



Ministério da
Ciência e Tecnologia



sid.inpe.br/mtc-m19/2010/10.15.13.35-TDI

**USO DE TÉCNICAS DE ANÁLISE DE SÉRIES
TEMPORAIS PARA PREVER O COMPORTAMENTO
DO RUÍDO DE FUNDO NA INTERNET BRASILEIRA
USANDO DADOS DO CONSÓRCIO BRASILEIRO DE
HONEYPOTS**

Eduardo Gomes de Barros

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada,
orientada pelos Drs. Stephan Stephany, e Antonio Montes Filho, aprovada em 19
de novembro de 2010

URL do documento original:
<<http://urlib.net/8JMKD3MGP7W/38E36T5>>

INPE
São José dos Campos
2010

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE

Gabinete do Diretor (GB)

Serviço de Informação e Documentação (SID)

Caixa Postal 515 - CEP 12.245-970

São José dos Campos - SP - Brasil

Tel.:(012) 3208-6923/6921

Fax: (012) 3208-6919

E-mail: pubtc@sid.inpe.br

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELLECTUAL DO INPE (RE/DIR-204):**Presidente:**

Dr. Gerald Jean Francis Banon - Coordenação Observação da Terra (OBT)

Membros:

Dr^a Inez Staciarini Batista - Coordenação Ciências Espaciais e Atmosféricas (CEA)

Dr^a Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação

Dr^a Regina Célia dos Santos Alvalá - Centro de Ciência do Sistema Terrestre (CST)

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Dr. Ralf Gielow - Centro de Previsão de Tempo e Estudos Climáticos (CPT)

Dr. Wilson Yamaguti - Coordenação Engenharia e Tecnologia Espacial (ETE)

Dr. Horácio Hideki Yanasse - Centro de Tecnologias Especiais (CTE)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon - Coordenação de Observação da Terra (OBT)

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Deicy Farabello - Centro de Previsão de Tempo e Estudos Climáticos (CPT)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)

EDITORAÇÃO ELETRÔNICA:

Vivêca Sant'Ana Lemos - Serviço de Informação e Documentação (SID)



Ministério da
Ciência e Tecnologia



sid.inpe.br/mtc-m19/2010/10.15.13.35-TDI

**USO DE TÉCNICAS DE ANÁLISE DE SÉRIES
TEMPORAIS PARA PREVER O COMPORTAMENTO
DO RUÍDO DE FUNDO NA INTERNET BRASILEIRA
USANDO DADOS DO CONSÓRCIO BRASILEIRO DE
HONEYPOTS**

Eduardo Gomes de Barros

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada,
orientada pelos Drs. Stephan Stephany, e Antonio Montes Filho, aprovada em 19
de novembro de 2010

URL do documento original:
<<http://urlib.net/8JMKD3MGP7W/38E36T5>>

INPE
São José dos Campos
2010

Dados Internacionais de Catalogação na Publicação (CIP)

- B278 Barros, Eduardo Gomes de.
 Uso de técnicas de análise de séries temporais para prever o comportamento do ruído de fundo na internet brasileira usando dados do consórcio brasileiro de honeypots / Eduardo Gomes de Barros. – São José dos Campos : INPE, 2010.
 xxvi+ 148 p. ; (sid.inpe.br/mtc-m19/2010/10.15.13.35-TDI)

 Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2010.
 Orientadores : Drs. Stephan Stephany, e Antonio Montes Filho.

 1. Honeypots. 2. Consórcio Brasileiro de Honeypots (CBH).
 3. Ruído de fundo. 4. Predição de eventos. I.Título.

CDU 681.324

Copyright © 2010 do MCT/INPE. Nenhuma parte desta publicação pode ser reproduzida, armazenada em um sistema de recuperação, ou transmitida sob qualquer forma ou por qualquer meio, eletrônico, mecânico, fotográfico, reprográfico, de microfilmagem ou outros, sem a permissão escrita do INPE, com exceção de qualquer material fornecido especificamente com o propósito de ser entrado e executado num sistema computacional, para o uso exclusivo do leitor da obra.

Copyright © 2010 by MCT/INPE. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, microfilming, or otherwise, without written permission from INPE, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use of the reader of the work.

Aprovado (a) pela Banca Examinadora
em cumprimento ao requisito exigido para
obtenção do Título de Doutor(a) em
Computação Aplicada

Dr. Luciano Vieira Dutra



Presidente / INPE / SJC Campos - SP

Dr. Stephan Stephany



Orientador(a) / INPE / SJC Campos - SP

Dr. Antonio Montes Filho



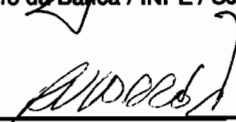
Orientador(a) / CTI/RA / Campinas - SP

Dr. Ulisses Thadeu Vieira Guedes



Membro da Banca / INPE / SJC Campos - SP

Dr. Ricardo Varela Correa



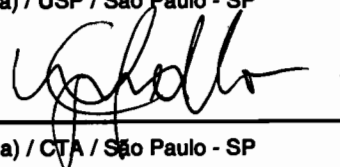
Membro da Banca / INPE / São José dos Campos - SP

Dr. Alberto Camilli



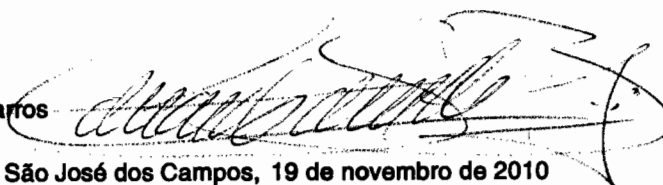
Convidado(a) / USP / São Paulo - SP

Dr. Cláudio Gomes de Mello



Convidado(a) / CTA / São Paulo - SP

Aluno (a): Eduardo Gomes de Barros



São José dos Campos, 19 de novembro de 2010

À minha família que, mais uma vez, soube entender o quão importante era este passo para mim sacrificando-se para que ele pudesse se concretizar.

AGRADECIMENTOS

Ao meu grande amigo Ricardo Varela que sempre esteve presente, nos momentos mais fáceis e nos mais difíceis, apoiando sempre, mesmo quando não se pedia.

A toda equipe de projeto do Consórcio Brasileiro de Honeypots, do CTI, CENPRA à época em que este trabalho se iniciou, sempre que solicitada, teve boa vontade para atender os pedidos feitos.

Em especial meus pais sem os quais eu não estaria aqui hoje.

À minha esposa, filhos e neta sem os quais minha vida não teria significado.

RESUMO

O tráfego capturado pelos sensores do Consórcio Brasileiro de Honeypots (CBH) revela a existência de um tráfego que existe na Internet independentemente do tipo de máquina sendo usada ou do tipo de serviço sendo prestado: o ruído de fundo – todo tráfego não produtivo, seja ele malicioso ou não. As atividades maliciosas que ocorrem na parcela brasileira da Internet estão embutidas neste tráfego. Conhecer-lo, caracterizá-lo e descrevê-lo é um desafio que auxiliará na emissão de alertas precoces, paradigma de segurança necessário para a defesa das infraestruturas críticas de uma Nação que vem complementar o atual, o reativo. A partir dos dados do Consórcio Brasileiro de Honeypots criou-se uma metodologia para sanitização dos mesmos que permitiu que servissem de base para construção de séries temporais. A partir destas séries foi possível a caracterização e a descrição deste tráfego na parcela brasileira da Internet. A modelagem matemática utilizada permitiu a projeção de eventos futuros e a análise de quando alertas precoces devem ser emitidos.

