UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP
Faculdade de Tecnologia e Inovação - FT

Mestrado em Tecnologia

Modelagem do Tráfego de Redes de Computadores através de Métodos Estatísticos

Dissertação de Mestrado

Renata Lussier Spagnol

Orientador: Dr. André F Angelis

Coorientadora: Dra. Laura Léticia R. Ramos.

Limeira - SP

Outubro/2011

UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP
Faculdade de Tecnologia e Inovação - FT

Modelagem do Tráfego de Redes de Computadores através de Métodos Estatísticos

Este exemplar corresponde a Dissertação que será apresentada na Faculdade de Tecnologia da UNICAMP como requisito parcial para a obtenção do Título de Mestre em Tecnologia e Inovação.

Banca Examinadora:

Limeira - SP
Outubro/2011

Agradecimentos

Sumário

1 Introdução **6**

2 Revisão Bibliográfica **8**

3 Bases Teóricas **13**

3.1 Análise de Correlação Linear 13

3.2 Modelos de Séries Temporais 14

3.2.1 Processos de Média Móvel 15

3.2.2 Processos Auto-regressivos 16

3.2.3 Modelos de Box-Jenking - ARMA 16

3.2.4 Modelos de Box-Jenking - ARIMA 17

3.2.5 Modelos ARFIMA 18

3.2.6 Critérios de seleção de Modelos 19

3.3 Gráficos de Controle Estatístico do Processo 21

3.3.1 Gráficos de Controle da Soma Cumulativa - CUSUM 21

3.3.2 Gráficos de Controle da Média Móvel Exponencialmente Ponderada - MMEP 23

4 Materiais e Métodos **24**

4.1 Materiais 24

4.2 Métodos 25

4.2.1 Amostragem e dados 25

4.2.2 Organização dos Dados para Análise Estatística 27

4.2.3 Análise de Correlação Linear 27

4.2.4 Ajustando o Modelo de Séries Temporais 28

4.2.5 Gráficos de CEP - CUSUM E MMEP 29

5 Análise Estatística da Rede **29**

5.1 Análise de correlação e escolha do modelo do fluxo de rede - fonte destino (Série 1) 29

5.1.1 **Cenário 1** 32

5.1.2 **Cenário 2** 33

5.1.3 **Cenário 3** 38

5.1.4 **Cenário 4** 45

6 Discussões e Resultados Finais **50**

7 Referências **53**

8 Apêndice **53**

Lista de Figuras

1 Fluxo desde a coleta dos dados até a análise (Fonte: ANGELIS[1]) 26
2 a-)Gráfico da Série ao longo do tempo. b-)Boxplot da Série. 30
3 a-)Gráfico da Série 1 ao longo do tempo. b-)Boxplot da Série 1. 31
4 a-)Gráfico da Função de Autocorrelação da Série 1. b-) Gráfico da Função de Autocorrelação Parcial da Série 1. 32
5 a-)Gráfico da ACF da Série 1 diferenciada. b-) Gráfico da PACF da Série 1 diferenciada. 33
6 a-)Gráfico dispersão do fluxo por dia(início às 0:00 e termino às 23:58). b-) Dotplot - Fluxo de pacotes por dia. 34
7 a-)Boxplot da base Série 2. b-)Gráfico da Série 2 ao longo do tempo. 35
8 a-)ACF da Série 2. b-)PACF da Série 2 35
9 a-)ACF da Série 2 diferenciada. b-)PACF da Série 2 diferenciada 36
10 a-)Boxplot da Série 3. b-)Gráfico da série Série 3 ao longo do tempo. 37
11 a-)ACF da Série 3. b-) PACF da Série 3. 37
12 a-)Gráfico da Série 3 diferenciada uma vez. b-)ACF da Série 3 diferenciada 38
13 a-)Boxplot da Série 4. b-)Gráfico da Série 4 ao longo do tempo 40
14 a-)ACF de Série 4. b-) PACF de Série 4. 40
15 Gráfico dá Série 4 diferenciada. 41
16 a-)ACF de Série 4 diferenciada. b-) PACF de Série 4 diferenciada 42
17 a-)Gráfico da Série 5. b-)Boxplot da Série 5. 42
18 a-)ACF da Série 5. b-)PACF da Série 5. 43
19 Gráfico da Série 5 diferenciada. 44
20 a-)ACF da Série 5 diferenciada. b-)PACF da Série 5 diferenciada. 44
21 a-)A Série 6 ao longo do tempo. b-)Boxplot da Série 6. 45
22 a-)ACF da Série 6. b-)PACF da Série 6. 46
23 a-) Série 6 diferenciada. 46
24 a-)ACF da Série 6 diferenciada b-)PACF da Série 6 diferenciada. 47
25 a-)Série 7 ao longo do tempo b-) Boxplot da Série 7. 48
26 a-)ACF Série 7 b-) PACF Série 7. 48
27 a-) Série 7 diferenciada 49
28 a-)ACF Série 7 diferenciada b-) PACF Série 7 diferenciada 49
29 a-)Gráfico de probabilidade da Série 6 b-) Gráfico de probabilidade da Série 6 diferenciada 51

30 a-)Gráfico de probabilidade da Série 7 b-) Gráfico de probabilidade da Série 7 diferenciada 51
31 a-) Fluxo intenso de pacotes. 52
32 a-) ACF fluxo intenso diferenciado. b-)PACF fluxo intenso diferenciado. 53
33 a-) Gráfico de probabilidade do fluxo intenso diferenciado. 53

Lista de Tabelas

1 Amostra da tabela de dados tratada 27
2 Variáveis correlacionadas 28
3 Características das funções ACF e PACF para processos estacionários 29
4 Variáveis correlacionadas 30
5 Pontos discrepantes da amostra. 31
6 Medidas de localização e dispersão. 31
7 Tabela de critério AIC e BIC para modelo Série 1. 33
8 Tabela de critério AIC e BIC para modelo da Série 2 36
9 Tabela de critério AIC e BIC para modelo estudados em Série 3. 39
10 Tabela de critério AIC e BIC para modelo estudados em Série 4. 41
11 Tabela de critério AIC e BIC para modelos estudados na Série 5. 43
12 Tabela de critério AIC e BIC para modelos estudados na Série 6. 47
13 Tabela de critério AIC e BIC para modelos estudados na Série 7. 50
14 Tabela Parâmetros do modelo. 51

Resumo

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

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. Este processo trouxe evidentes ganhos à sociedade, desde o acesso global a dados até as comodidades do governo e do comércio eletrônico. 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 esta dependência, tem-se como conseqüência 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), neste texto considerado a figura responsável pelo projeto, implementação e operação da RC.

Se um problema ocorre com a rede, o AR deve resolvê-la 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, da qual se comentou 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 (virtuais). Nesta direção, foi encontrado na literatura um modelo RCs para aplicação de CEP [1], aqui 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, no caso verificada a não-normalidade dos dados. Estas adaptações, embora corretas do ponto de vista matemáticos, 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. 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, este número representa dificuldades operacionais para os ARs.

Esta pesquisa se propôs investigar e aperfeiçoar o mrCEP, dados os pontos anteriormente men-

cionados, 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 da base dos dados a análise de correlação resultou em resultados positivos, diminuindo realmente o número de variáveis. A segunda linha investigativa seguiu pela melhoria da informação disponível. Buscou-se responder se cartas 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 a 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 se têm os métodos estatísticos utilizados, justifica-se esta enumeração pela complexidade dos métodos utilizados em função das diferenças entre as áreas de redes (formação do AR) e estatística (foco desta dissertação). Materiais e Métodos, mostra em detalhes os recursos de software utilizados e aspectos relevantes da base de dados. Em seguida, análise da rede, que contempla desde da análise descritiva dos dados, o modelos de previsão e forma de monitoramento até os resultados abtidos. Em Discussão dos Resultados aponta-se a diferença entre os modelos ajustados. Na Conclusão apontando o modelo proposto para a série do fluxo de rede da fonte para o destino. O texto é fechado com Referências Bibliográficas.

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, sendo elas causadas por quedas ou ataques de intrusos. A seguir estão exemplificados alguns desses estudos, sendo que a maioria utiliza algum método estatístico como ferramenta principal para o entendimento da rede.

Uma dos meios para melhorar a qualidade da rede é oferecer um serviço mais seguro, robusto 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[7], 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 este desenvolvimento dificulta a sua identificação.

Um ataque pode se manifestar de várias maneiras. Uma delas é como um congestionamento intencional na rede. Alguns métodos de defesa procuram detectar este 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[26], é proposto um método sofisticado, onde se usa um teste que permite a detecção rápida e precisa de ataques através dos métodos clássicos de grupo de teste (GT). Para a contrução 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, como descrito em Paxson[19]. Porém, este 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 distribuição 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[19]. 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 Poissiano e explica como o modelo de auto-similar difere dos tradicionais.

Rutka[20] 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 tratativa a aplicação de modelos auto-regressivos de média móvel (ARIMA) para prever futuros tráfegos.

Elagha[11], estudou o tráfego da rede ATM buscando encontrar algumas particularidades do tipo de série. A princípio, tinha-se a suspeita que os dados possuiam 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 auto-similaridade. Não foram encontrados auto-similaridades em alguns pontos, mas as observações apresentaram características de estruturas de fractais ou multifractais. E, 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 Isaksson[13], para identificar as tentativas ou ataques sobre um sistema de computadores ou na rede. Tem-se como objetivo identificar as anomalias da rede, através da mudança de comportamento dos dados. O modelo proposto por ele detecta os problemas baseando-se na freqüência de ocorrência e, em seguida, mede o desvio da distância de cada anomalia no espaço dos dados. O nível de risco com que a anomalia está associada é avaliado pelo desvio, de quanto os dados que foram previstos ficaram longe dos dados sem avaria. O modelo estatístico foi construído com base no Modelo *Markov Extensible* (EMM), que é baseado em uma técnica de modelagem no espaço-temporal, e trabalha com base nas ocorrências das transições de estados na cadeia de *Markov* gerada. Foi incluído no estudo técnicas de *clustering*, associada 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 à 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 comprender o comportamento da rede, que é bem variado, outro ponto que dificulta seu monitoramento é que geralmente têm-se várias variáveis que influenciam no tráfego e, para criar um modelo que o represente bem, é necessário escolher as variáveis significativas. Em Silva [21], 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 deste número existem alguns métodos estatísticos, por exemplo, é retirar do modelo as variáveis correlacionadas.

Em Angelis [1], estudou-se um método para monitorar as redes de computadores e assim prever futuras interrupções através 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). 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 a rede observada(interna ou externa), de pacotes e dos fluxos do protocolo IP e outros e de protocolos de trnsporte 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 entre as variáveis pois, com esta análise, ele conseguiria otimizar seu estudo, evitando de monitorar ao mesmo tempo variáveis que estão correlacionadas, vista 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 os dados não seguem uma distribuição Gaussiana, 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, através da desigualdade de Chebyshev, cujo cálculo é pouco conhecido e pouco trivial.

Um método indicado para facilitar os cálculos dos limites é a utilização de gráficos de Média Móvel Exponencialmente Ponderada e de Soma Cumulativa (Montgomery [15]), 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[27] estudou a fundo os métodos de CEP para gerar cartas mais sensíveis às variações do tráfego da rede, sem considerar e possível redução da quantidade das variáveis que deveriam ser monitoradas. Inicialmente utilizou limites variáveis para os gráficos de atributos, porém isto não foi satisfatório, já que estes gráficos são utilizados para a construção itens com a classificação conformes ou não conforme. Esta classificação segundo seus estudos é impraticável para um tráfego de uma RC. Analisando outros tipos de gráficos de CEP, Yokoyama [27], 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 este gráfico é mais trabalhoso que o R.

Um problema encontrado por Angelis [1] foi que os dados não seguiam uma distribuição normal. Este fato as vezes pode ser alterado pela maneira como os dados foram amostrados. A amostragem dos dados pode ser feita de duas maneiras: tomar-se amostras pequenas e freqüentes ou tomar-se amostras grandes e pouco freqüentes. Yokoyama [27] optou pelo primeira maneira. Neste caso através do Teorema do Limite Central, os dados deveriam convergir para uma distribuição normal,mas não se verificou esta convergência. A maneira adotada para minimizar o efeito de não-normalidade foi construir limites variáveis. Assim conseguiu que os limites acompanhassem a flutuações do tráfego com boa sensibilidade, sendo cortados somente em situações aparentemente atípicas do comportamento decorrente da rede. Para construir limites variáveis foi necessário alterar a maneira como eram calculados os limites em Angelis[1]. Para facilitar os cálculos poderia ter utilizado o gráfico de MMEP, sendo que uma das principais características deste gráfico é que seus limites são variáveis.

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 [16], diz que uma delas é que as observações devam ser independentes.

Alguns processos possuem observações que apresentam autocorrelação. Para este caso, as ferramentas do CEP como $\bar{X}$ e R não funcionam corretamente, podendo resultar em 'falsos alarmes'. Moreira [16] sugere que, no caso de informações autocorrelacionadas, é necessário tratá-las antes e depois controlá-los estatisticamente. Como forma de tratamento, sugere primeiramente fazer um ajuste de um modelo de série temporal ARIMA, que remova a autocorrelação dos dados. E deste modelo obtem-se os resíduos independentes conhecidos como 'ruído branco', e depois aplica-se um gráfico de controle. Uma sugestão para monitorar os resíduos é usar os gráficos de controle de Soma Acumulativa (CUSUM) ou o Média Móvel Exponencialmente Ponderada (MMEP), já que estes detectam pequenas mudanças na média do processo. Discutiu-se que 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, segundo Moreira[16].

Anteriormente a esta pesquisa, realizou-se no ano de 2009, um estudo prévio sobre o comportamento da rede, Spagnol[22]. Utilizou-se dados de uma rede (LAN), disponíveis de acordo com Cowperwait[5], e obtiveram-se resultados satisfatórios quanto ao monitoramento ARFIMA. Foi comprovado que os dados de um fluxo de rede são autocorrelacionados e possuem uma grande variabilidade, sugerindo um modelo. Aplicou-se primeiramente o modelo auto-regressivo 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 foi 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 neste caso não era preciso escolher variáveis, como é o caso de parte dos estudos mostrados acima.

Este trabalho propôs diminuir o números 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, e assim 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 e não um ARFIMA, como o sugerido em Spagnol[22]. E por fim utilizou-se gráficos de controle, CUSUM e MMEP, para monitorar a série.

Nas seções seguintes são detalhados os procedimentos e métodos usados para concretizar o estudo.

3 Bases Teóricas

3.1 Análise de Correlação Linear

Definição 3.1 *O coeficiente de correlação mede a relação linear existente entre as variáveis, ou seja, o quão fortemente elas estão relacionadas linearmente. E é expressado como:*

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

onde X e Y são variáveis aleatórias com variância σ_x^2 e σ_y^2.

Propriedades:

- *$\rho(X,Y)$ é sempre entre -1 e 1;*
- *Quanto maior a correlação entre as variáveis, mais próximo de 1 ou de -1 o coeficiente estará.*
- *Quando $\rho(X,Y)$ estiver perto de zero, não existe correlação linear.*

Contextualizando com o 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 uma maneira foi estudar a correlação linear entre elas.

Em Spagnol [23] 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 seria necessário um

estudo mais detalhado para certificar quais variáveis podem ser retiradas sem perda de efetividade no monitoramento. Este cuidado foi tomado neste trabalho.

3.2 Modelos de Séries Temporais

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 probabilistico adequado para a série, analisando a existência de periodicidade, entre outros.

