modelo considerou-se as características da ACF e PACF e sempre procurou-se um menor número de parâmetros. Cada modelo ARIMA (3,1,1) da Série 6 e Série 7

possui valores distintos das constantes ϕ e θ da parte autorregressiva e de média móvel, na Tabela 6.1. Em geral são usados para o ajuste completo do modelo.

Tabela 6.1: Tabela Valores dos Parâmetros dos Modelos.

Modelo	AR(1)	AR(2)	AR(3)	MA(1)
Série 6	0.1989	0.0388	-0.1102	-0.9018
Série 7	0.0503	-0.1483	-0.1593	-0.9776

O CUSUM não se mostrou muito adequado para dados muito autocorrelacionados e vários pontos ultrapassaram os limites sem necessidade. O fato de usar o resíduo do modelo não é um ponto positivo para o gerenciamento da ferramenta. O MMEP mostrou resultados mais satisfatórios que o CUSUM, sendo que apresentou um número adequado de alertas. Evidenciando apenas os pontos que mais variaram da amostra. Isso provavelmente se deve às propriedades do MMEP, já que é indicado para monitorar séries temporais. Ele incorpora um ARIMA (0, 1, 1); além de ser aconselhável para observações autocorrelacionadas e livres de distribuições. Comparando o desempenho do MMEP com os gráficos usados no mrCEP, nota-se que no mrCep houve inúmeros falsos alarmes quando usados os limites convencionais das cartas de $\bar{X}$ e amplitude, provavelmente isto ocorreu porque estas cartas não representam bem dados autocorrelacionados. Já quando usa-se o mrCEP com os limites não convencionais (ajustados especialmente para o conjunto de dados), nota-se que o resultado foi muito mais satisfatório do que com os limites convencionais, possuindo poucos alarmes. Uma vantagem do MMEP para as cartas com limites não convencionais, é que facilmente pode-se deixar os limites mais sensíveis ou mais robustos, do mesmo conjunto de dados alterando apenas o valor de L ou e o de λ, no caso dos limites do mrCEP eles são únicos para aquele conjunto de dados. As cartas de MMEP com os parâmetros escolhidos são mais sensíveis que o mrCEP com os limites ajustados. Caso o AR queira limites mais robustos, uma maneira seria aumentar o valor dos parâmetros do MMEP.

Capítulo 7

Conclusão

O primeiro objetivo desta pesquisa foi determinar possibilidade de se monitorar estatisticamente uma rede de computadores com um número menor de variáveis em relação ao modelo mrCEP, que examina 23 medidas. Foi possível identificar apenas 4 variáveis suficientes para o monitoramento estatístico e, com uso da análise de correlação linear, verificou-se que uma delas pode representar bem as demais. Obteve-se assim uma redução muito satisfatória da quantidade de variáveis a processar, simplificando os processos de monitoramento da rede.

O segundo objetivo era aprimorar a qualidade do monitoramento com uso de melhores técnicas estatísticas, tomado o mrCEP como base. Foi feito um estudo de caso de uma rede departamental de uma grande Universidade, considerada representativa para o conjunto de redes similares em estrutura ou operação. Os métodos de análise de séries temporais, em especial o modelo ARIMA, se mostraram adequados para modelar a rede cuja série não se mostrou estacionária e não apresentou características de longa dependência.

Consequentemente, foi possível monitorar a rede por meio de cartas de CUSUM e MMEP. As cartas de CUSUM exigem técnica mais complexa para construção e interpretação e apresentaram mais alarmes falsos, sendo consideradas inadequadas para os propósitos deste estudo. As cartas MMEP mostraram melhores resultados, com menor complexidade, e são indicadas para dados autocorrelacionados, conseguindo uma melhor aderência em relação às cartas de amplitude e média aplicadas no mrCEP. Portanto,

esta investigação concluiu que as cartas MMEP podem ser utilizadas com benefícios para substituir as anteriores no monitoramento de redes de computadores com características similares àquela examinada neste trabalho.

Por fim, a pesquisa atingiu os objetivos propostos e mostrou novas formas de analisar a rede, abrindo possibilidades de monitoramento e entendimento até então não investigadas.

Como trabalhos futuros, sugere-se monitorar uma rede em tempo real por meio da carta MMEP, comparando os valores obtidos e os valores previstos pelos modelos ARIMA para avaliar o desempenho destes últimos.