USE OF TEMPORAL SERIES ANALYSIS OVER THE BRAZILIAN HONEY- POT ALLIANCE DATA TO PREDICT THE BRAZILIAN INTERNET BACK- GROUND RADIATION BEHAVIOR

ABSTRACT

The traffic captured by the sensors of the Brazilian Honeypots Alliance (CBH) reveals the existence of a traffic that exists on the Internet regardless of the type of machine or the service being provided: the background noise - all non-productive traffic, whether malicious or not. The malicious activities occurring in the Brazilian portion of the Internet are embedded in this traffic. Know it, characterize it and describe it is a challenge that will help the issue of early warnings, the security paradigm necessary for the protection of Nation's critical infrastructures and which complements the current, the reactive. From the Brazilian Honeypots Alliance data a methodology was created for data sanitization and allowing its use for constructing time series. From these series the characterization and description of the Brazilian Internet traffic was possible. The mathematical model used allows the projection of future events and the analysis of when early warnings should be issued.

LISTA DE FIGURAS

	<u>Pág.</u>
Figura 1.1 - Programas maliciosos e softwares não desejados na parcela brasileira da Internet, classificados por categoria, pela SourceFire, no primeiro semestre de 2009.....	2
Figura 2.1 - Localização dos Honeypots.....	12
Figura 3.1 - Curva de distribuição gaussiana com os limites usados na segunda métrica para seleção da amostra de trabalho.....	28
Figura 3.2 - Histograma de distribuição dos pontos obtidos após aplicação dos critérios de seleção.....	41
Figura 4.1 - Percentual de fluxos dos protocolos TCP, UDP e ICMP no ruído de fundo da parcela brasileira da Internet.....	43
Figura 4.2 - Percentual de fluxos TCP com menos de 3 e com 3 ou mais pacotes no ruído de fundo da parcela brasileira da Internet.....	44
Figura 4.3 - (a) Fluxo TCP com menos de 3 pacotes na primeira semana de janeiro de 2005, usando janela de amostragem hora, nas 26 portas mais significativas; (b) Fluxo TCP com 3 ou mais pacotes na mesma situação.....	49
Figura 4.4 - Quantidade de fluxos TCP com menos de 3 pacotes agrupados por endereço IP de origem.....	52
Figura 4.5 - Quantidade média de fluxos TCP com 3 ou mais pacotes agrupados por endereço IP de origem.....	53
Figura 4.6 - Relacionamento entre endereços IP de origem e as portas de destino, janela de amostragem hora, fluxo TCP com menos de 3 pacotes, durante a primeira semana de janeiro de 2005.....	54
Figura 4.7 - Distribuição dos fluxos TCP com menos de 3 pacotes em relação aos horários do dia.....	55
Figura 4.8 - Distribuição dos fluxos TCP com 3 ou mais pacotes em relação aos horá-	

	rios do dia.....	55
Figura 4.9 -	Distribuição de frequência dos fluxos TCP com menos de 3 pacotes.....	56
Figura 4.10 -	Distribuição de frequência dos fluxos TCP com 3 ou mais pacotes.....	57
Figura 7.1 -	Fluxo dos protocolos TCP, UDP e ICMP, janela de amostragem dia, como série temporal.....	81
Figura 7.2 -	Boxplot do fluxo do protocolo TCP, janela de amostragem dia, agrupado a cada 7 dias.....	82
Figura 7.3 -	Média versus desvio padrão para agrupamentos de fluxos médios TCP, janela de amostragem dia, a cada 7 dias.....	83
Figura 7.4 -	Logaritmo dos fluxos dos protocolos TCP, UDP e ICMP, janela de amostragem dia.....	84
Figura 7.5 -	Logaritmo dos fluxos dos protocolos TCP, UDP e ICMP, janela de amostragem hora.....	84
Figura 7.6 -	Análise da série temporal do fluxo do protocolo TCP já aplicada a função de transformação logarítmica, janela de amostragem dia: série, correlograma, correlograma parcial e densidade espectral.....	90
Figura 7.7 -	Análise da série temporal do protocolo TCP com janela de amostragem dia: histograma e gráfico Q-Q.....	91
Figura 7.8 -	Análise da série temporal do fluxo do protocolo TCP já aplicada a função de transformação logarítmica, janela de amostragem hora: série, correlograma, correlograma parcial e densidade espectral.....	92
Figura 7.9 -	Análise da série temporal do protocolo TCP com janela de amostragem hora: histograma e gráfico Q-Q.....	93
Figura 7.10 -	Gráfico de dispersão do fluxo TCP, com janela de amostragem dia, para intervalos de 1 a 4 dias.....	94
Figura 7.11 -	Gráfico de dispersão do fluxo TCP, com janela de amostragem hora, para intervalos de 12, 24, 36 e 48 horas.....	94

Figura 7.12 - Análise da série temporal resultante da aplicação da primeira diferença com intervalo 1, janela de amostragem dia: série, correlograma, correlograma parcial e densidade espectral.....	95
Figura 7.13 - Análise da série temporal resultante da aplicação da primeira diferença com intervalo 24, janela de amostragem hora: série, correlograma, correlograma parcial e densidade espectral.....	96
Figura 7.14 - Análise da série temporal resultante da aplicação de filtragem com janela de tamanho 3, pesos 0,85, 0,1 e 0,05, janela de amostragem dia.....	98
Figura 7.15 - Análise da série temporal resultante da aplicação de filtragem com janela de tamanho 3, pesos 0,85, 0,1 e 0,05, janela de amostragem hora.....	99
Figura 7.16 - Análise da série temporal resultante da aplicação de suavização exponencial dupla com $\alpha = 0.493056$ e $\beta = 0.1682243$, janela de amostragem dia. .	100
Figura 7.17 - Análise da série temporal resultante da aplicação de suavização exponencial dupla com $\alpha = 0.557775$ e $\beta = 0.2012012$, janela de amostragem hora	101
Figura 7.18 - Gráfico da densidade espectral, suavizado e limitado ao intervalo [0, 0.2] das frequências, para fluxo TCP com janela de amostragem dia.....	102
Figura 7.19 - Gráfico da densidade espectral, suavizado e limitado ao intervalo [0, 0.05] das frequências, para fluxo TCP com janela de amostragem hora.....	103
Figura 7.20 - Análise da série temporal resultante da aplicação de suavização exponencial tripla com $\alpha = 0.493056$, $\beta = 0.1682243$ e $\gamma = 0.3850921$, janela de amostragem dia.....	104
Figura 7.21 - Análise da série temporal resultante da aplicação de suavização exponencial tripla com $\alpha = 0.557775$, $\beta = 0.2012012$ e $\gamma = 0.03205071$, janela de amostragem hora.....	105
Figura 7.22 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 1$, d	

	=0 e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem dia.....	106
Figura 7.23 -	Análise do resíduo resultante da aplicação de auto-regressão com $p = 1$, $d = 1$ e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem hora.....	107
Figura 7.24 -	Análise do resíduo resultante da aplicação de auto-regressão sazonal com $p = 1$, $d = 1$, $q = 2$, $P = 1$, $D = 0$ e $Q = 1$ ao fluxo TCP, já aplicada a transformação logarítmica, com janela de amostragem hora.....	108
Figura 8.1 -	Previsão do comportamento do ruído de fundo, do fluxo TCP, da parcela brasileira da Internet, já aplicada a transformação logarítmica, com janela de amostragem dia.....	113
Figura 8.2 -	Previsão do comportamento do ruído de fundo, do fluxo TCP, da parcela brasileira da Internet, já aplicada a transformação logarítmica, com janela de amostragem hora.....	114
Figura 8.3 -	Previsão do comportamento do ruído de fundo da parcela brasileira da Internet, com $p = 1$, $d = 0$ e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem dia.....	115
Figura 8.4 -	Previsão do comportamento do ruído de fundo da parcela brasileira da Internet, com $p = 1$, $d = 1$ e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem hora.....	116
Figura 9.1 -	Distribuição do fluxo TCP para portas 135,139, 445, 1080 e 1433, já aplicada a transformação logarítmica, com janela de amostragem hora.....	121
Figura 9.2 -	Representação do gráfico QQ para portas 135,139, 445, 1080 e 1433, já aplicada a transformação logarítmica, para as janelas de amostragem dia e hora considerando tráfego de varredura e de conexão.....	122
Figura 9.3 -	Representação do gráfico de correlação cruzada para portas 135,139, 445, 1080 e 1433, já aplicada a transformação logarítmica, para as janelas de	