Uma questão importante para a construção do modelo é selecionar o mais simples, ou seja, buscando um número mínimo possível de parâmetros.

Para a escolha do modelo é necessário saber se a série é estacionária, e uma série é estacionária quando ela expressa certas características constantes no tempo. Uma definição mais formal é que a lei de probabilidade do processo não sofre alterações com o tempo.

Definição 3.2 *Um processo estocástico Z(t) é estritamente estacionário se a distribuição conjunta* $Z(t_1)$, $Z(t_2)$,...,$Z(t_n)$ *é a mesma de* $Z(t_{1-k})$, $Z(t_{2-k})$,...,$Z(t_{n-k})$, *para todo n, para todos os tempos* (t_1, t_2, ..., t_n) *e para todos os ''lags' (posições) k, Silvester[]. O processo é dito fracamente estacionário, ou simplemente estacionário, se o primeiro e o segundo momentos se mantém contantes no tempo, ou seja, se*

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

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 alguma 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.

A classe de modelos autoregressivos-integrados-médias-móveis(ARIMA) é capaz de descrever uma certa classe de séries estacionárias ou derivadas de uma estacionária. Os modelos puramente autoregressivo ou de média móvel não possuem esta característica.

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 mas 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, por exemplo modelo de regressão linear. Porém, para as características que os dados do fluxo apresentaram, o modelo auto-regressivo integrado e de média móvel foi o mais indicado. Outro ponto, apesar do modelo escolhido ser um ARIMA, também foi estudado o modelo ARFIMA, já que algumas referências indicavam este modelo como uma opção para ajustar dados de redes de computadores devido à alta variabilidade e à longa dependência.

A análise inicial para a escolha do modelo observou-se o comportamento da média e verificou-se que havia necessidade de transformação nos dados, ou seja, as observações da RC necessitaram de diferenciação.Por fim o modelo ARIMA foi uma parte essencial para a análise da série do estudo.

3.2.1 Processos de Média Móvel

Definição 3.3 *O processo de Média Móvel é representado se apenas um número finito pesos ψ são diferentes de zero, isto é , ψ_1 = -θ_1, ψ_2 = -θ_2, ..., ψ_q = -θ_q, e ψ_k = 0, quando k>q. Neste caso, dizemos que o processo MA(q) possui uma memória de q períodos, denotado como MA(q). O processo MA(q) é definido como :*

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

ou

$$Z_t = \theta(B)a_t = (1 - \theta_1 B - ... - \theta_q B^q)a_t \tag{3.2.2}$$

onde Z_t é um processo média-móvel de ordem q e a_t tem distribuição N(0, σ^2) e θ_1 ...θ_q são constantes.

Para 1+ θ_1^2+ ... + θ_q^2, um processo MA é sempre estacionário,Wei[]. Já para a condição de invertibilidade, é necessário que as raízes da equação característica $\theta(B)$ = 0. Uma explicação mais detalhada sobre a estacionáriedade e invertibilidade pode ser encontrada em Cowpertwait[] ou Wei[].

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, pórem não foi o mais adequado dentre os outros. Analisando a média constatou-se que ela não era estacionária no tempo.

3.2.2 Processos Auto-regressivos

Definição 3.4 *Um modelo auto-regressivo é dito de ordem p, AR(p), se somente um número finito de pesos são diferentes de zero, ou seja,* $\pi_1 = \phi_1$, $\pi_2 = \phi_2$, ..., $\pi_p = \phi_p$ *e* $\pi_k = 0$ *para k>p. E AR(p) é dado por:*

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Zt - 2 + \phi_p Zt - p + a_t \tag{3.2.3}$$

ou

$$\phi_p(B) Z_t = a_t \tag{3.2.4}$$

Onde Z_t *corresponde à observação da série temporal no tempo t,* ϕ_t *corresponde ao parâmetro do modelo AR de ordem p,* ϕ_t *(B) = (1-* ϕ_1 *B - ... -* $\phi_p B^p$*) e* a_t *ao ruído branco.*

Para $\sum_{j=1}^{\infty} |\pi_j| = \sum_{j=1}^{\infty} \phi_j < \infty$ *indica que o processo será sempre invertível. Para a condição de estacionáriedade as raízes do processo AR(p) devem cair fora do círculo unitário, ou seja, as raízes da função característica (*$\phi_p(B)$*) deve sempre exceder o valor absoluto para ser estacionária. Tem-se em Cowpertwait[] exemplos de casos onde se tem e não se tem estacionariedade e Wei[] em detalhes a explicação de círculo unitário.*

Neste trabalho ajustou-se um auto-regressivo 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-Jenking - ARMA

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

Definição 3.5 *Seja um processo estocástico estacionário e este apresenta características não só auto-regressiva ou puramente média móvel e sim características de ambos, como o proposto por Box & Jenkins[]*

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

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi p Z_{t-p} - a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \tag{3.2.5}$$

ou

$$\phi(B) Z_t = \theta(B) a_t \tag{3.2.6}$$

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 tem fatores comuns, Brockwell[].

Para um ARMA(p,q), as condições de estacionariedade são somente estabelecidas pela parte do processo auto-regressivo, já que o processo média móvel não predispõem nenhuma exigência para a estacionariedade. Portanto o processo será estacionário se $\phi(B) = 0$ estiver com todas as raízes fora do círculo unitário. No caso do processo ser inversível necessita-se que as raízes de $\theta(B) = 0$ fiquem fora do círculo unitário, Wei[].

Em resumo o processo é estacionário se $\phi(a) =$ (1 - $\phi_1 a$ - ... -$\phi_p a^p$) $\neq 0$ para todo $|a| \leq 1$ e o processo será invertível se $\theta(a) =$ (1 - $\theta_1 a$ - ... -$\phi_q a^q$) $\neq 0$ para todo $|a| \leq 1$

3.2.4 Modelos de Box-Jenking - 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 'I' é justamente um parâmetro que indica o número de diferenças que a série sofreu para se tornar estacionária.

Definição 3.6 *Se d é inteiro e não negativo então Z_t é um processo ARIMA(p, d, q).*

Um processo ARIMA(p, d, q) pode ser expresso por:

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

Se W_t for estacionária, pode-se representar W_t por um processo ARMA(p, q), e se tem:

$$W_t = (1 - B)^d Z_t = \Delta^d Z_t \tag{3.2.8}$$

e

$$\phi(B)W_t = \theta(B)a_t \tag{3.2.9}$$

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

Supõe-se que a d-ésima diferença de Z_t pode ser descrita por um modelo ARMA estacionário e inversível.

No caso de processos não-estacionários homogêneos o usual é tomar de uma a duas diferenças para que a série se torne estacionária, Morettin[]:

- 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.

Contextualizando com o estudo foi utilizado os procedimentos descritos anteriormente. Em especial, verificou-se a necessidade de diferenciação da série. Ajustou-se um modelo misto auto-regressivo e média móvel (ARMA) e um ARIMA, constatou-se que a série necessita de uma diferenciação, sendo então o ARIMA é mais adequado que um ARMA.

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". E onde o parâmetro d é um inteiro, que estabelece o nível de diferenciações necessárias para tornar uma série temporal estacionária. Estudos mais recentes Trevisan[24] propõem uma generalização desta modelagem em relação ao parâmetro d, podendo este assumir não só valores inteiros, mas também representar graus de diferenciação fracionários. Modelos com esta propriedade permitem estudar séries caracterizadas por longas dependências temporais, processos de "memória longa".

Um processo de memória longa é um processo estacionário em que a função de autocorrelação 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). Estes modelos intitulam-se processo auto-regressivo fracionário integrado de média móvel, ARFIMA (p,d,q) ou FARIMA, onde F significa "fractional".

Definição 3.7 *Um processo ARFIMA (p,d,q) com -1/2<d <1/2, se Z_t for estacionário e satisfazer a equação*

$$\phi(B)(1-B)^d Z_t = \theta(B)a_t \tag{3.2.10}$$

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

Se $d = 0$, Z_t é um modelo auto-regressivo 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 $|k|$ $\rightarrow \infty$. As autocorrelações originadas de um modelo ARMA(p, q) têm um decaimento exponencial $p(k) \sim a_t$, 0 < a < 1 (Box & Jenkins [3]). No caso do ARFIMA, um processo de "longa dependência", se $0 < d < 0.5$. Um estudo mais detalhado pode ser encontrado em Cowpertwait[].

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

Ás vezes mais de um modelo se ajusta bem aos dados, a escolha pode ser fácil ou até mesmo difícil, nestes casos 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 critério de AIC. O melhor modelo é aquele que tiver o menor valor de AIC calculado. Esse critério também se aplica ao BIC.

Definição 3.8 *Suposto que um modelo estatístico de k parâmetros é ajustados aos dados. Para avaliar a qualidade do ajuste é utilizado o critério de AIC, Wie[].*

O AIC é dado por:

$$AIC = -2ln\hat{L} + 2k \tag{3.2.11}$$

onde $\hat{L}$ é o estimador de máxima verossimilhança e k o número de parâmetros do modelo e o número de observações igual a n, para o processo ARMA. E a função de máxima verossimilhança é dada por:

$$lnL = -\frac{n}{2}ln2\pi\sigma_a^2 - \frac{1}{2\sigma_a^2}S(\phi,\mu,\theta). \tag{3.2.12}$$

onde a soma dos quadrados:

$$S(\phi,\mu,\theta) = \sum_{t=-\infty}^{n}[E(a_t|\phi,\mu,\theta,Z)]^2 \tag{3.2.13}$$

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* $\sigma_a^2 = \frac{S(\phi,\mu,\theta)}{n}$,

$$ln(\hat{L}) = -\frac{n}{2}ln\sigma_a^2 - \frac{n}{2}(1 + ln2\pi) \tag{3.2.14}$$

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

$$AIC(k) = nln\sigma_a^2 + 2k \tag{3.2.15}$$

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 BIC.

Através de estudos afirmou-se que o BIC é menos provável de superestimar a ordem da parte autoregressiva, Wei[].

Definição 3.9 *O BIC é defindo como:*

$$BIC(k) = -2ln\sigma_a^2 + kln(n). \tag{3.2.16}$$

Contextualizando com o estudo de fluxo de rede os conceitos acima apresentados foram essenciais para ajustar modelos de séries temporais. Para encontrar um modelo adequado aplicou-se a diversidade de modelos acima (AR(p), MA(q), ARMA(p,q), e ARIMA(p,d,q) e ARFIRMA(p,d,q)) e também se necessário variou-se o número de parâmetros, sendo que números de parâmetros diferentes indicam modelos diferentes. 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 do fluxo da RC em questão, buscou-se um método que pudesse monitorá-las. Últimamente utilizam-se os gráficos de Controle Estatístico do Processo(CEP) para monitorar processos diversos. Nesta trabalho, buscou-se encontrar o gráfico de CEP que melhor suporta os dados de RC.

Em CEP, um dos gráficos mais conhecidos e utilizados é o gráfico de controle de Shewhart ($\bar{X}$ e R), tendo como vantagem a sua fácil aplicação. Porém, uma grande desvantagem destes gráficos é que eles utilizam apenas as informações contidas no último ponto plotado, descartam qualquer informação fornecida pela seqüência inteira de pontos. Devido a esta característica, esse tipo de gráfico é insensível a pequenas mudanças no processo (da ordem de 1,5 σ ou menos) Montgomery[15]. E o mais importante para os dados em questão é que as observações devem ser independentes e normalmente distribuídas, Moreira[16]. Estudando os dados da rede nota-se que os dados possuem autocorrelação, ou seja, eles não são independentes e a informação contida nos pontos anteriores é importante para entender o comportamento do fluxo dos dados.

O Gráfico de Controle da Soma Cumulativa (CUSUM) e o Gráfico de Controle de Média Móvel Exponencialmente Ponderada(MMEP ou 'EWMA') também são gráficos 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 distribuídas, admitindo observações correlacionadas.

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

Contextualizando como estudo do tráfego de rede devido as características do fluxo da rede analisada optou-se por monitorar os dados da RC os gráficos de MMEP e CUSUM, já que como descrito acima são mais sensíveis a pequenas variações e principalmente não exigem normalidade.

3.3.1 Gráficos de Controle da Soma Cumulativa - CUSUM

Na Soma Cumulativa, 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 [8],(n= 1 signfica que

houve apenas um único valor observado várias vezes). No estudo em questão, foi considerado apenas uma amostra com 4700 observações.

Para melhor entendimento deste trabalho, detalharam-se alguns conceitos de CUSUM baseados em Montgmomery [15].

Definição 3.10 *O gráfico de CUSUM incorpora diretamente toda informação na sequência dos valores da amostra em relação a um valor-alvo. Supondo que tem-se amostras de tamanho n≥ 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 Cumulativa é formado, plotando-se a quantidade versus a amostra i. C_i é a soma cumulativa até, e incluindo, 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 definida abaixo 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 para cima ou positiva que se desenvolverá na soma cumulativa C_i e inversamente se a média se desloca para baixo para um valor $\mu_1 < \mu_0$ então uma tendência para baixo ou 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 dos 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 3.11 *O CUSUM tabular acumula desvios de μ_0 que estão acima do alvo, com outra estatística C^+, e acumula desvios de μ_0 que estão abaixo do alvo, com outra 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:

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

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. Geralmente assume-se que um valor razoável para H é 5 vezes o desvio padrão do processo.

3.3.2 Gráficos de Controle da Média Móvel Exponencialmente Ponderada - MMEP

Para melhor entendimento deste trabalho, detalharam-se alguns conceitos MMEP foram baseados em Montgmomery [15]. O desempenho do MMEP é parecido com o CUSUM, porém tem a vatagem de ser mais fácil de operar e de estabelecer parâmetros. Para o monitoramento do fluxo de rede possui alguns pontos muito relevantes:

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

Definição 3.12 *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 $z_0 = \mu_0$.

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, pode-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 bem pequeno, o ideal é reduzir a largura dos limites, adotando L entre 2.6 e 2.8, segundo Montgmomery[15].

4 Materiais e Métodos

4.1 Materiais

Angelis[1] contruiu um processo/programa no 'C++ Builder 5' como tratativa inicial da base de dados. O Sistema de Gerenciamento de Banco de Dados (SGBD) utilizado foi o 'SQL server 2000', através deste processo foi possível ler a base de dados'.

Após os dados terem passados pelo estágio de ''leitura' anteriormente especificado, foram tabulados no programa Microsoft Office Excell 2007, a fim ficarem com o formato adequado para se utilizar os modelos de séries temporais.

Para a análise estatística foram utilizados os seguintes programas:

- Minitab 15
- JUMP 7.0
- R2.9.2 project

E processados em um Processador Intel Pentium Dual-Core T2390 com 2 GB de memória e sistema operacional Microsoft Windows Vista

4.2 Métodos

4.2.1 Amostragem e dados

- **Dados - Tráfego de Rede Ethernet**

  A amostragem dos dados esteve fora do escopo deste estudo. O conjunto de dados estudado foi o mesmo da tese de Angelis[1], de forma a estabelecer comparações significativas entre os trabalhos. Os dados já passaram por tratamentos adequados e agrupados em fluxos.

  Os dados observados, Angelis[1], referem-se ao fluxo de uma rede local do Instituto de Física da Universidade Federal de São Carlos (IFSC), que foi observada no período de novembro de 2000 à outubro de 2001. Obteve cerda de 10 milhões de registros, pelo software NeTraMet, que atua como um 'sniffer', lendo todo o tráfego da rede.