Referências Bibliográficas

[1] Angelis, A. F., Um modelo de tráfego de rede para aplicação de técnicas de Controle Estatístico de Processos. 2003. 176 f. Tese (Doutorado), Instituto de Física de São Carlos, Universidade de São Paulo, 2003.

[2] Becchi, M. From Poisson Processes to Self-Similarity: a Survey of Network Traffic Models. Technical report, Citeseer, 2008.

[3] Box, G. E. P.; JENKINS, G. M. Time series analysis forecasting and control. San Francisco: Holden- Day, 1976. Edição revisada.

[4] Brockwell P. J; Davis R. A, Introduction to time series and forecasting. Springer Texts in Statistics, Book, Second Edition 2002.

[5] Bussab, W.O.; Morettin, P.A. Estatística básica. 5 ed. São Paulo: Saraiva, 2005. 321 p.

[6] Chanet, Reinaldo; Bonvino, Heloísa; Freire, Clarice Azevedo de Luna; Chanet, Eugênia M. Reginato. Análise de modelos de regressão linear com aplicações. Campinas: Editora da Unicamp, 1999, 356p.

[7] Cowperwait and A.V Metcalfe, Introductory Time Series with R, Springer, 2009.

[8] Crato, N. Aplicações de Modelos de Memória Longa, 9^a Escola de Séries Temporais e Econometria Belo Horizonte, material didático, 2001.

[9] Creeger, M. CTO Roundtable:Malware Defense. Communications of the acm. April 2010.

[10] Cunha A. C. Gráficos de Controle CUSUM: um enfoque dinâmico para a análise estatística de processos. Universidade Federal de Santa Catarina, Florianópolis, 2003.

[11] Elagha, H.; AlShafee, M. On the Self-Similar Nature of ATM Network Traffic,Royal Scientific Society, Information Technology Centre. Issues in Informing Science and Information Technology Volume 4, 2007.

[12] Guiesi,G. J. Caracterização de Tráfego de Rede, Trabalho de Conclusão de Curso, Univesidade Estadual de Londrina, 2004.

[13] Isakssont C., Mengtt Y. and Dunhamt, M.H. Risk Leveling of Network Traffic Anomalies,IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.6, June 2006.

[14] Montgomery C. D. Introdução ao Controle Estatístico da Qualidade. 4^a ed, Universidade do Arizona. 2004, Rio de Janeiro 2003.

[15] Moreira F.J, Schwengber C. Proposta de uma carta de controle estatístico de dados autocorrelacionados. XIV Encontro Nac. de Eng. de Produção - Florianópolis, SC, Brasil, ENEGEP 2004.

[16] Morettin, P. A.; Toloi, C. M. C. Análise de Séries Temporais. 2. ed. São Paulo: Atual Editora, 2004.

[17] Paxson, V. Fast Approximation of Self-Similar Network Traffic. Technical report,Lawrence Berkeley Laboratory and EECS Division, University of California, 1995.

[18] Rutka G. Some Aspects of Traffic Analysis used for Internet Traffic Prediction, Faculty of Electronics and Telecommunication, Riga Technical, University, 2009.

[19] Silva, A. Análise de Tendência de Tráfego Origem Destino em Redes IP utilizando Estatística Multivariada, Centro de Informática, Universidade Federal de Pernambuco (UFPE).

[20] Spagnol, R. L. Monitoramento de Tráfego de Redes de Computadores através de Métodos Estatísticos, Trabalho de conclusão de curso, Universidade Estadual de Campinas, 2009.

[21] Spagnol, R. L. Redução do número de variáveis a serem monitoradas do tráfego de rede de computadores, Universidade de Campinas, 2010.

[22] Trevisan E.S., Souza R.C, Souza L.R, Estimação do parâmetro "d"em modelos ARFIMA. Pesquisa operacional, Vol.20, 2000.

[23] Wei. William. W. S, Time Series Analysis - Univariate and Multivariate Methods, Departament of Statistcs The Fox School od Business and Management, Second edition, 2006.

[24] Xuan, Y., Shin, I. Detecting Application Denial-of-Service Attacks: A Group Testing Based Approach, 2009.

[25] Yokoyama , R. Limites variáveis para controle estatístico de processos aplicado a redes de Computadores. 2006. Trabalho de Graduação (TG), Instituto de Geociências e Ciências Exatas de São Carlos, Universidade Estadual Paulista Júlio de Mesquita Filho', 2006.