	amostragem dia e hora considerando tráfego de varredura e de conexão....	123
Figura 9.4 -	Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 1$ e $q = 2$, no fluxo de conexão para porta 135, já aplicada a transformação logarítmica, com janela de amostragem hora.....	125
Figura 9.5 -	Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 1$ e $q = 2$, no fluxo de conexão para porta 139, já aplicada a transformação logarítmica, com janela de amostragem hora.....	126
Figura 9.6 -	Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 0$ e $q = 2$, no fluxo de conexão para porta 445, já aplicada a transformação logarítmica, com janela de amostragem hora.....	127
Figura 9.7 -	Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 1$ e $q = 2$, no fluxo de conexão para porta 1080, já aplicada a transformação logarítmica, com janela de amostragem hora.....	128
Figura 9.8 -	Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 0$ e $q = 2$, no fluxo de conexão para porta 1433, já aplicada a transformação logarítmica, com janela de amostragem hora.....	129
Figura 9.9 -	Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 135, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo.....	131
Figura 9.10 -	Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 139, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo.....	132
Figura 9.11 -	Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 445, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo.....	133
Figura 9.12 -	Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 1080, já aplicada a transformação logarítmica, com janela	

	de amostragem hora e modelo auto-regressivo.....	135
Figura 9.13 -	Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 1433, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo.....	135

LISTA DE TABELAS

	<u>Pág.</u>
Tabela 2.1 - Sensores do CBH e suas localizações.....	13
Tabela 3.1 - Registros de tráfego encontrados nos arquivos sanitizados do CBH.....	19
Tabela 3.2 - Quantidade de dias que cada sensor do CBH forneceu dados, por ano...23	23
Tabela 3.3 - Percentuais de fluxos dos diferentes protocolos de transporte em relação ao fluxo total.....	24
Tabela 3.4 - Primeiro critério de seleção de amostra: número de dias com fluxo.....	27
Tabela 3.5 - Segunda métrica: fluxo TCP em 2005. Comparação com valor mínimo....	30
Tabela 3.6 - Segunda métrica: fluxo UDP em 2005. Comparação com o valor máximo	31
Tabela 3.7 - Segunda métrica: fluxo ICMP em 2005. Comparação com valor máximo.	32
Tabela 3.8 - Segunda métrica: fluxo TCP em 2006. Comparação com valor mínimo....	33
Tabela 3.9 - Segunda métrica: fluxo UDP em 2006. Comparação com valor máximo...	34
Tabela 3.10 - Segunda métrica: fluxo ICMP em 2006. Comparação com valor máximo.	35
Tabela 3.11 - Terceira métrica: fluxo TCP < 3 em 2005. Comparação com valor mínimo	36
Tabela 3.12 - Terceira métrica: fluxo TCP >= 3 em 2005. Comparação com valor máximo.....	37
Tabela 3.13 - Terceira métrica: fluxo TCP < 3 em 2006. Comparação com valor mínimo	38

Tabela 3.14 - Terceira métrica: fluxo TCP ≥ 3 em 2006. Comparação com valor máximo.....	39
Tabela 3.15 - Pontuação obtida pelos sensores após aplicação das métricas.....	40
Tabela 4.1 - Percentual de fluxos TCP com menos e com mais de 3 pacotes, por porta, para as janelas hora e dia.....	46
Tabela 4.2 - Portas mais acessadas, seus usos corretos e possíveis vulnerabilidades/ataques.....	47
Tabela 4.3 - Espaço de endereçamento IP, por sensor, do CBH.....	50
Tabela 4.4 - Espaço de endereçamento IP remoto, por sensor.....	51
Tabela 4.5 - Médias e desvios padrões para os fluxos TCP com menos e com mais de 3 pacotes em relação ao número de sensores.....	57

LISTA DE ABREVIATURAS E SIGLAS

ACF	Autocorrelation Function – Função de Autocorrelação
AIC	Akaike Information Criterion – Critério de Ajustamento da Informação de Akaike
ARMA	Autoregressive Moving Average Model – Modelo de Médias Móveis Auto-regressivo
ARIMA	Autoregressive Integrated Moving Average – Modelo Integrado de Médias Móveis
AR(p)	Modelo auto-regressivo de ordem p
BSM	Basic Structural Model – Modelo Estrutural Básico
CBH	Consórcio Brasileiro de Honeypots
CenPRA	Centro de Pesquisas Renato Archer
CERT.br	Centro de Estudos, Resposta e Tratamento de Incidentes
CRAN	Comprehensive R Archive Network – Rede Abrangente de Arquivos R
CTI	Centro de Tecnologia da Informação Renato Archer
DCOM	Distributed Component Object Model – Modelo de Componentes de Objetos Distribuídos
DFT	Discrete Fourier Transformation – Transformada Discreta de Fourier
DNS	Domain Name System – Sistema de Nomes de Domínio
DoS	Deny of Service – Negação de Serviço
epmap	Endpoint Mapper – Software para o Windows NT cuja finalidade é a mesma do RPC
FFT	Fast Fourier Transformation – Transformada Rápida de Fourier
HTTP	Hypertext Transfer Protocol – Protocolo de Transferência de Hipertexto
IANA	Internet Assigned Numbers Authority – Autoridade para Atribuição de Números na Internet
ICMP	Internet Control Message Protocol – Protocolo de Mensagens de Controle na Internet
IDS	Intrusion Detection System – Sistema de Detecção de Intrusão

iid	independentes e identicamente distribuídas
IIQ	Intervalo Inter-Quartílico
IP	Internet Protocol – Protocolo da Internet
IPS	Intrusion Prevention System – Sistema de Prevenção de Intrusão
MA(q)	Modelo de média móvel de ordem q
MAD	Mean Absolute Deviation – Desvio Absoluto da Média
MSE	Mean Square Error – Erro Médio Quadrático
NAT	Network Address Translation – Tradução de Endereços de Rede
NETBIOS	Network Basic Input/Output System – Sistema Básico de Entrada/Saída para Rede
opencap	Protocolo para troca de informações de calendário na Internet
PACF	Partial Autocorrelation Function – Função de Autocorrelação Parcial
Q-Q	Gráfico Quantile-Quantile
TCP	Transmission Control Protocol – Protocolo de Controle de Transmissão
radmin	Software para administração remota
RPC	Remote Procedure Call – Chamada Remota de Procedimentos
SARIMA	Seazonal ARIMA – Modelo Integrado de Médias Móveis Sazonal
SMB	Server Message Block – Bloco de Mensagens do Servidor
SQL	Structured Query Language – Linguagem de Consulta Estruturada
SSH	Secure Shell – Shell Seguro
UDP	User Datagram Protocol – Protocolo de Datagramas de Usuário

SUMÁRIO

	<u>Pág.</u>
1 INTRODUÇÃO.....	1
1.1 Objetivo.....	5
2 REVISÃO BIBLIOGRÁFICA.....	7
2.1 Honeypot e Honeynet.....	9
2.2 Ruído de fundo.....	10
2.2.1 Redes Telescópio.....	10
2.2.2 Honeyfarms.....	11
2.2.3 CBH.....	11
2.3 Análise de séries temporais.....	13
2.3.1 Teoria De Previsão.....	15
3 DADOS DO CBH.....	17
3.1 Pré-processamento dos dados.....	18
3.2 Fluxo.....	21
3.3 Dados utilizados.....	21
3.4 Situações especiais.....	23
3.5 Seleção da amostra de trabalho.....	25
3.5.1 Métrica: Número De Dias.....	25
3.5.2 Métrica: Relação Entre Fluxos De Protocolos.....	27
3.5.3 Métrica: Relação Entre Fluxos TCP.....	35
3.5.4 Conjunto De Sensores Seleccionados Pela Aplicação Da Métrica.....	39
4 CARACTERÍSTICAS DO RUÍDO DE FUNDO DA INTERNET BRASI-	