  As observações foram classificadas em fluxos (conjunto de pacotes trocados entre dois computadores) para facilitar a análise. Os dados brutos das extrações não estavam em um formato adequado, devido a este fato, eles foram tratados através de aplicativos criados especificamente para esta finalidade, Angelis[1]. No período de coleta dos dados não houve nenhum problema na rede, ou seja, a rede estava sobre controle.

  Nas extrações observou-se variáveis e que descrevem grandezas associadas ao tráfego presente na rede, por exemplo: número de bytes, número de pacotes, fluxos IP, fluxos TCP, fluxos UDP, fluxos não-IP, etc. Foram estudas variáveis são significativas para a construção de um modelo de monitoramento da rede. Para construção inicialmente foi estuda a correlação entre as variáveis, levando em consideração a informação a priori sobre cada variável para realização dos comparativos.

  Os 10 milhões de registros amostrados estão divididos em 3 amostras, denotadas como Amostra 1, Amostra 2 e Amostra 3, maiores detalhes sobre o conteúdo de cada amostra está disponível

em Angelis[1]. A base Amostra 3 foi a base analisada neste trabalho, ela contempla 8 dias de coleta começando às 10:00 do dia 15 de outubro de 2001 até às 11:32 de 22 de outubro de 2001. Os dados foram observados e registrados a cada 2 minutos.

A partir desta seção quando o texto citar base de dados do fluxo da rede está se referenciando a base Amostra 3.

Na Figura 1, tem-se o fluxo desde a extração dos dados até a análise, nele é possível observar o fluxo que a informação seguiu.

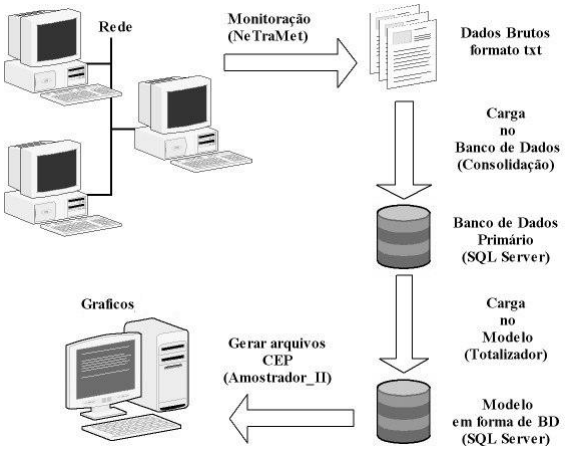

Figura 1: Fluxo desde a coleta dos dados até a análise (Fonte: ANGELIS[1])

A nível de experimentação, em um primeiro momento, foi estuda inicialmente uma série de tráfego de redes de computadores - LAN(Cowperwait [5]) mais simples do que a Amostra 3, porém com características semelhantes. Os dados são números de pacotes (conjunto de informações em bits) medidos em bitis observados em um intervalo de 10-ms em 4000 consecutivas observações de uma rede Ethernet do Instituto de pesquisa "'Bellcore Morristown and Engineering"'. Foi feita uma análise nestes dados apenas com o intuito de ajudar a entender o comportamento do fluxo de uma rede para facilitar na escolha do método estatístico indicado para esse tipo de série de dados. Através do estudo destas observações, Spagnol[22], notou-se que os dados são altamente correlacionados o que indica que um modelo de previsão que representaria bem os dados é um modelo ARIMA. E que, devido aos dados serem autocorrelacionados, os gráficos de CEP 'úsuais', que necessitam de observações independentes e normalmente distribuídas não são os mais indicados, já que autocorrelação compromete a independência entre observações. Para um melhor desempenho do resultado, é indicado a utilização do gráfico de média móvel exponencialmente ponderada (MMEP) e o de soma cumulativa (CUSUM) como alternativa.

Esta primeira análise ajudou no direcionamento dos métodos a serem utilizados neste estudo.

4.2.2 Organização dos Dados para Análise Estatística

Base de dados estava agrupada em fluxo, porém apenas este tratamento não foi suficiente para analisar os dados estatisticamente. Como havia vários computadores sendo monitorados na rede foi necessário somar o fluxo de todos eles. Outro ponto relevante foi que os dados eram acumulativos a cada dois segundos, ou seja, no instantes 6 segundos após o inicio da coleta, tinha-se a informação acumulada do instante 2, 4 e 6. Para aplicar o modelo de séries temporais foi necessário desacumular os dados e então somar os fluxos.

O tratamento para desacumular os dados seguiu a seguinte lógica :

O fluxo inicia-se no instante t_0 e se tem a primeira coleta no instante t_1.

Se o instante t_i sendo i diferente de 0 e 1 tem-se a seguinte equação.

$$\text{QTD no instante } t_i = \text{QTD no instante } t_i \text{ - QTD no instante } t_{i-1}$$

onde QTD é a quantidade de dados.

O mesmo foi feito para a quantidade de bytes. O maior desafio foi o de não misturar as informações de cada conexão, por exemplo, entre o computador A e o computador B não perder sua referência inicial.

Com este tratamento foi possível organizar os dados para serem analisados estatisticamente.

Na Tabela 1 tem-se uma amostra da disposição dos dados

Tabela 1: Amostra da tabela de dados tratada

Instante	Número de pacotes
2001Oct21.01:00:00	85
2001Oct21.01:02:00	69
2001Oct21.01:04:00	69

4.2.3 Análise de Correlação Linear

Para diminuir o número de variáveis do modelo Mrcep, proposto por Angelis[1], realizou-se uma análise de correlação linear que mediu o grau de relação entre elas.

Calculou-se o coeficiente de correlação linear e para ilustrar a correlação linear representou-se os dados em gráficos, sendo que no eixo x imputou-se uma variável e no eixo y a outra variável (Tabela 2).

Tabela 2: Variáveis correlacionadas

Variável (eixo x)	Variável (eixo y)
Número de pacotes fonte destino	Número de bytes fonte destino

4.2.4 Ajustando o Modelo de Séries Temporais

Para representar o fluxo da rede de computadores estudada, foi escolhido a variável fonte-destino em número de pacotes.

Primeiramente analisou-se a série ao longo do tempo através de gráficos de dispersão como *scaterplot* e *boxplot*, em ambos é possível observar como os dados estão distribuídos ao longo do tempo. E verificou-se que algumas informações eram muito discrepantes (*outliers*) em relação ao restante, sendo que alguns pontos estavam longe da média. Como no período de observação dos dados não ocorreu nenhuma interrupção ou mesmo problema na rede optou-se por retirar tais pontos para não distorcer o modelo estatístico, foram retirados 4 pontos no total. Todas as análises a partir deste momento foram feitas sem tais observações.

Analisando os dados graficamente observou-se que a série possui variações ao longo do tempo, porém entendeu-se que este é o comportamento natural dela, já que era esperado que em alguns horários o fluxo de dados fosse mais intenso do que outro.

Para analisar o fluxo em detalhe optou-se por dividir a análise em 4 visões, a fim de encontrar a maneira mais adequada para entender e monitorar a rede.

Os cenários foram estruturados da seguinte maneira:

1. A série foi analisada seguindo a seqüência do registro dos dados.

2. A série foi separada em períodos, das 6:00 as 20:00 maior fluxo e o restante com menor fluxo.

3. Estruturada como o cenário 2, porém calculou-se a média de pacotes dos dias no mesmo instante.

4. Estruturada como o cenário 2, porém calculou-se a mediana do mesmo instante dos dias observados.

Para os 4 cenários realizou-se o seguinte procedimento:

- Analisou-se o comportamento da série ao longo do tempo.

- Verificou-se se a média era constante durante o fluxo, para identificar se a série era estacionária ou não.

- Ajustou-se vários modelos variando o tipo de modelo e o número de parâmetros. Iniciou-se por modelos mais simples, AR(p) e MA(q) ajustando primeiro com menor número de parâmetros para os maiores, observando o AIC e o BIC de cada um. Se com o aumento do número de parametros o AIC e BIC diminuem aumentava-se o número de parâmentos até que a queda fosse expressiva. Depois foi feito o mesmo procedimento para o os modelos mistos ARMA, ARIMA e ARFIMA.

- Utilizou-se os critérios de AIC e BIC e a análise da série para a escolha do modelo final.

Para auxiliar na escolha do modelo considerou-se as seguintes características da ACF e PACF, descritas na Tabela 3,Wei[] .

Tabela 3: 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)

Após estas etapas escolheu o modelo ARIMA para representar o fluxo da rede.

4.2.5 Gráficos de CEP - CUSUM E MMEP

5 Análise Estatística da Rede

5.1 Análise de correlação e escolha do modelo do fluxo de rede - fonte destino (Série 1)

Realizou-se primeiramente um estudo com o AR para entender qual das variáveis era necessária para o monitoramento da rede. As variáveis mais relevantes foram : fonte destino em pacotes, fonte destino em bytes, destino fonte em pacotes e destino fonte em bytes. Na Tabela 4 nota-se a correlação entre as variáveis e que não é necessário monitorar todas, as variáveis fonte-destino pacote e fonte-destino bytes são correlacionadas e o mesmo para destino-fonte pacote com destino-fonte bytes.

Tabela 4: Variáveis correlacionadas

Variáveis	Coeficiente de correlação
Fonte-destino pacotes vs Fonte-destino bytes	0,85
Fonte-destino pacotes vs Destino-fonte pacotes	0,85
Destino-fonte pacotes vs Destino-fonte bytes	0,94

Para este trabalho optou-se por monitorar e estimar um modelo para o fluxo da fonte para o destino em número de pacotes, (Série 1), o mesmo tipo de análise pode ser replicado para as outras variáveis.

Observando a Série 1 notou-se que alguns pontos variaram significativamente em relação à média da série, Figura 2 a-). Na Figura 2 b-) é possível notar esta variação através do boxplot. Este 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 a grande dispersão dos pontos atípicos. Em especial tem-se 4 pontos que são os que mais estão discrepantes em relação ao conjunto de dados, outiliers. Outilires são pontos que não se enquadram no padrão do restante dos dados; eles podem ser erros de medida ou um comportamento diferente do padrão.

.

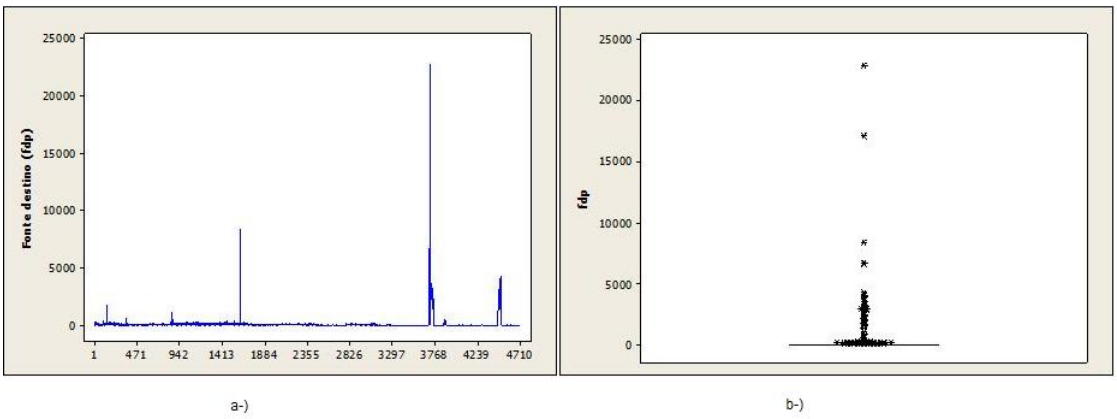

Figura 2: a-)Gráfico da Série ao longo do tempo. b-)Boxplot da Série.

Baseado na Figura 2 b-) e analisando a variação em relação à média, optou-se por retirar 6 pontos, Tabela 5, sendo 1 porque é a primeira observação da coleta que está com um valor muito inferior comparado aos seguintes, provavelmente porque na primeira observação não deve ter sido carregado devidamente o "sniffer- leitor dos dados e os demais 5 devido a grande variação pontual. Estes 5 pontos foram retirados levando em consideração a premissa que em nenhum momento a rede analisada

sofreu queda/interrupção, devido a este fato entendeu-se que esta 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 a partir deste ponto foram sem tais observações.

Tabela 5: 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

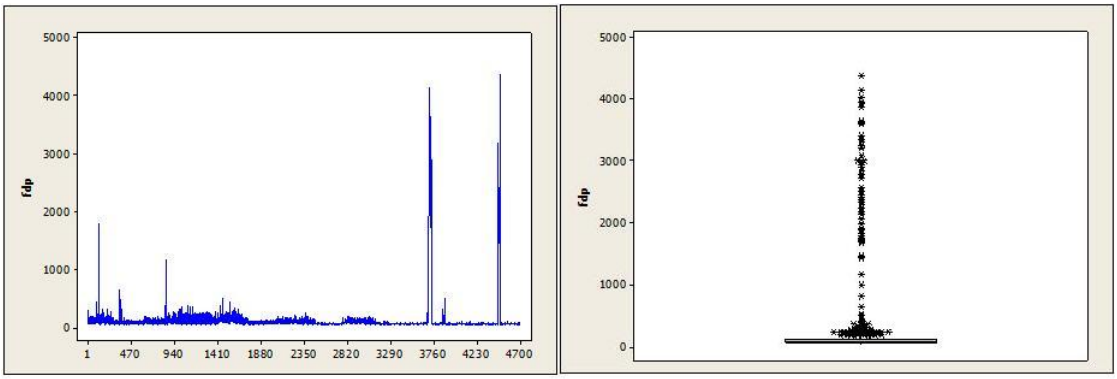

Figura 3: a-)Gráfico da Série 1 ao longo do tempo. b-)Boxplot da Série 1.

Tabela 6: Medidas de localização e dispersão.

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

Analisando novamente a série de dados, Figura 3 a-), é possível notar que ainda existe variação do fluxo de pacotes ao longo do tempo, como por exemplo os 'picos' no final da série, porém entendeu-se que esta variação é característica da rede. Na Figura 3 b-) pode-se notar que o boxplot ainda apresenta pontos atípicos, porém antes o ponto máximo era 22845 e agora é de 4377, Tabela 6.

Na Tabela 6 analisando o desvio padrão, o minímo e o máximo dos valores observados, nota-se que a variabilidade das observações diminuiram com a retirada de 6 pontos.

Para ajustar um modelo mais aderente optou-se por analisar a Série 1 distribuindo os dados em 4 cenários, como citados na seção de métodos.

5.1.1 Cenário 1

Para ajustar um modelo de séries temporais, primeiramente analisou-se a média da Série 1 era constante ao longo do tempo. E pela Figura 4 a-), nota-se que ao longo do tempo existe variações, estas variações indicavam que a série deveria ser diferenciada. Antes de diferenciar a série foi analisada a função de autocorrelação(ACF) e a função de autocorrelação parcial (PACF). Na Figura 4 observa-se o comportamento delas respectivamente, na ACF o decai lentamente e na PACF o número de lags significativos indicaram a ordem do modelo, um ARMA(5,1).

Na Figura 5, tem-se a ACF e PACF diferenciada, e com a Tabela 7 nota-se que o ARIMA(5,1,3) é o modelo que pela ACF e PACF e pelo AIC é o mais indicado.