LEIRA.....	43
4.1 Percentual de fluxos TCP, UDP e ICMP no ruído de fundo.....	43
4.2 Percentual de fluxos TCP.....	44
4.3 Portas mais acessadas.....	44
4.4 Endereçamento IP.....	50
4.5 Comportamento dos fluxos TCP.....	54
4.6 Modelagem do ruído de fundo.....	57
5 ANÁLISE DE SÉRIES TEMPORAIS.....	61
5.1 Conceitos básicos.....	63
5.2 Componentes de séries não estacionárias.....	65
5.2.1 Componente De Tendência.....	65
5.2.1.1 Regressão.....	66
5.2.1.2 Diferenciação.....	67
5.2.1.3 Suavização/Filtragem.....	68
5.2.2 Componente De Sazonalidade.....	69
5.2.2.1 Diferenciação.....	70
5.2.2.2 Suavização.....	70
5.2.3 Componentes Estruturados.....	71
5.2.3.1 Processos auto-regressivos.....	71
5.2.3.2 SARIMA.....	73
6 PREDIÇÃO DE EVENTOS USANDO SÉRIES TEMPORAIS.....	75
6.1 Erros	76
6.2 Médias móveis.....	77

6.3	Suavização exponencial.....	77
6.4	Modelos auto-regressivos.....	79
7	ANÁLISE DE SÉRIES TEMPORAIS APLICADA AOS DADOS DO CBH....	81
7.1	Visualização dos dados do CBH.....	81
7.2	Análise do Fluxo TCP.....	85
7.3	Remoção do Componente de Tendência.....	95
7.3.1	Diferenciação.....	95
7.3.2	Suavização/Filtragem.....	97
7.3.2.1	Suavização exponencial dupla.....	99
7.4	Remoção do componente de sazonalidade.....	101
7.4.1	Suavização Exponencial Tripla.....	103
7.5	Remoção de estruturas auto-regressivas.....	105
7.5.1	SARIMA.....	107
8	PREDIÇÃO DE EVENTOS FUTUROS.....	111
8.1	Filtragem.....	112
8.2	Processo auto-regressivo integrado.....	115
9	FLUXOS POR PORTAS.....	119
9.1	Aderência dos dados à distribuição normal.....	120
9.2	Análise conjunta dos fluxos TCP.....	121
9.3	Correlação entre os fluxos TCP.....	123
9.4	Análise por filtragem.....	124
9.5	Análise por auto-regressão integrada.....	124

9.6	Predição.....	129
10	CONCLUSÃO.....	137
10.1	Modelo preditivo de segurança.....	137
10.2	Características do ruído de fundo.....	138
10.3	Análise estatística dos dados como séries temporais.....	140
10.4	Geração de alertas.....	141
10.5	Sugestões para pesquisas futuras.....	143
	REFERÊNCIAS BIBLIOGRÁFICAS.....	145

1 INTRODUÇÃO

Nós estamos vivendo na sociedade da informação. A dependência que temos de sistemas automatizados e, principalmente, de sistemas interconectados é cada vez maior.

A Internet é a maior ferramenta de ligação entre as informações e os usuários, sejam eles simples consumidores, grandes corporações ou mesmo provedores de serviços de infraestrutura críticas da nação.

Este meio de comunicação e de troca de dados tão útil também se tornou fonte de problemas. Num mundo sem fronteiras e com pouca vigilância atividades não legítimas podem se desenvolver no seu meio.

Numa sociedade utópica todo o tráfego da Internet deveria ser legítimo, isto é, de usuários bem intencionados relacionando-se com outros usuários e/ou prestadores de serviços que lhes forneceria a informação desejada pura e simplesmente.

Na sociedade real o tráfego legítimo está imerso em outro tipo de tráfego, que circula pelos mesmos meios físicos, o ruído de fundo. Segundo Pang et al (2004) ruído de fundo é todo tráfego não produtivo seja ele malicioso (tentativas de DoS, varreduras, sondagens, ...) ou benigno (tráfego gerado a partir de uma máquina ou serviço mal configurado...).

O tráfego que pode ser caracterizado como complexo, altamente automatizado, malicioso e que sofre mutações num curto espaço de tempo, tais como:

- a) varreduras: são programas usados para buscar portas abertas em sistemas remotos. Os mais especializados tentam reconhecer o programa servidor e o sistema operacional remoto verificando se há vulnerabilidades conhecidas;
- b) worms: segundo Arce e Levy (2003), são programas que se executam sozinhos, isto é, não necessitam da interferência do ser humano, e que propagam uma versão totalmente funcional de si mesmos para outras máquinas; e,

c) bots: segundo The Honeynet Project (2010a), são programas que, uma vez instalados numa máquina, possuem um mecanismo de comunicação com o invasor permitindo que sejam controladas remotamente.

são os principais componentes do ruído ou radiação de fundo.

Como se pode ver na Figura 1.1, desenvolvida a partir de estudos realizados por empresa de detecção de intrusão que, junto com o Exército Brasileiro, montaram um sistema de prevenção de incidentes de redes a nível nacional, worms são a segunda maior forma de malware perdendo somente para roubos de senhas e ferramentas de monitoração (varreduras).

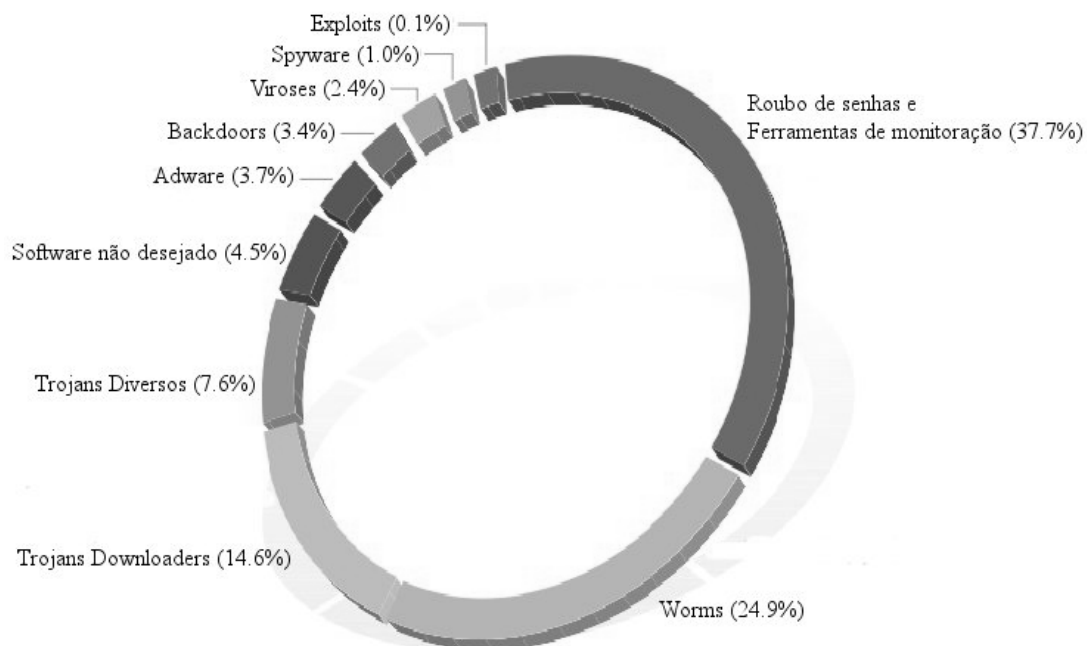


Figura 1.1 - Programas maliciosos e softwares não desejados na parcela brasileira da Internet, classificados por categoria, pela SourceFire, no primeiro semestre de 2009
Fonte: Adaptada de Kirk (2010).

Os efeitos maléficos de um ataque maciço de bots ou de worms já pode ser dimensionado graças a ataques recentes. Entretanto, não se pode dimensionar, com segurança, quais

seus efeitos quando a infraestrutura crítica de uma nação¹ é ameaçada. Uma das formas de ameaçá-la é com o uso maciço de ataques de worms e de bots. Portanto, a detecção precoce deste tipo de evento torna-se, cada vez mais, um fator decisivo para segurança das infraestruturas críticas de uma nação.

Segundo Savage et al (2010), se preocupa com a repercussão que bots e worms possam causar de forma global. Ele afirma que o grande desafio para analisar fenômenos mundiais como os worms e os bots é a aquisição de dados em quantidade suficiente que garanta a representatividade do fenômeno sendo observado.

Uma possível ferramenta a ser usada para fornecer a visibilidade necessária é a rede telescópio – redes que monitoram o tráfego enviado para porções não alocadas do endereçamento IP. Como não há motivo legítimo para que uma máquina envie pacotes para estes endereços todo o tráfego a eles direcionado sugere a ocorrência de atividades maliciosas.

Segundo Pang et al (2004), a maioria dos estudos que envolvem a coleta de dados maliciosos ou de ruído de fundo da Internet têm sido realizadas em redes telescópio. O espaço dos endereços IP não usados se tornou uma importante fonte de informação sobre intrusão e atividades de ataque.

A primeira tentativa de caracterizar o ruído de fundo da Internet foi feita por Pang et al (2004) usando uma rede telescópio.

O uso de redes telescópio possui pelo menos dois inconvenientes: a necessidade de um grande número de máquinas para que o resultado seja expressivo, e a atitude passiva das máquinas não permite que se capturem informações mais completas sobre um ataque realizado.

O uso de honeypots distribuídos é uma tentativa de superar as restrições existentes nas

1 as Infraestruturas Críticas são as instalações, serviços, bens e sistemas que, se forem interrompidos ou destruídos, provocarão sério impacto social, econômico, político, internacional ou à segurança do Estado e da sociedade. Portaria nº 2-GSI, de 8 de fevereiro de 2008, publicado no DOU nº 27, de 11 de fevereiro de 2008.

redes telescópio. Dentre os modelos de redes de honeypots distribuídos existentes no mundo, o adotado no Brasil, através do CBH, é único porque ele permite a captura de informações no espaço de endereçamento real da parcela brasileira da Internet.

Segundo CBH (2010), o Consórcio Brasileiro de Honeypots é um projeto que objetiva aumentar a capacidade de detecção de incidentes, correlação de eventos e determinação de tendências de ataques no espaço da Internet brasileira. É implementado através da aliança de 40 instituições, aproximadamente, distribuídas por todo o Brasil, coordenadas pelo CTI e pelo CERT.br.

Honeypots, se considerados individualmente, não são as ferramentas mais adequadas para gerar avisos precoces. Entretanto, a colocação de vários honeypots viabiliza a detecção das fases iniciais de alguns ataques permitindo, assim, a geração de alertas precoces.

Este trabalho usa o sistema distribuído de honeypots – CBH – para caracterizar e modelar o ruído de fundo na área abrangida pelos seus sensores e gerar avisos precoces de atividades maliciosas na parcela brasileira da Internet.

Segundo The HoneyNet Project (2010b) é possível realizar predição de eventos malignos usando dados de um honeypot, entretanto, na revisão bibliográfica não foram encontrados outros trabalhos nesta área.

Neste trabalho, para tratar do problema da caracterização, da modelagem e da predição, foram usadas várias técnicas: programação dinâmica, mineração de dados, redução de dimensionalidade de séries temporais, discretização dos dados reduzidos e análise de séries temporais.

Das abordagens testadas as técnicas de análise de séries temporais apresentaram bons resultados e foram escolhidas como ferramenta de análise, modelagem e predição neste trabalho.

1.1 Objetivo

Gerar estatísticas de diferentes parâmetros contidos nos dados do CBH para caracterizar o ruído de fundo na parcela brasileira da Internet. Esta caracterização fornece um panorama, pelo menos na média, aos administradores de sistemas conectados à parcela brasileira da Internet, particularmente os responsáveis por redes e segurança, do que se pode esperar de tráfego não legítimo em qualquer situação.

Usar técnicas de análise de séries temporais para modelar o ruído de fundo e, a partir do modelo adotado, gerar previsões de eventos futuros. A visão de futuro é que o modelo matemático levantado venha a permitir a criação de uma ferramenta para geração de ruído de fundo em laboratório de forma a reproduzir, o mais próximo possível, o ruído de fundo real e permitir testes mais robustos de segurança nos sistemas computacionais. A geração de previsões de eventos futuros pode permitir a emissão de alertas precoces para eventos de grande amplitude que ocorram na parcela brasileira da Internet e, mais particularmente, nas infraestruturas críticas nacionais.

Para permitir a consecução dos objetivos levantados foi necessário o desenvolvimento de uma metodologia de tratamento dos dados do CBH que permitem sua caracterização, modelagem e previsão, qualquer que seja o intervalo de tempo de interesse.

Para alcançar os objetivos propostos esta tese foi dividida nos seguintes capítulos:

- a) Capítulo 1: introdução;
- b) Capítulo 2: revisão bibliográfica;
- c) Capítulo 3: dados recebidos e pré-processamento realizado;
- d) Capítulo 4: características levantadas do ruído de fundo na parcela brasileira da Internet;

- e) Capítulo 5: técnicas para análise de séries temporais;
- f) Capítulo 6: técnicas para predição de eventos usando séries temporais;
- g) Capítulo 7: aplicação das técnicas de análise de séries temporais e de predição de eventos nos dados do CBH. Resultados encontrados e metodologia utilizada para selecionar os modelos matemáticos a serem empregados na predição;
- h) Capítulo 8: predição de eventos futuros para os fluxos do CBH usando modelos de séries temporais;
- i) Capítulo 9: predição de eventos futuros para os fluxos por porta, usando modelos de séries temporais, e a metodologia levantada até este ponto; e,
- j) Capítulo 10: conclusões.

2 REVISÃO BIBLIOGRÁFICA

Para a consecução dos objetivos da tese foi realizada uma revisão bibliográfica extensa sobre vários assuntos que são direta ou indiretamente ligados ao problema sendo tratado.

Sobre caracterização do ruído de fundo foram analisados os seguintes artigos:

- a) Pang et all (2004), que trata da caracterização do ruído de fundo usando uma rede telescópio;
- b) Richardson et all (2005), que usa uma combinação de modelagem analítica, simulação e medidas para entender o relacionamento entre ruído de fundo e a fidelidade das detecções das varreduras dos worms; e,
- c) Grizzard et all (2005), que compara o tráfego obtido por um conjunto de máquinas residenciais de usuários com o tráfego obtido por uma honeynet usando uma série de combinações de parâmetros úteis para caracterização do fluxo de dados, particularmente na representação visual dos dados sendo comparados; e
- d) Viinikka et all (2004), que usam mapas de médias móveis exponenciais ponderadas para ajudar o operador de sistemas de detecção de intrusão na determinação de quais alertas são mais significativos. O foco principal deste artigo não está na caracterização mas sim na redução do número de alertas gerados por sistemas de detecção de intrusão.

Sobre o uso de séries temporais na modelagem dos dados com vistas a sua aplicação na área de segurança tem-se o trabalho de Viinikka et all (2006).

Sobre previsão tem-se The Honeynet Project (2010b) que testa o conceito de predição e alerta precoce inerente aos sistemas de honeypot concluindo que é possível predizer um ataque dias antes dele ocorrer.