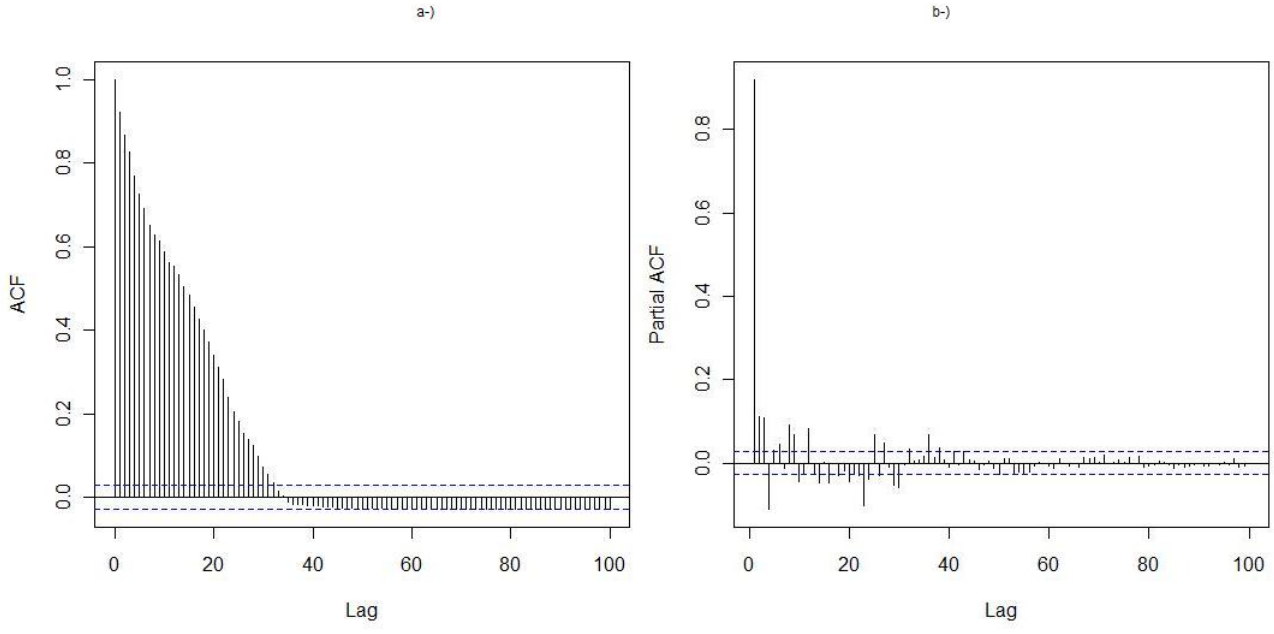

Figura 4: a-)Gráfico da Função de Autocorrelação da Série 1. b-) Gráfico da Função de Autocorrelação Parcial da Série 1.

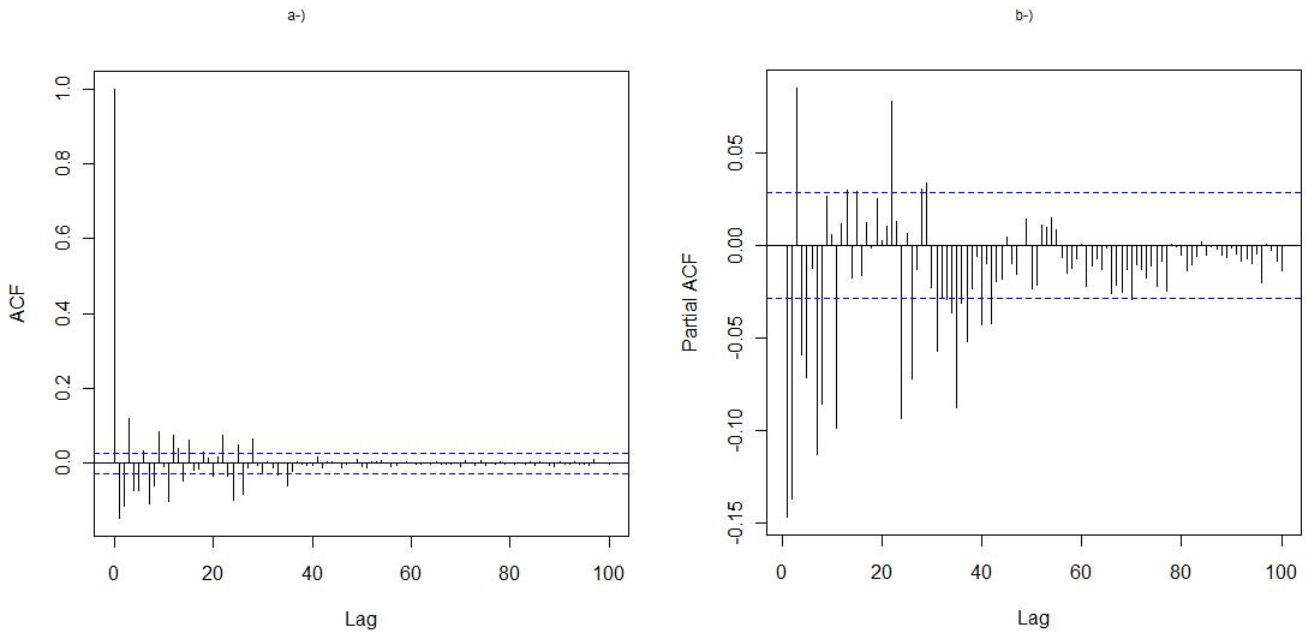

Figura 5: a-)Gráfico da ACF da Série 1 diferenciada. b-) Gráfico da PACF da Série 1 diferenciada.

Tabela 7: 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

5.1.2 Cenário 2

No Cenário 2 a série foi separada em períodos, maior fluxo (das 6:00 ás 20:00) e o restante contendo o menor fluxo. Chegou-se nesta divisão analisando-se como era o fluxo ao decorrer de cada dia para

encontrar qual o período que tinha o maior fluxo. Na Figura 6 a-), tem-se cada dia expresso no instante de cada coleta. Analisando-os nota-se que ao longo do fluxo tem-se horários que possuem maior quantidade de pacotes do que o restante correspondendo ao horário entre 6:00 ás 20:00.

Na Figura 6 b-), observa-se a dispersão do fluxo de pacotes por dia, nota-se que alguns dias tiveram uma variação superior que os restantes, sendo eles: 20 e 21.

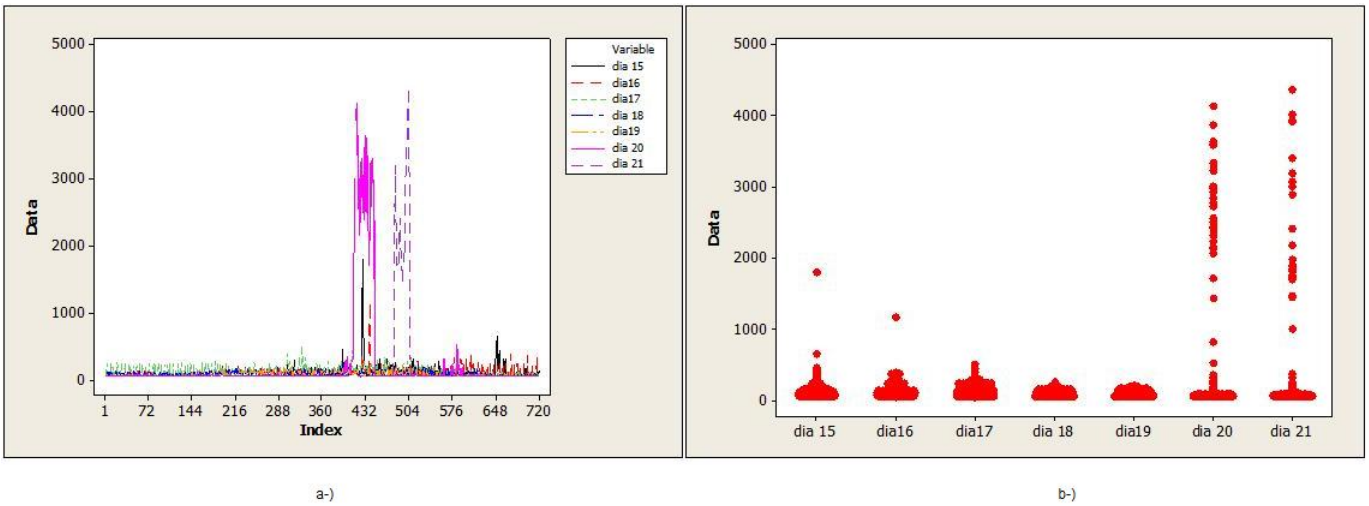

Figura 6: a-)Gráfico dispersão do fluxo por dia(início às 0:00 e termino às 23:58). b-) Dotplot - Fluxo de pacotes por dia.

A análise iniciou-se pelo período com maior fluxo, Série 2. O período de maior fluxo de cada dia foi consolidado em sequência. Por exemplo no dia 15 a última observação do período como maior quantidade de pacotes foi às 20:00 em seqüência agrupou-se as observações do dia 16, das 6:00 às 20:00 e o mesmo para o dias restantes.

Na Figura 7 a-) tem-se o boxplot da Série 2, onde há uma grande dispersão dos dados depois do $q3$. Tal dispersão após q3 se deve pela 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, a média é praticamente constante, porém após este período tem-se dois momentos que possuem uma quantidade superior ao início, tais observações pertencem aos dias 20 e 21, Figura 7.

Observando a ACF e a PACF da Série 2, respectivamente Figura 8, 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)).

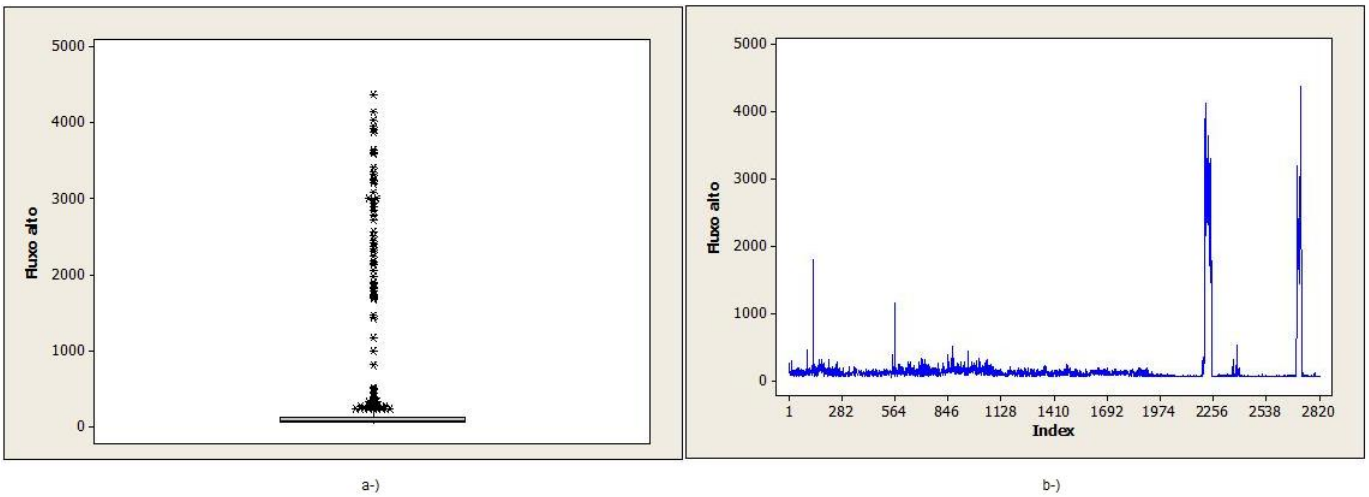

Figura 7: a-)Boxplot da base Série 2. b-)Gráfico da Série 2 ao longo do tempo.

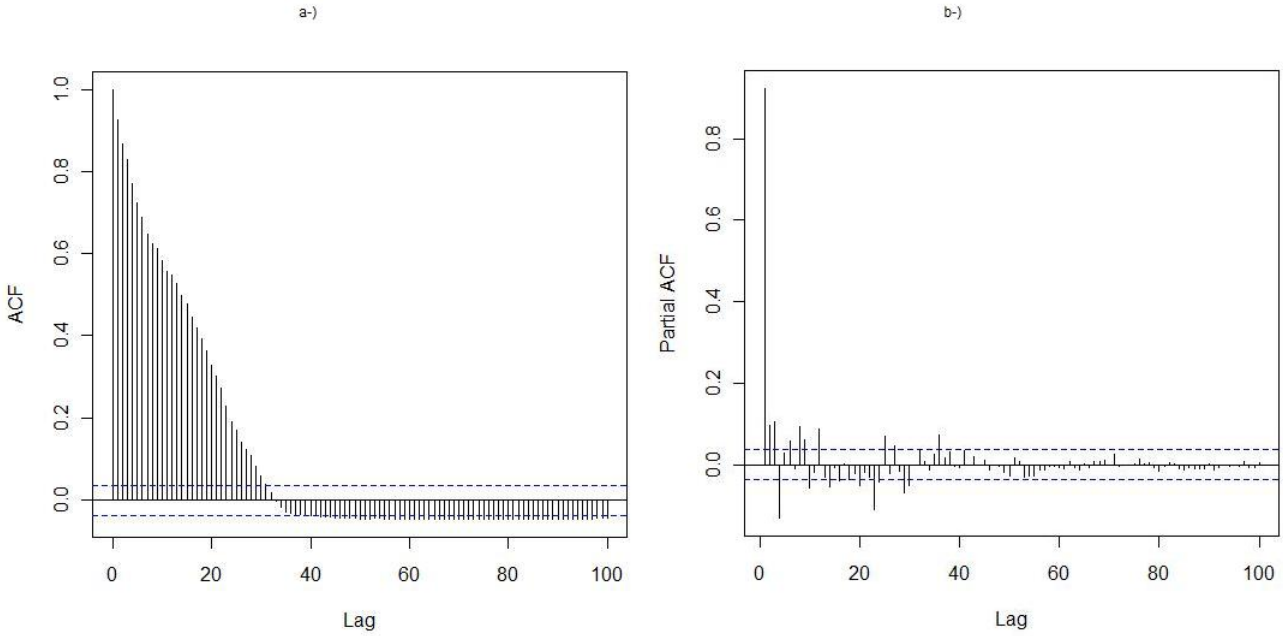

Figura 8: a-)ACF da Série 2. b-)PACF da Série 2

Na Figura 9 tem-se a ACF e PACF diferenciada e na Tabela 8 nota-se que o modelo diferenciado é o que se ajusta melhor aos dados. O ARIMA(5,1,3) é o mais indicado sendo que possui o menor AIC.

Tabela 8: 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**

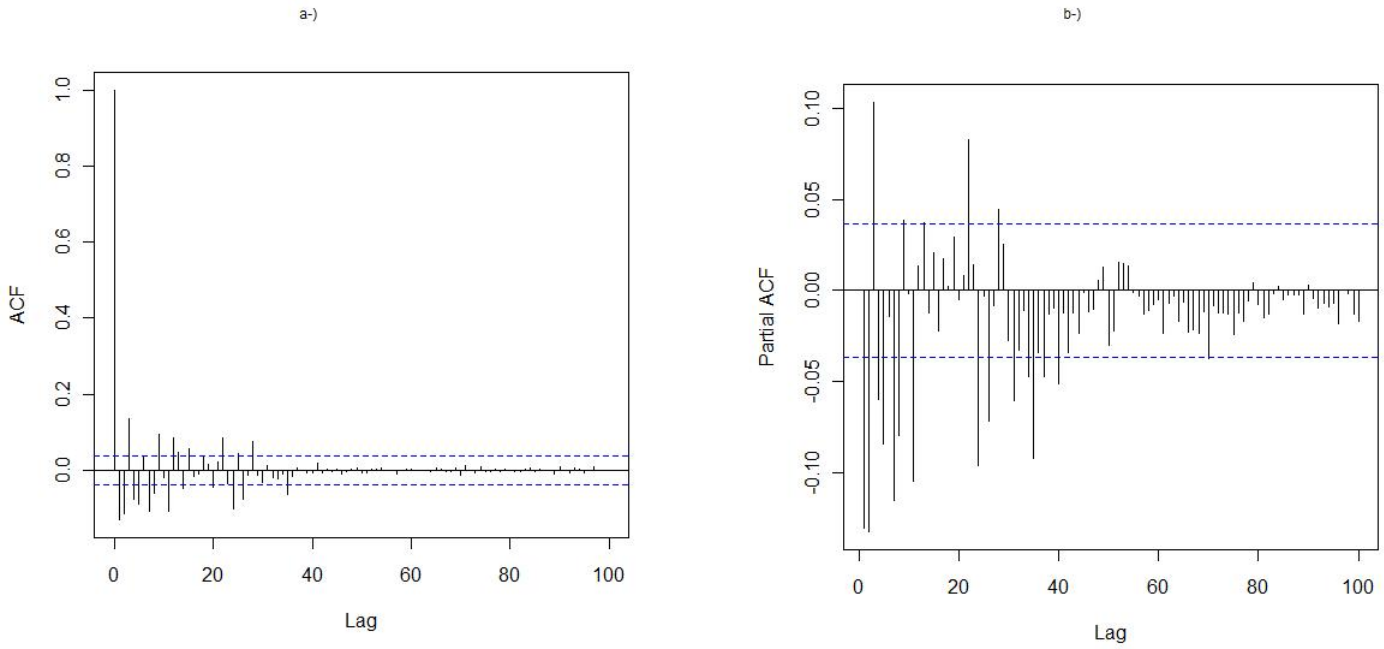

Figura 9: a-)ACF da Série 2 diferenciada. b-)PACF da Série 2 diferenciada

Para o restante do fluxo deu-se o nome de fonte destino de pacotes baixo (Série 3). A mesma análise da Série 2 foi realizada para a Série 3, e analisando a dispersão dos dados, a Série 3 tem um menor desvio padrão, na Figura 10 a-), nota-se que a 'caixa' do boxplot expandiu e os pontos

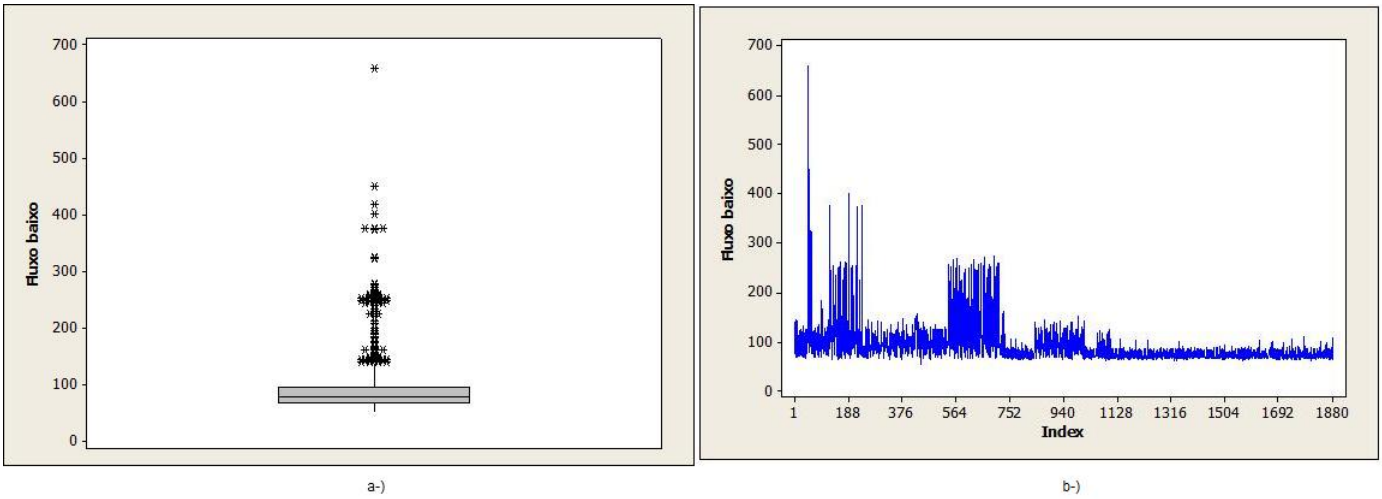

Figura 10: a-)Boxplot da Série 3. b-)Gráfico da série Série 3 ao longo do tempo.

atípicos diminuíram em relação a Série 2, indicando uma maior concentração dos dados. Apesar da variação não ser tão expressiva quanto a Série 2 na Figura 10 b-) é possível notar que a série tem um maior fluxo e variação do começo da amostra até por volta da observação 1130, indicando que a média não é constante em grande parte do fluxo.

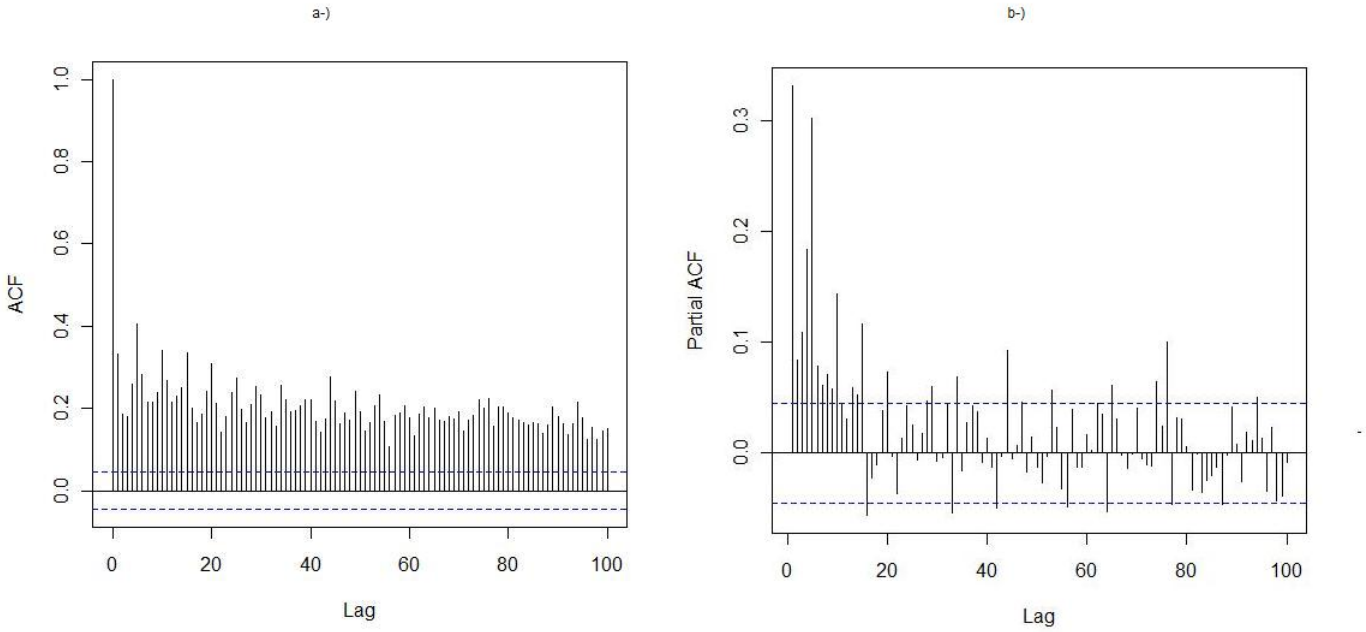

Figura 11: a-)ACF da Série 3. b-) PACF da Série 3.

A ACF da Série 3 não apresentou nenhum decaimento exponencial, mas sim todos os lags até 100º significativos, acima dos limites, Figura 11 a-). Na PACF, 11 b-), não se tem uma decaimento expressivo e há alguns lags significativos. Analisando a série e as funções, optou-se por diferencia-la, sendo que em grande parte do fluxo a média não era constante.

Analisando a ACF da série diferenciada, Figura 12 a-), nota-se que possui um decaimento exponencial depois do primeiro lag. Na PACF, Figura 12 b-), tem-se indicação que a parte autoregressiva do modelo possui ordem de 4 a 6.

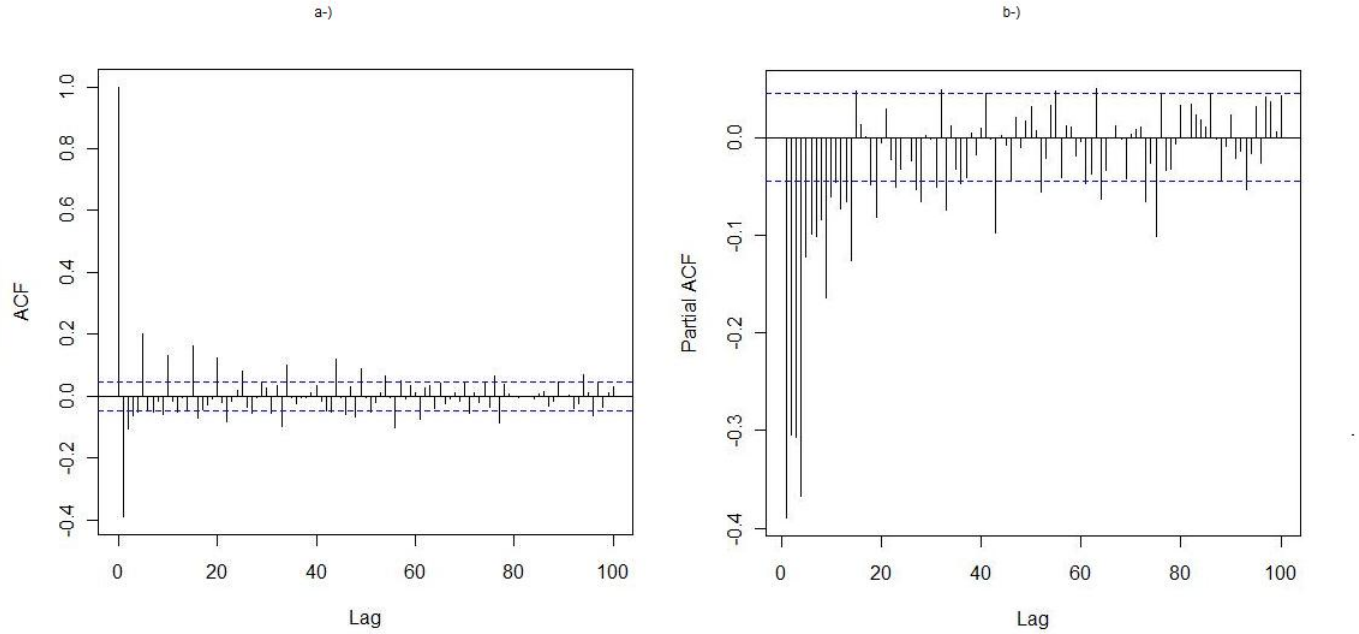

Figura 12: a-)Gráfico da Série 3 diferenciada uma vez. b-)ACF da Série 3 diferenciada

Observando a Tabela 9 o modelo mais indicado é um ARIMA(5,1,1), já que possui o menor AIC e BIC. Realizou-se um teste para identificar se a série possuía longa dependência, porém com o 'd' sugerido pela a análise, não indicou longa dependência, 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 distintos de modelo para cada período do fluxo, um ARIMA(3,1,2) e um ARIMA (5,1,1), isto mostra a necessidade de se dividir os dados, já que cada série necessitou de modelos diferentes.

Tabela 9: Tabela de critério AIC e BIC para modelo estudados em 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

5.1.3 Cenário 3

Para o Cenário 3 utilizou-se a mesma separação de períodos do Cenário 2, porém ao invés 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 13 a-), que apresentou uma maior variação entre $q2$ e $q3$ do que em $q1$ e $q2$. Ainda se tem alguns pontos atípicos, mesmo utilizando a média dos dados a variação ainda é significativa.

Analisando a Figura 13 b-) é nítida a influência da variação dos dias 20 e 21 mesmo utilizando a média dos dias do mesmo instante a variação ainda é significativa. Um ponto importante é que os pontos restantes dos dois picos na Figura 13 b-), o fluxo está com pouca variação.

A ACF da Série 4, Figura 14 a-), não possui decaimento exponencial e a PACF, Figura 14 b-), possui 3 lags significativos indicando um auto-regressivo de ordem 3. Porém como observado na Figura 13a-), a série não tem média constante em alguns momentos, portanto analisou-se modelos com diferenciação também.

A série foi diferenciada uma vez, Figura 15, e as ACF e PACF foram refeitas, Figura 16, respectivamente. E continuaram indicando um auto-regressivo ordem 3, porém com uma ordem de

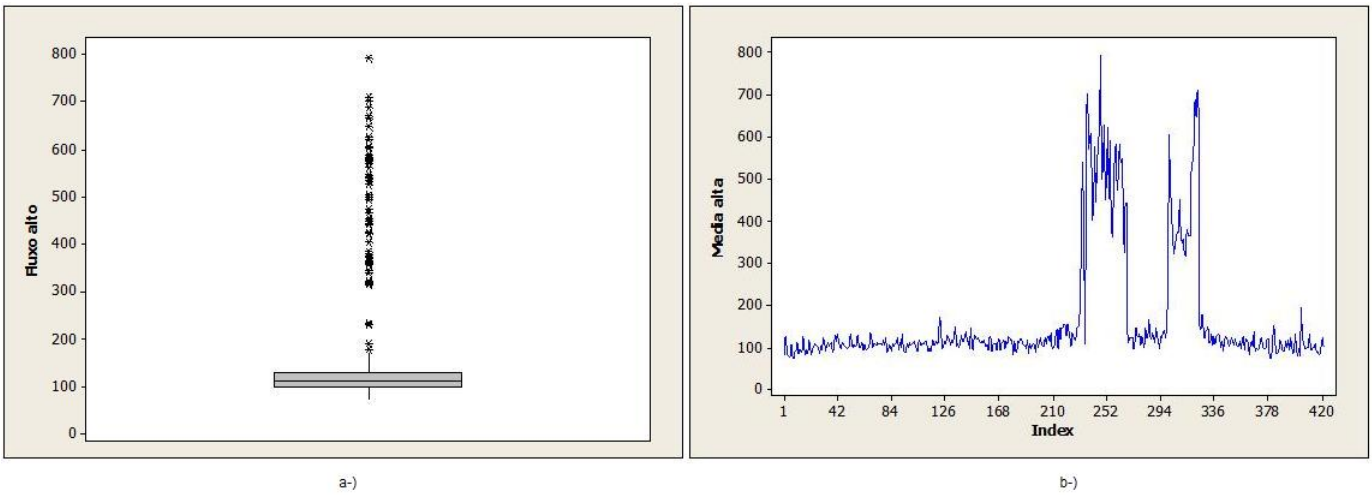

Figura 13: a-)Boxplot da Série 4. b-)Gráfico da Série 4 ao longo do tempo

diferenciação e a ACF com 3 lags significativos.