Sobre a propagação e o comportamento de worms e bots tem-se os trabalhos de:

- a) Staniford et al. (2010), que apresentam uma análise da magnitude da ameaça causada por worms e sugerem um modelo matemático para sua propagação;
- b) CAIDA (2010), que detalha o comportamento das três versões do Code-Red;
- c) Arce e Levy (2003), que detalham o comportamento do Slapper¹;
- d) Bayley et al (2005), que explicam o comportamento do Blaster²;
- e) Moore et al (2003), que descrevem o Sapphire³;
- f) The Honeynet Project (2010a), que explica o que são bots, como eles podem infectar uma máquina, como se propagam, os tipos de ataques que podem ser executar além de sugerir uma metodologia para rastrear e observar botnets com o auxílio de honeypots; e,
- g) McCarty (2003) que descreve o uso de uma honeynet para detectar bots.

Sobre a coleta de dados na Internet tem-se o trabalho de Vanderavero (2004) que propõe um sistema de honeypots, chamado de HoneyTank, para coletar grande quantidade de informação de tráfego malicioso simulando a presença de sistemas em endereços IP não usados de uma rede.

Sobre o uso de honeypots para detectar e/ou auxiliar sistemas de detecção de intrusão tem-se:

-
- 1 Worm que se propaga em máquinas Linux usando uma falha da biblioteca OpenSSL.
 - 2 Worm que se propaga em computadores rodando sistemas operacional Microsoft explorando uma vulnerabilidade do tipo estouro de pilha (buffer overflow) do sistema operacional.
 - 3 Também chamado de Slammer foi o worm de mais rápida propagação na história. Ele conseguiu infectar 90% das máquinas vulneráveis em menos de 10 minutos. Explorou uma vulnerabilidade do tipo estouro de pilha (buffer overflow) em computadores rodando Microsoft SQL Server ou MSDE 2000.

- a) Yin et all (2004), que apresentam uma aplicação e um projeto de honeypot capaz de ser usado em colaboração com sistemas de detecção de intrusão para gerar um sistema capaz de detectar varreduras de portas;
- b) Levine et all (2003), que discutem formas de uso de honeynets com a finalidade de auxiliar o administrador de uma grande organização a identificar tráfego malicioso; e,
- c) Dagon et all (2004) apresentam um sistema local capaz de fornecer avisos precoces na detecção de worms, o HoneyStat, que usa honeypots modificados para gerar um fluxo de alertas preciso e com baixa taxa de falso-positivos.

2.1 Honeypot e Honeynet

Segundo Hoepers et all (2010), honeypot é um recurso computacional de segurança dedicado a ser sondado, atacado e comprometido. Eles podem ser de dois tipos:

- a) de baixa interatividade: emula sistemas operacionais e serviços com os quais os atacantes interagem; e,
- b) de alta interatividade: interação dos atacantes com sistemas operacionais, aplicações e serviços reais.

Ainda segundo Hoepers et all (2010), uma honeynet é uma ferramenta de pesquisa, que consiste de uma rede projetada especificamente para ser comprometida, e que contém mecanismos de controle para prevenir que seja utilizada como base de ataques contra outras redes. Também é conhecida por honeypot de pesquisa.

Honeynets são conceitualmente muito simples; são uma rede com um ou mais honeypots. Como honeypots não possuem sistemas de produção uma honeynet não possui atividade de produção e/ou serviço autorizado. Logo, toda interação com uma honeynet implica em atividade maliciosa ou não autorizada.

2.2 Ruído de fundo

Segundo Pang et al (2004), o monitoramento de qualquer endereço IP na Internet revela uma atividade incessante. Esta atividade ocorre mesmo quando se está monitorando endereços que não estão em uso. A este tráfego dá-se o nome de radiação ou ruído de fundo. Este ruído é constituído fundamentalmente de tráfego não-produtivo seja ele maligno (varreduras, worms, ...) ou benigno (má configuração, ...).

Segundo The Honeypot Project (2010a) o mecanismo de propagação usado pelos bots é um dos principais causadores do ruído de fundo da Internet uma vez que eles varrem grandes espaços de endereçamento procurando computadores vulneráveis agindo de forma similar a um worm ou um vírus.

2.2.1 Redes telescópio

Segundo Savage et al (2010), o grande desafio para analisar fenômenos mundiais, como os worms, é adquirir dados em quantidade suficiente para garantir a representatividade do fenômeno sendo observado.

Ao se monitorar um único local pode-se perder os estágios iniciais da propagação de um ataque. É necessário, portanto, de uma visão mais global que leve a entender o comportamento dos fenômenos sendo observados em escala global (quantas máquinas foram infectadas, qual a velocidade de propagação, quais as alterações no seu comportamento, etc.).

Para diminuir o tempo de detecção de um ataque e minimizar os erros causados por sinais que parecem ser ataques mas que são legítimos, deve-se investir no aumento do monitoramento e consequente coleta de mais dados. Os recursos estatísticos necessários para prever ataques são mais confiáveis à medida que o tamanho da população aumenta.

Uma ferramenta que fornece dados em quantidade suficiente para a análise estatística é a rede telescópio. Elas monitoram o tráfego enviado para porções não alocadas do endereçamento IP. Uma vez que não há nenhum motivo legítimo para que uma máquina en-

vie pacotes para estes endereços todo o tráfego nestas redes sugere a ocorrência de atividades maliciosas.

Segundo Pang et al (2004), a maioria dos estudos que envolvem a coleta de dados maliciosos ou de ruído de fundo da Internet têm sido realizadas em redes telescópio. O espaço dos endereços IP não usados é uma importante fonte de informação sobre intrusão e atividades de ataque. Sistemas de medida que usam os endereços IP ainda não atribuídos são chamados de *sink-holes* ou de redes telescópio.

2.2.2 Honeyfarms

Redes telescópio são entidades totalmente passivas. Mesmo que elas consigam observar um aumento de atividade elas não permitam emitir uma resposta a um evento maligno detectado.

Segundo Savage et al (2010), a combinação da capacidade de detecção em larga escala da rede telescópio com a capacidade de resposta dos honeypots permite a obtenção de grande quantidade de informação porém, com melhor qualidade uma vez que, como há possibilidade de interação com o usuário, mais dados podem ser coletados.

O uso de honeypots no espaço de endereçamento das redes telescópio formam um sistema que, coletivamente, é chamado de *honeyfarm*.

2.2.3 CBH

Ao se pensar em escala global os honeypots não são as ferramentas mais adequadas para monitoramento e obtenção de dados uma vez que são recursos individuais.

Supera-se esta deficiência com a utilização de um grande número de honeypots. O que difere a solução do CBH das *honeyfarms* é o espaço de endereçamento sendo monitorado. O Consórcio Brasileiro de Honeypots monitora o espaço de endereçamento válido da Internet.

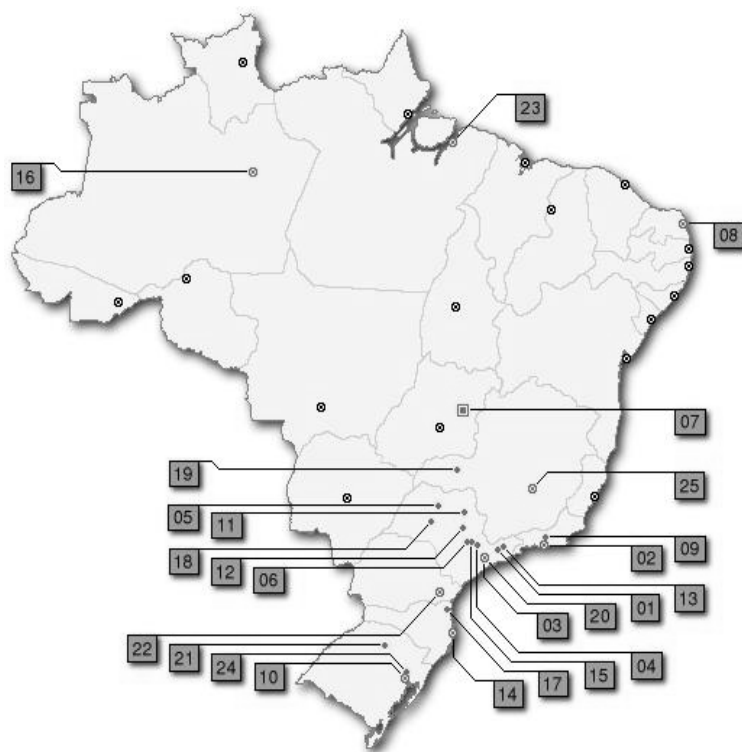


Figura 2.1 - Localização dos Honeypots
 Fonte: Adaptada de CBH (2010).