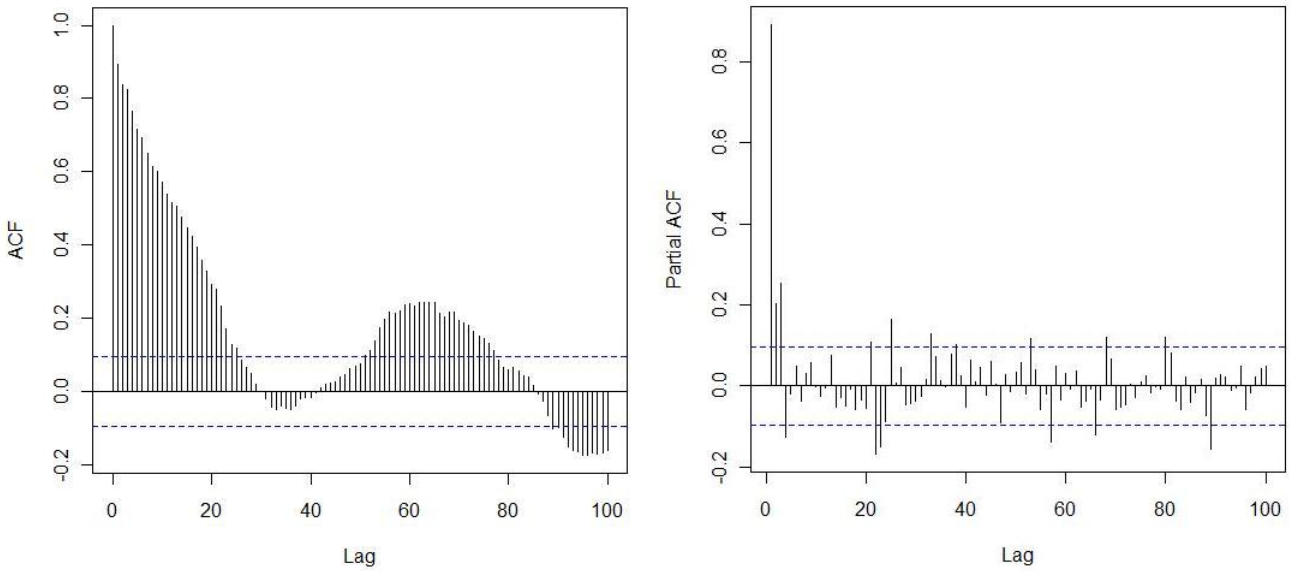

Figura 14: a-)ACF de Série 4. b-) PACF de Série 4.

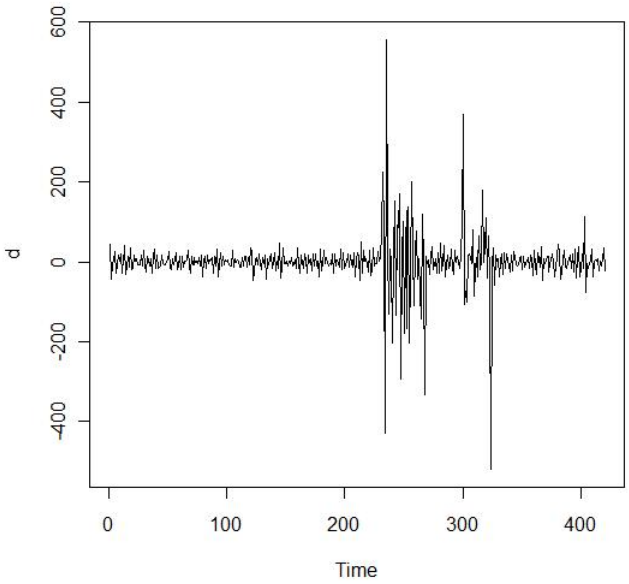

Figura 15: Gráfico dá Série 4 diferenciada.

Para a escolha do modelo realizou-se simulações com ordens diferentes, na Tabela 10 tem-se o *AIC* e *BIC* para cada simulação. Observando a Tabela 10, observa-se que a indicação de um ARIMA(3,1,3) pela ACF e PACF também é indicado pelo AIC sendo o menor.

Tabela 10: Tabela de critério AIC e BIC para modelo estudados em 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 o mesma análise da Série 4.

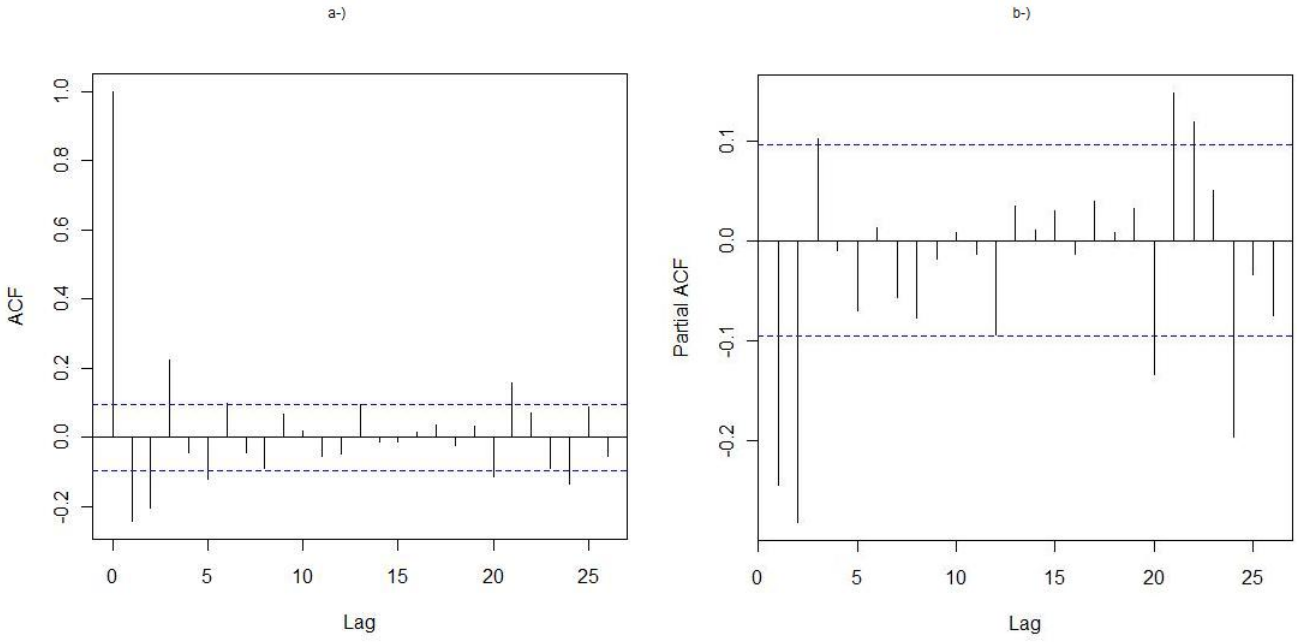

Figura 16: a-)ACF de Série 4 diferenciada. b-) PACF de Série 4 diferenciada

Observando a Série 5, Figura 17 a-), nota-se que está bem mais estável que a Série 3, como esperado. Na Figura 17 b-) nota-se que a quantidade de valores acima de q3 é bem menor em relação as séries já analisadas.

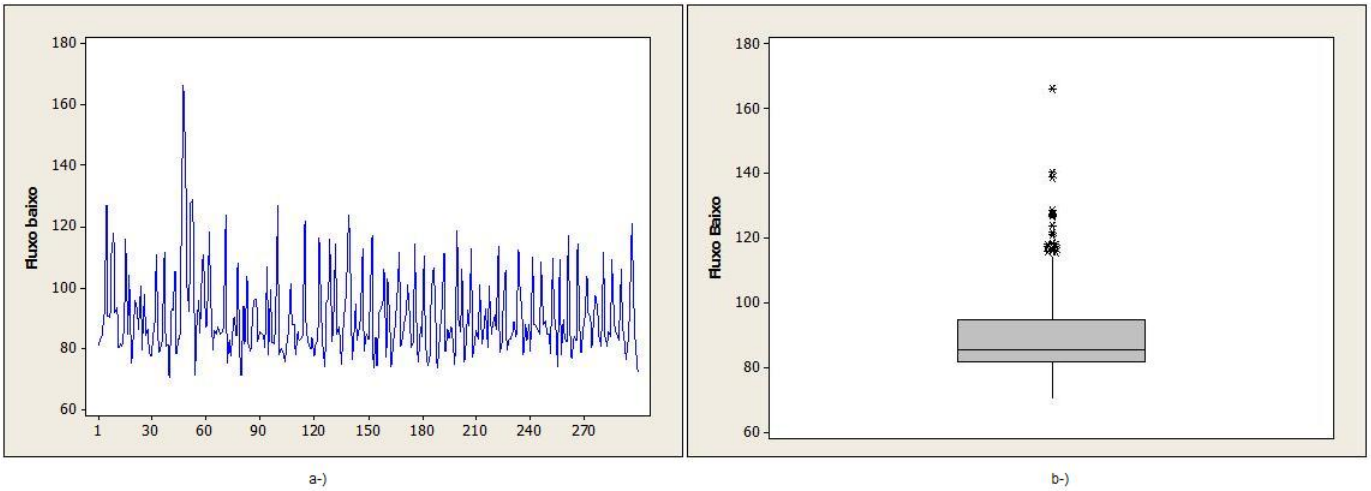

Figura 17: a-)Gráfico da Série 5. b-)Boxplot da Série 5.

Analisando a ACF e PACF, Figura 18, não indica claramente nenhum modelo. Observando a Tabela 11, 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 19 tem-se a série diferenciada, e nela é possível notar que os dados estão ao redor do eixo x =0.

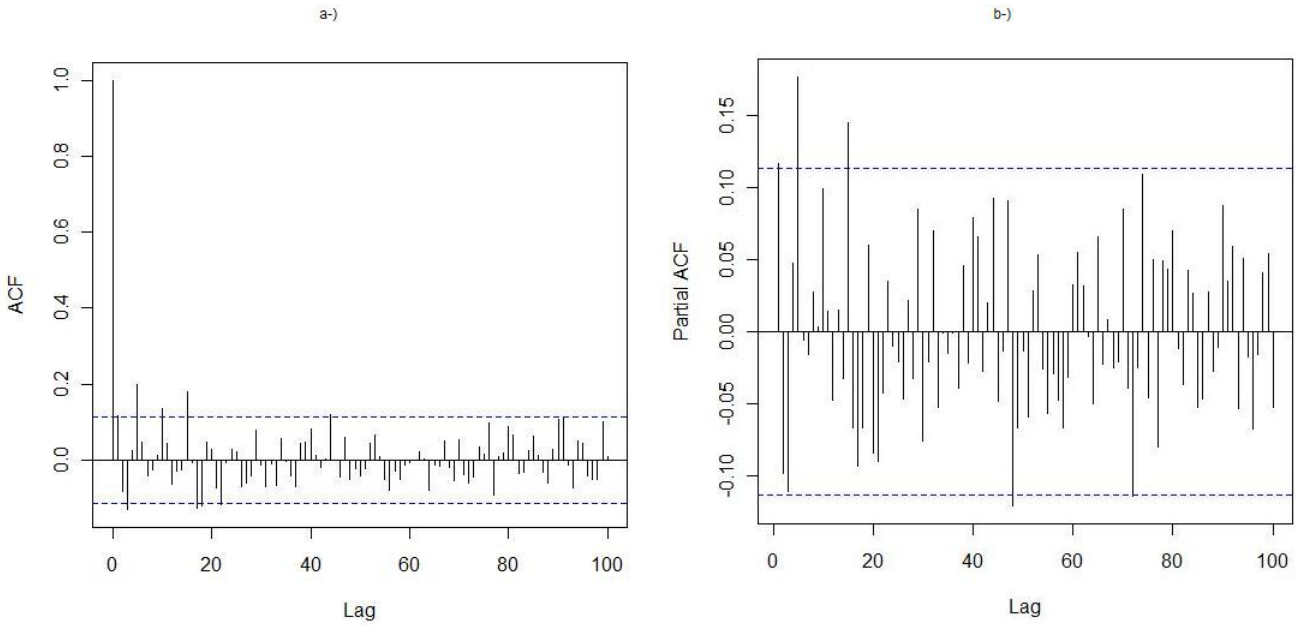

Figura 18: a-)ACF da Série 5. b-)PACF da Série 5.

Observando a ACF e PACF da série diferenciada, Figura 20, confirma-se o modelo proposto pelo AIC. Na ACF tem-se um lag significativo indicando um MA (1) e a PACF tem 5 lags significativos indicando que parte autoregressiva de ordem 5, então por fim tem-se ARIMA (5,1,1).

Para o Cenário 3 então tem-se um ARIMA(3,1,3) para a Série 4 e um ARIMA(5,1,1) para a Série 5.

5.1.4 Cenário 4

No Cenário 4 utilizou-se a mesma divisão do fluxo dos dados Cenário 2, porém ao invés de todos os dados estarem representados unitariamente, a mediana dos dias de cada instante foi usada como

Tabela 11: Tabela de critério AIC e BIC para modelos estudados na 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

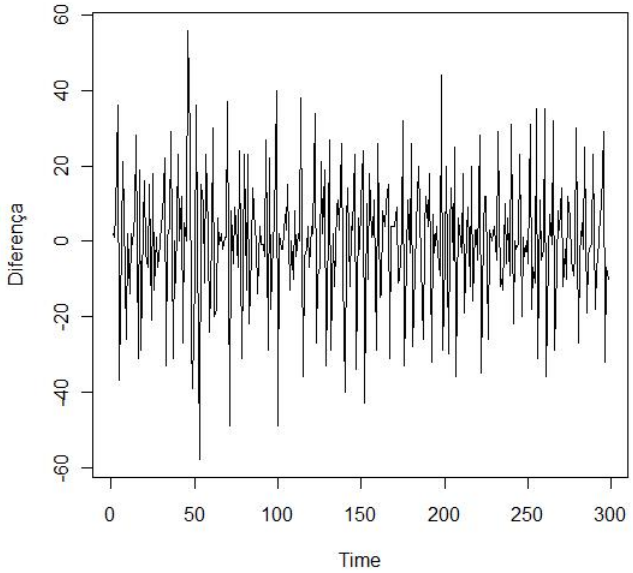

Figura 19: Gráfico da Série 5 diferenciada.

medida representativa. A mediana é uma medida de localização do centro da distribuição dos dados, ou seja, é uma medida de tendência central. 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). Quando se tem valores "muito grandes"ou "muito pequenos", mesmo que estes valores sejam em

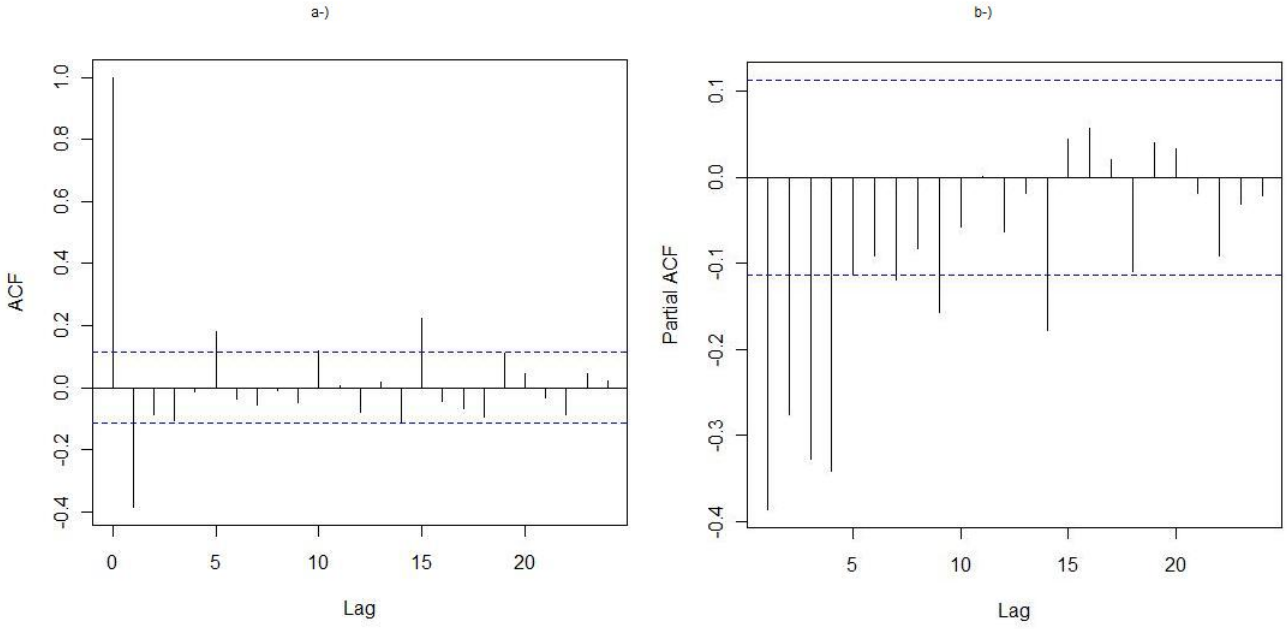

Figura 20: a-)ACF da Série 5 diferenciada. b-)PACF da Série 5 diferenciada.

pequena quantidade na amostra o mais indicado é utilizar a mediana, pois ela é uma medida mais robusta do que a média não se carreganto tanto esta variação que pode prejudicar a construção do modelo e/ou levar a decisões errôneas.

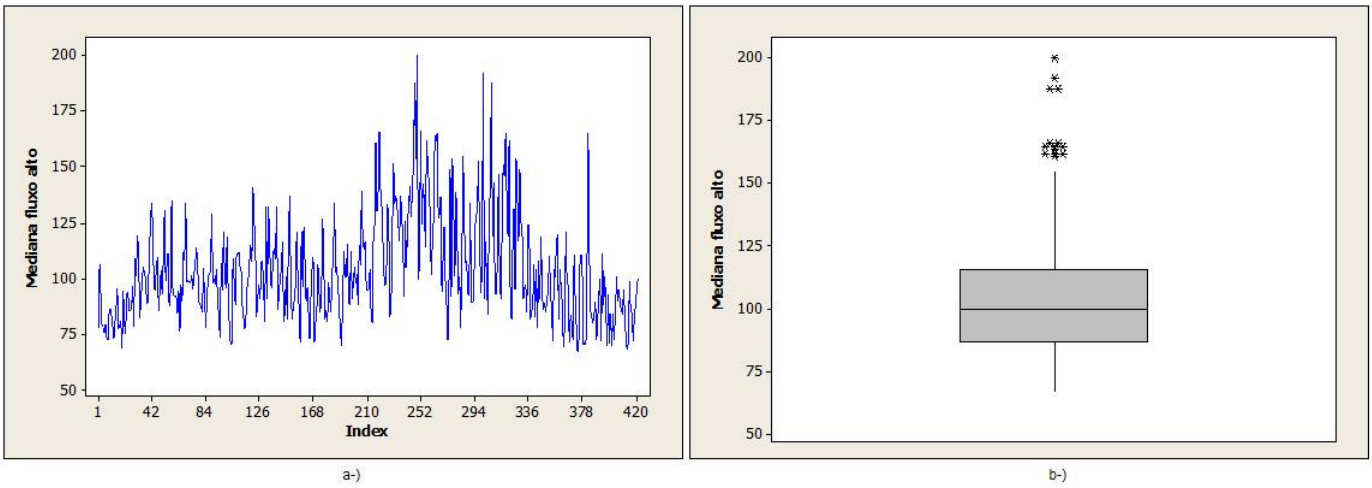

Figura 21: a-)A Série 6 ao longo do tempo. b-)Boxplot da Série 6.

Na Figura 21 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, Figura13 b-). Na Figura 21 b-), a variação do $q3$ até limite superior é menor comparado com a Figura13 a-) com estas duas comparacões é nítido que a mediana não absorve tanto as variações dos dados como a média, levando a um modelo mais robusto.

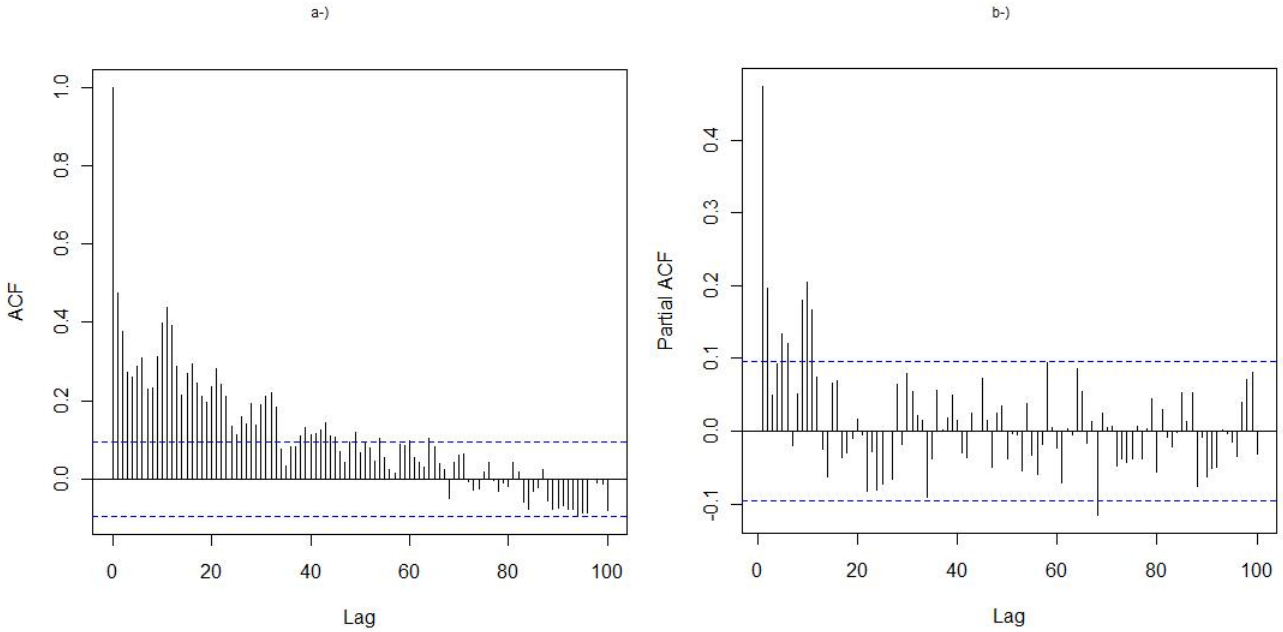

Figura 22: a-)ACF da Série 6. b-)PACF da Série 6.

Na Figura 22 tem-se a ACF e PACF da Série 6 respectivamente, analisando tem-se indícios que a série não é estacionária. Mesmo com a robustez da mediana os dados ainda necessitam de transformação para a série seja estacionária. Para a série se tornar estacionária aplicou-se uma diferença, na Figura 23 tem-se a Série 6 diferenciada.

Figura 24, a ACF indica um lag significativo indicando um MA de ordem 1, já a PACF indica um AR de ordem 3 ou 4.

Na Tabela 12 tem-se o AIC e BIC dos modelos ajustados, nota-se que o modelos ARIMA(4,1,1) e ARIMA (3,1,1) possuem mesmo AIC, porém o BIC do ARIMA(3,1,1) é menor, portanto optou-se por ele.

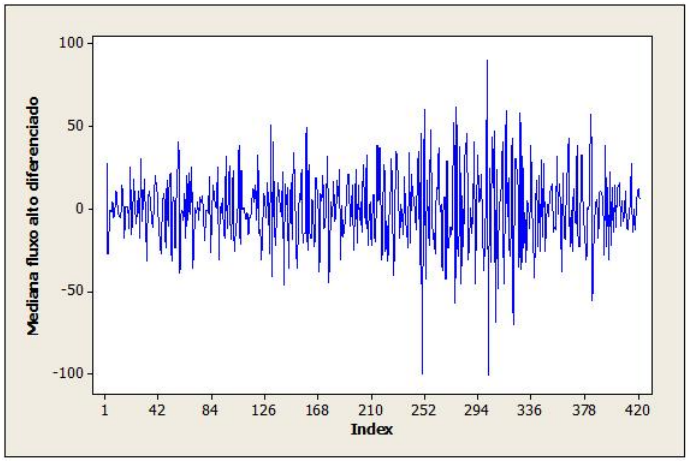

Figura 23: a-) Série 6 diferenciada.

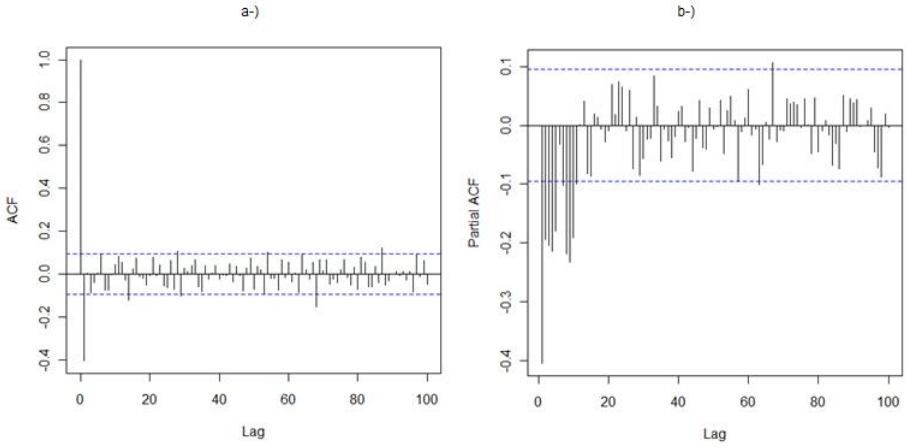

Figura 24: a-)ACF da Série 6 diferenciada b-)PACF da Série 6 diferenciada.

Tabela 12: Tabela de critério AIC e BIC para modelos estudados na 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

Para os dados com a quantidade de fluxo de pacotes menor(20:02 as 5:58), Série 7, realizou-se o mesmo estudo do fluxo maior. Na Figura 25 a-) nota-se que a série em sua maioria tem pouca variação ao longo do fluxo. Na Figura 25 b-) tem apenas um ponto discrepante, na Série 5 que utiza a média das observações tem -se tem mais pontos fora do $q3$ que usando a mediana.

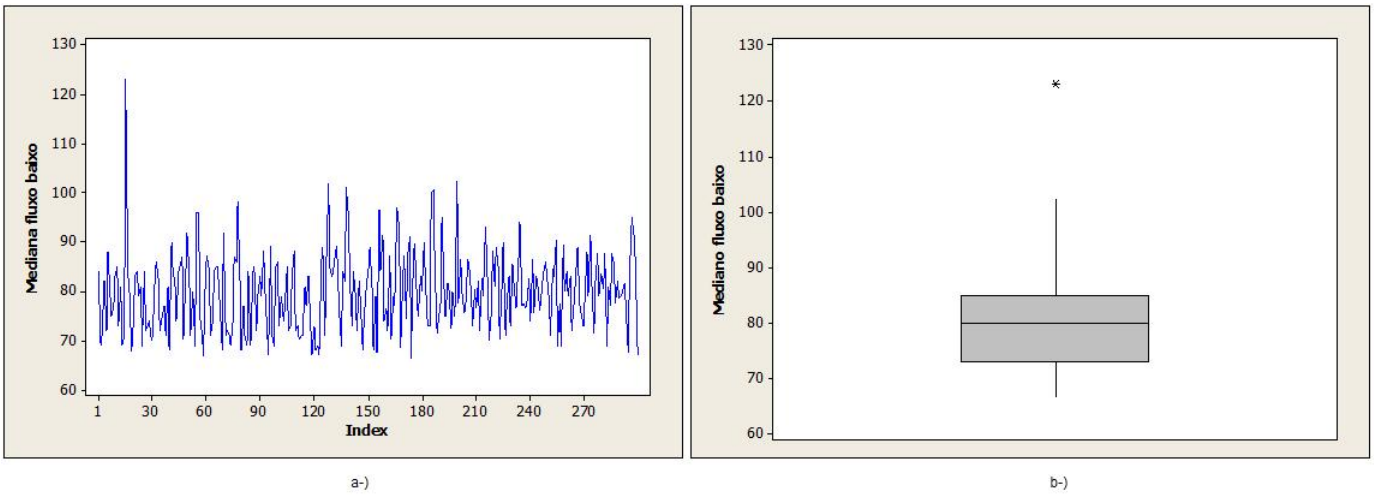

Figura 25: a-)Série 7 ao longo do tempo b-) Boxplot da Série 7.

Para ajustar um modelo observou-se a ACF e PACF da Série 7, Figura 26, indicando que a série necessita de uma diferença. Na Figura 27, tem-se a Série 7 com uma diferença. Na seqüência observou-se a ACF e PACF da série diferenciada , Figura 28, indicando um ARIMA(4,1,1) sendo que o na ACF tem-se um lag significativo e na PACF indica ordem 4 na parte autoregressiva.

Para confirmar o modelo mais adequado calculo-se o AIC e BIC de alguns modelos propostos. Na Tabela 13, nota-se os modelos que possuem maior ordem tem menor AIC, porém conciliando as análises da ACF e PACF , indica um modelo de ordem menor, um ARIMA(4,1,1). Mas o ARIMA (3,1,1) não está fora de ser um modelo proposto já que a diferença dos AIC e BIC é baixa, como já dito anteriormente quanto menos parâmetros melhor.

Para facilitar a estimativa do modelo escolheu-se o ARIMA(3,1,1) para a Série 7 para facilitar a estimativa dos dados sendo que o modelo da Série 6 também é um ARIMA(3,1,1), tendo apenas os pesos diferentes ($\phi e \theta$).

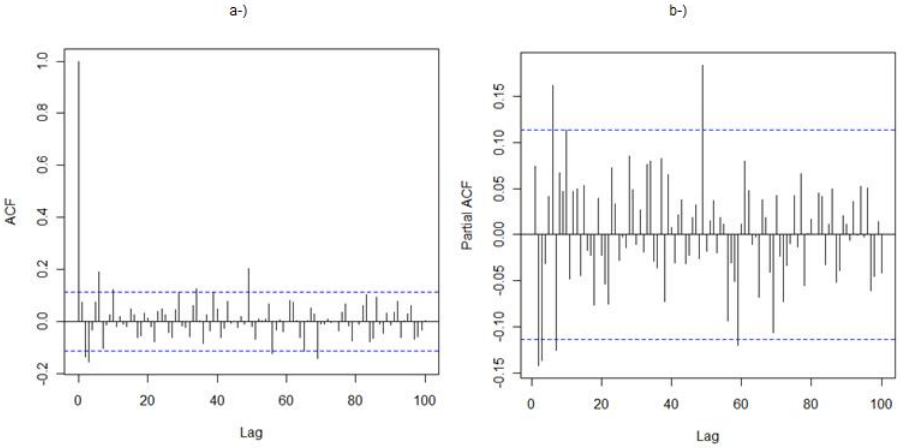

Figura 26: a-)ACF Série 7 b-) PACF Série 7.