Segundo CBH (2010), o CBH tem o objetivo de aumentar a capacidade de detecção de incidentes, correlação de eventos e determinação de tendências de ataques no espaço Internet brasileiro. Em 2007 os 44 sensores componentes do CBH estavam distribuídos ao longo do Brasil conforme a Figura 2.1.

Cada uma das localidades apresentadas na Figura 2.1 está listada na Tabela 2.1 com seus respectivos sensores:

Tabela 2.1 - Sensores do CBH e suas localizações

	Cidade	Instituições/Sensores
01	São José dos Campos	INPE, ITA
02	Rio de Janeiro	CBPF, CBPF 2, EBT, Fiocruz, IME, PUC-RIO, RedeRio, RedeRio 2, UFRJ
03	São Paulo	ANSP, CERT.br, Diveo, Durand, UNESP, USP
04	Campinas	CTI, ITAL, UNICAMP
05	São José do Rio Preto	UNESP
06	Piracicaba	USP
07	Brasília	Brasil Telecom, Ministério da Justiça, TCU
08	Natal	UFRN
09	Petrópolis	LNCC
10	Porto Alegre	CERT-RS
11	Ribeirão Preto	USP
12	São Carlos	USP
13	Taubaté	UNITAU
14	Florianópolis	UFSC DAS
15	Americana	VIVAX
16	Manaus	VIVAX
17	Joinville	UDESC
18	Lins	FPTE
19	Uberlândia	CTBC Telecom
20	Santo André	VIVAX
21	Passo Fundo	UPF
22	Curitiba	PoP-PR
23	Belém	UFPA
24	São Leopoldo	Unisinos
25	Belo Horizonte	Diveo

Fonte: Adaptada de CBH (2010).

2.3 Análise de séries temporais

Segundo Statsoft (2010), a análise de séries temporais possui dois grandes objetivos:

- a) identificar a natureza do fenômeno representado pela sequência de observações;

e,

- b) prognosticar (predizer valores futuros da variável da série temporal)¹.

Para que seja possível tanto a identificação da natureza do fenômeno como para prognosticar é necessário que se tente encontrar um modelo matemático que faça parecer que a série é composta de uma parte determinística mais um ruído. Para tal procura-se extrair as estruturas existentes na série temporal, os dados dependentes e transforma-se a série numa sucessão de valores independentes.

A parcela do ruído desejado, não é qualquer. Espera-se que constitua o ruído branco, isto é, aquele composto de uma série de variáveis aleatórias não correlacionadas com expectativa nula e variância constante.

Pode-se, portanto, inferir que o objetivo da análise de séries temporais é transformar os dados de tal forma que o resultado final das transformações seja uma nova série cujos resíduos pareçam gaussianos.

A análise de séries temporais comporta 4 etapas:

- a) a transformação do conjunto de dados, eventual;
- b) a estimação do componente de tendência (μ_t);
- c) a estimação do componente de sazonalidade (S_t);
- d) a análise dos resíduos $e_t = x_t - \mu_t - S_t$; onde x_t é o valor observado:
 - se $\{e_1, e_2, \dots, e_n\}$ é um ruído branco, a análise termina;

¹ Os termos prognosticar (*forecast*), predição (*prediction*), projeção (*projection*), prognóstico (*prognosis*) e previsão são usados indistintamente neste trabalho com o mesmo significado.

- caso contrário faz-se necessário propor modelos mais sofisticados para estimação dos parâmetros.

Viinikka et al (2004) usa uma técnica de controle, passível de ser empregada em filtros de séries temporais, para monitorar o ruído de fundo gerado pelo sistema de detecção de intrusão visando minimizar o número de falsos positivos. Já Viinikka et al (2006) usa técnicas de análise de séries temporais para filtrar os alertas oriundos do sistema de detecção de intrusão com o mesmo objetivo.

2.3.1 Teoria de previsão

Segundo Jensen (2010), a análise de séries temporais fornece ferramentas para selecionar um modelo que descreve a série temporal e que permite prever eventos futuros.

A análise de uma série temporal objetiva determinar o melhor modelo para uma situação em particular. Uma vez selecionado o modelo e tendo os dados à disposição, o problema se resume a achar parâmetros que melhor ajuste o modelo aos dados históricos.

Portanto, modelar uma série temporal é um problema estatístico. Os dados observados são computacionalmente tratados para estimar coeficientes de um modelo matemático que descreve os dados observados e que permite extrapolação para tempos futuros fornecendo uma previsão.

A realização de previsões de eventos futuros tem sido empregada em vários campos de atuação. Entretanto, em computação e, mais particularmente, na área de segurança computacional, foi encontrado somente um resultado em The Honeynet Project (2010b).

3 DADOS DO CBH

A estrutura de gerência do CBH faz a transmissão dos dados coletados nos sensores a cada 24 horas para o Centro de Tecnologia da Informação Renato Archer (CTI). Segundo Barros (2010b), cada sensor transmite dois arquivos com dados coletados das 00:00 até as 23:59: um completo com todo o tráfego incluindo o *payload*; e, um resumido, com um resumo do tráfego. Para preservar os sensores estes arquivos são sanitizados antes de serem utilizados, isto é, os endereços IP dos sensores são substituídos por outros ou simplesmente apagados.

Após a transmissão os arquivos são armazenados no diretório correspondente ao nome do sensor que coletou os dados. Dentro de cada diretório há subdiretórios. Cada um corresponde a um dia. Dentro de cada sub-diretório há um ou mais arquivos comprimidos nomeados como *honeyd.log.ano.mes.dia:hora_minuto*.

Pode ocorrer que os registros que correspondem a um dia de coleta estejam distribuídos em mais de um arquivo. O complemento *hora:minuto* é usado para recompor os eventos na ordem em que ocorreram.

Um registro do arquivo sanitizado tem uma das seguintes estruturas:

- se protocolo de transporte TCP ou UDP:

data hora protocolo [E|S|-] IP_Origem Porta_Origem IP_Destino
Porta_Destino: Bytes [Flags|0] [S.O.]

- se protocolo ICMP:

data hora icmp(1) - IP_Origem IP_Destino type(code)

Registros do protocolo de transporte TCP são de três diferentes tipos:

- a) o que indica uma tentativa de conexão onde 1, ou 2 pacotes, são trocados. É indi-

cado por um “-”;

b) o que indica uma conexão estabelecida. É indicado por “S”; e,

c) o que indica o término de uma conexão estabelecida. É indicado por “E”.

Registros do protocolo de transporte UDP, embora o protocolo não seja orientado à conexão, são dos mesmos tipos do protocolo TCP. Para que isto seja possível é necessário abstrair o conceito de orientação à conexão e interpretar uma comunicação entre um endereço IP e uma porta de origem com um mesmo endereço IP e mesma porta de destino como conexão. A existência de um pacote sem resposta caracterizaria o “-”.