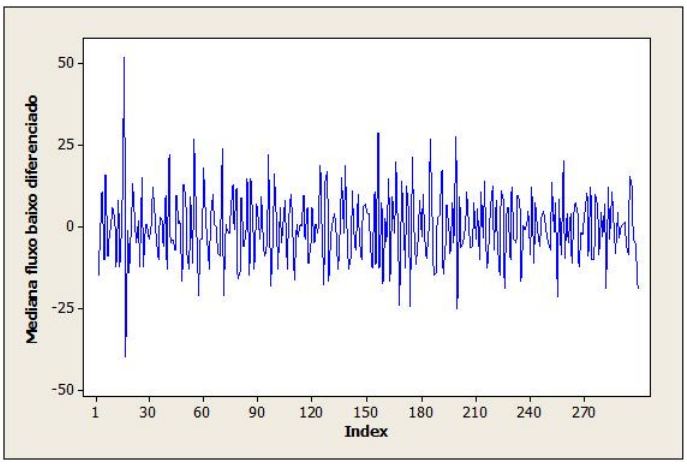

Figura 27: a-) Série 7 diferenciada

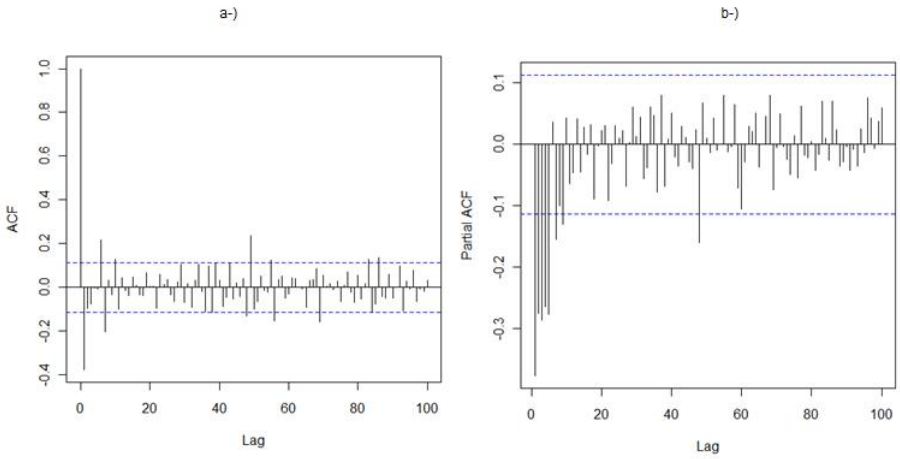

Figura 28: a-)ACF Série 7 diferenciada b-) PACF Série 7 diferenciada

Tabela 13: Tabela de critério AIC e BIC para modelos estudados na 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

6 Discussões e Resultados Finais

Analisando os 4 cenários propostos para a construção de um modelo aderente para o fluxo da RC, observou-se que a transformação foi essencial para o ajuste de modelos mais significativos. Observou-se também que os modelos de longa depedência (ARFIMA) não foram os mais indicados, não possuindo o menor AIC e também as funções não apresentavam sinais de decaimento muito lento dos lags como o visto em algumas referências.

Outro ponto considerado foi que mesmo após uma transformação os dados dos cenários 1 ao 3, não apresentaram normalidade, eles ficaram mais bem 'comportados' que os não diferenciados mas não chegaram a ter distribuição normal. No Cenário 4 após uma diferenciação apresentou normalidade ao um nível de confiança de 1%, Figura 29 e 30.

Além dos 4 cenários apresentados na seção anterior realizou-se outras análises. Considerando a alta variação dos dados aplicou-se *log* nas Série 1, 2 e 3 a fim de minimizá-la e tornar os dados mais bem ' comportados'. Outro ponto foi analisar a Série 1, separando os dados em dias típicos (dia 15 ao 19) e em dias de tráfego intenso (20 e 21) e o mesmo para a Série 2 e 3. Porém nenhum desses cenários apresentou um resultado tão satisfatório quanto a análise do Cenário 4 e também não apresentaram normalidade.

O fato do Cenário 4 utilizar a mediana como medida representativa do instante dos dias fez com que minimiza-se o impacto de algumas variações que ocorreram durante o fluxo analisado, estas

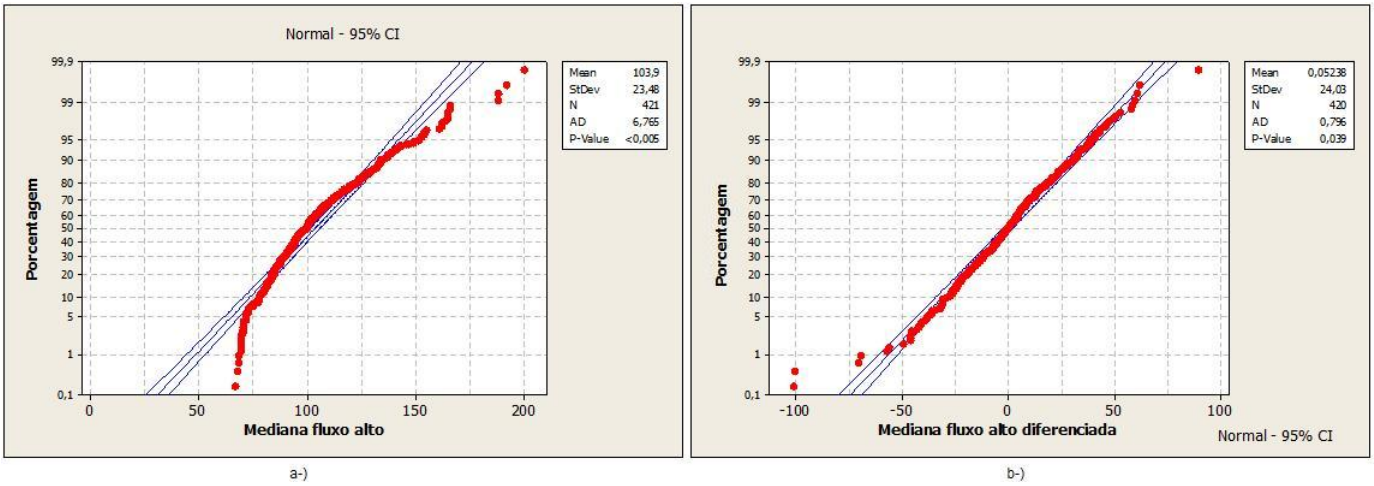

Figura 29: a-)Gráfico de probabilidade da Série 6 b-) Gráfico de probabilidade da Série 6 diferenciada

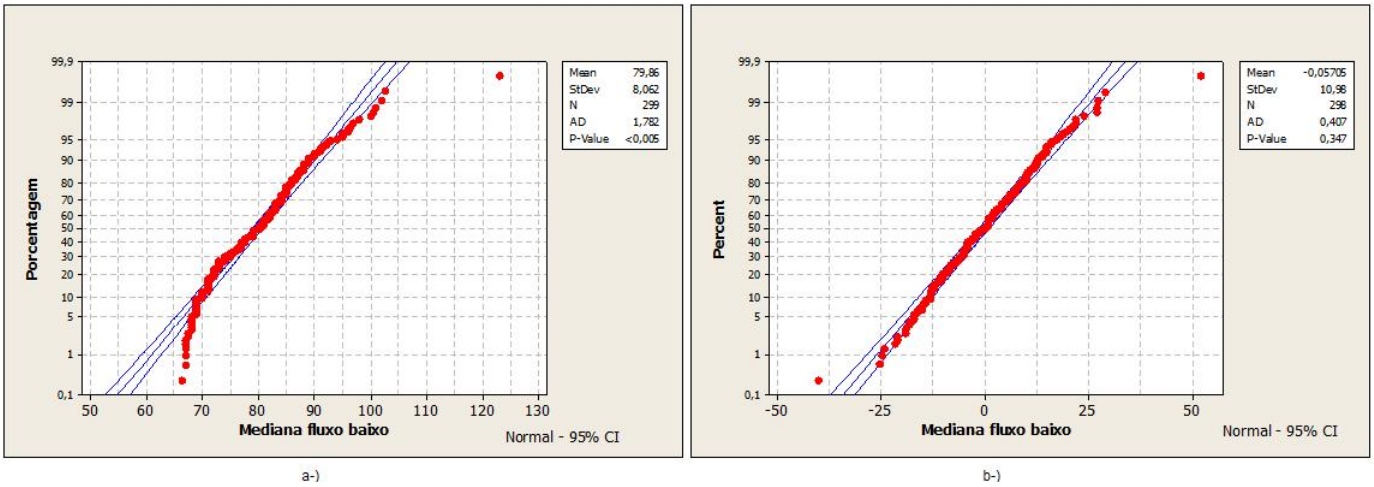

Figura 30: a-)Gráfico de probabilidade da Série 7 b-) Gráfico de probabilidade da Série 7 diferenciada

variações apenas dificultam o ajuste de um modelo, sendo que no período observado não houve nenhuma interrupção do tráfego. Portanto após todas as considerações acima escolheu-se como modelo mais indicado para os dados deste estudo o modelo proposto no Cenário 4, sendo ele um ARIMA (3,1,1). Na Tabela 14 tem-se os parâmetros respectivos aos modelos ARIMA(3,1,1) do fluxo alto, Série 6, e fluxo baixo, Série 7.

Tabela 14: Tabela Parâmetros do modelo.

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

Ajustou-se também um modelo para os períodos de tráfego mais intenso e com alta variabilidade

separadamente, observando toda a amostra nos dias 20 e 21 tem-se dois períodos durante estes dias que possuem um fluxo intenso que não foi contabilizado nos modelos no Cenário 4 como já dito. Optou-se por ajustar estes modelos caso o usuário possua um tráfego mais intenso que usual em um determinado período e necessite estimá-lo. Na Figura 31 observa-se o fluxo de pacotes ao longo do tempo, nota-se que além do número alto de pacotes comparado com a Série 6 e 7 uma alta variabilidade. Observando a ACF e PACF da série diferenciada, Figura 32, e analisando o AIC optou-se pelo modelo ARIMA(0,1,1). OS dados diferenciados tem distribuição normal, Figura 33.

Por fim para modelar a RC analisada para os dias típicos o modelo ARIMA(3,1,1) é o mais indicado e para os dias de tráfego muito intenso e com alta variabilidade indicou-se um ARIMA(0,1,1) ou um IMA(1,1).

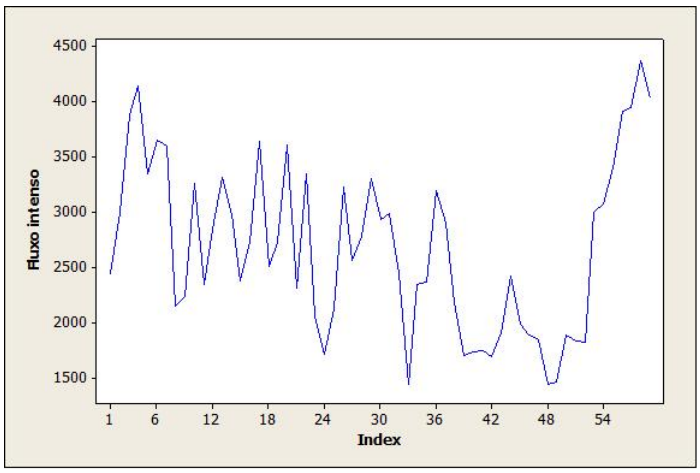

Figura 31: a-) Fluxo intenso de pacotes.

Após ajustar modelos para o tráfego da rede analisada, estudou-se dois métodos para monitorar a rede através do MMEP e CUSUM. Optou-se como método principal o MMEP devido ao resultados satísfatórios e de ser de mais fácil implementação.

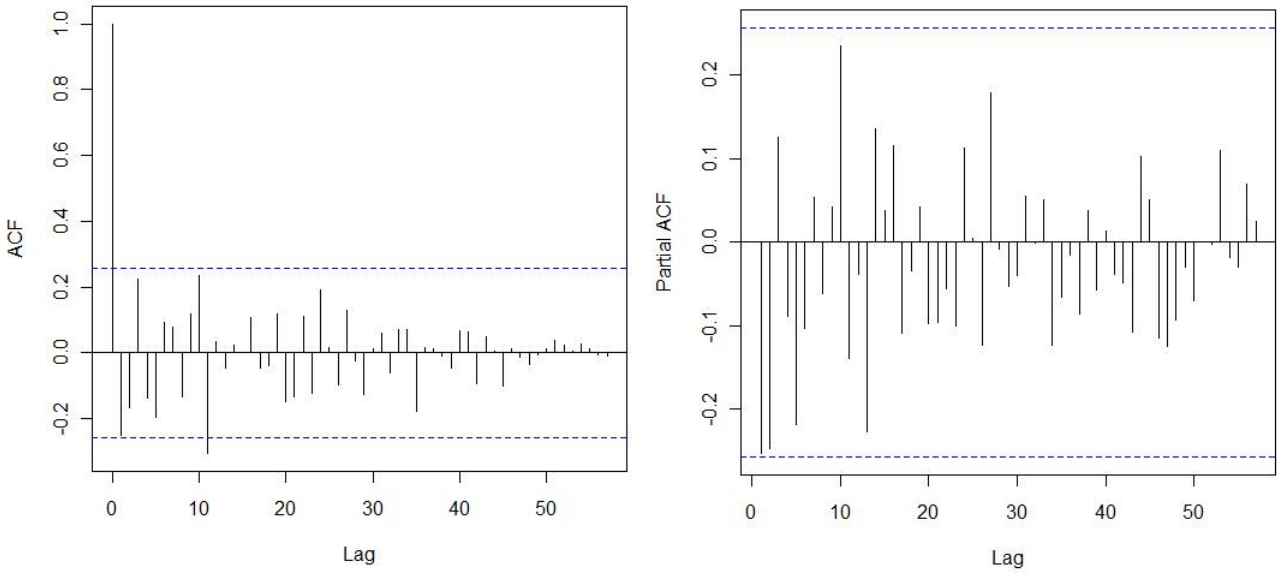

Figura 32: a-) ACF fluxo intenso diferenciado. b-)PACF fluxo intenso diferenciado.

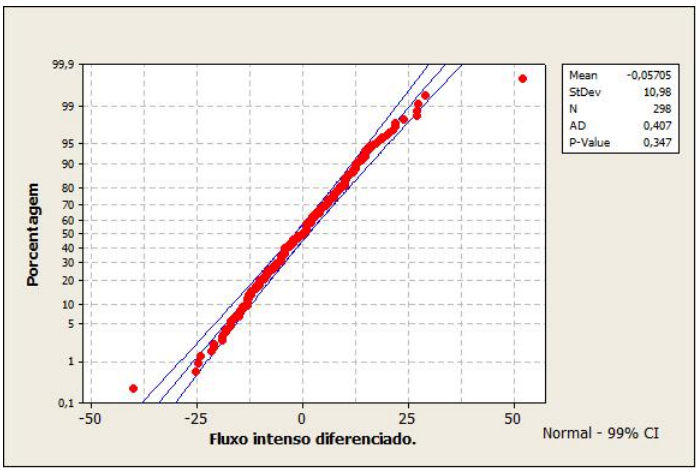

Figura 33: a-) Gráfico de probabilidade do fluxo intenso diferenciado.

7 Referências

8 Apêndice