Foi necessário, antes da análise dos dados, realizar um processo de sanitização dos mesmos, aqui chamado de processo de pré-processamento dos dados, para retirada de inconsistências nos arquivos causados por problemas de transmissão, repetições não desejadas de registros, dentre outras.

3.1 Pré-processamento dos dados

Os arquivos sanitizados são retirados dos sub-diretórios e colocados no diretório com o nome do sensor. Os arquivos são descompactados e, se houver mais de um arquivo para um mesmo dia, eles são unidos num único arquivo na ordem correta de ocorrência dos eventos registrados.

Durante este processo linhas de comentários são retiradas. Protocolos que não TCP, UDP ou ICMP também são retirados. Registros com o flag “S”, para os protocolos TCP ou UDP, também o são; entende-se que a ocorrência de um registro “S” implica, necessariamente, na ocorrência de um registro “E” tornando-os redundantes.

Alguns sensores podem estar configurados para interagir com um atacante de um determinado serviço. Neste caso pode ocorrer o registro de tráfego de saída com origem no sensor. Dos dados recebidos verificou-se a existência dos tráfegos listados na Tabela 3.1.

Tabela 3.1 - Registros de tráfego encontrados nos arquivos sanitizados do CBH

	Direção	IP Origem	IP Destino	Comentários/Exemplos
A	Entrada	CBH	CBH	Mensagens icmp do serviço Ping
B	Entrada	ñCBH	CBH	Tráfego esperado em sensor passivo
C	Saída	ñCBH	CBH	Ocorre se máquina CBH estiver comprometida
D	Saída	ñCBH	ñCBH	Ocorre se máquina CBH estiver comprometida
E	Saída	CBH	CBH	Mensagens opencap
F	Saída	CBH	ñCBH	Mensagens udp do serviço dns

Não se consegue, usando somente os dados sanitizados, distinguir entre os tráfegos A e E, e os de B e C. Sobram, portanto, 4 diferentes tipos de tráfegos.

Realiza-se a separação dos registros do arquivo sanitizado original colocando cada um dos tráfegos apresentados na Tabela 3.1 em quatro diretórios, a saber:

- a) oCBHdCBH: com o tráfego A ou E, isto é, tráfego com origem em máquinas com um endereço IP de um sensor do CBH destinado a máquinas com endereço IP do CBH;
- b) oCBHdNCBH: com o tráfego de F, isto é, tráfego com origem em máquinas com endereço IP de um sensor do CBH destinado a máquinas com endereço IP que não seja do CBH;
- c) oNCBHdCBH: com o tráfego de B ou C, isto é, tráfego com origem em máquinas com endereço IP que não sejam do CBH destinado a máquinas com endereço IP de um sensor do CBH; e,
- d) oNCBHdNCBH: com o tráfego de D, isto é, tráfego com origem em máquinas com endereço IP que não seja do CBH destinado a máquinas com endereço IP que também não sejam do CBH.

Além de criar estes diretórios e selecionar o conteúdo de acordo com o tipo de tráfego o registro em cada arquivo também é alterado e passa a ser:

- oCBHdCBH:

data proto flag IP_Orig Prt_Dest (ou Serv se proto = ICMP)

- oCBHdNCBH:

data proto flag IP_Dest Prt_Dest (ou Serv se proto = ICMP)

- oNCBHdCBH:

data proto flag IP_Orig Prt_Dest (ou Serv se proto = ICMP)

- oNCBHdNCBH:

data proto flag IP_Orig Prt_Orig IP_Dest Prt_Dest (TCP/UDP)

data proto flag IP_Orig IP_Dest Serv (ICMP)

Esta tese enfatiza o uso dos sensores do CBH como forma de obtenção de dados para geração de alertas precoces de eventos que estão ocorrendo na Internet, isto é, eventos cuja origem sejam máquinas com endereços IP não pertencentes ao CBH destinado a máquinas que pertençam ao CBH, isto é, que sejam sensores. Logo, os registros a serem usados são os classificados em oNCBHdCBH.

Uma vez classificados é possível agrupar os registros de acordo com uma janela de amostragem. Para este trabalho uma janela de amostragem corresponde ao intervalo de tempo no qual se considera uma unidade de apresentação da informação.

É utilizada a janela de amostragem em razão das dificuldades de ordem prática para se realizar processamentos em tempo real ou quase. Para exemplificar a dificuldade, já foi dito no início deste Capítulo que os dados dos sensores são transmitidos somente uma vez ao dia. Só esta característica já inviabiliza toda e qualquer forma de tratamento em

tempo real.

A janela de amostragem tem de ser tal que permita a realização da captura da informação, sua transmissão, processamento e geração de previsões dentro de um intervalo de tempo útil.

Esta tese utilizou diferentes janelas de amostragem: ano, mês, dia e hora. Para geração de alertas a janela de amostragem hora foi a que apresentou melhores resultados para a geração de alertas precoces.

Este processamento é realizado de forma automática e o resultado obtido é um arquivo resumo com o número de registros, de cada sensor, para cada intervalo de amostragem considerado. Além deste arquivo outro é gerado onde se consolida, baseado no *flag* existente em cada registro, o número de pacotes TCP que contenham “-”, os TCP que contenham “E”, os UDP com “-”, os UDP com “E” e ICMP.

3.2 Fluxo

Segundo Barros (2008), não é possível, a partir dos dados contidos nos arquivos sanitizados do CBH, identificar quantos pacotes foram trocados entre duas máquinas comunicantes. Assim, neste trabalho, será usado o conceito de fluxo como substituto de tráfego e de pacotes com o seguinte significado: fluxo é um conjunto de 1, 2, ..., n pacotes trocados por duas máquinas dentro do contexto de uma sessão de comunicação.

Um fluxo com um ou dois pacotes representa, geralmente, um tráfego de varredura (pacotes errados e mal configurados são de baixa frequência e estão sendo considerados como parte do tráfego de varredura). Um fluxo com três ou mais pacotes representa uma conexão (o termo conexão não é adequado para comunicações com o protocolo de transporte UDP mas foi utilizado significando tráfego/comunicação entre duas máquinas).

3.3 Dados utilizados

Foram obtidos os dados dos sensores listados na Tabela 2.1 entre 01/01/2005 e 30/06/2006. Neste período os sensores das cidades de Natal e de São Leopoldo ainda

não estavam operacionais e foram desconsiderados. Além dos sensores listados mais dois estavam funcionando à época da coleta: HP-PSC e Unicamp-Feec. Embora tenham contribuído para a pesquisa eles não mais integram o CBH.

Os dados obtidos têm as seguintes características gerais:

- a) Número total de dias de coleta: 546;
- b) Número de dias sem dados: 26;
 - Os dias 18/02/2005, 19/02/2005, 20/02/2005, 15/03/2005, 09/07/2005, 10/07/2005, 15/07/2005, 16/07/2005, 17/07/2005, 23/10/2005, 13/01/2006, 14/01/2006, 15/01/2006, 23/01/2006, 24/01/2006, 26/01/2006, 28/02/2006, 09/03/2006, 20/03/2006, 20/04/2006, 21/04/2006, 22/04/2006, 23/04/2006, 05/05/2006, 06/05/2006, 07/05/2006 não apresentaram dados para nenhum sensor. Provavelmente devido a alguma falha de sincronismo na hora da transferência dos dados.
- c) Número efetivo de dias com dados coletados: 520.

Cada sensor coletou dados em um número diferente de dias. O número de dias que cada sensor coletou dados, tanto para o ano de 2005 como para o ano de 2006 encontra-se listado na Tabela 3.2.

Tabela 3.2 - Quantidade de dias que cada sensor do CBH forneceu dados, por ano

	Dias				Dias		
Sensor	2005	2006	Total	Sensor	2005	2006	Total
ansp	283	83	366	puc-rio	350	165	515
brasiltelecom	296	160	456	rederio	349	139	488
cbpf-2	340	138	478	rederio-2	93	165	258
cbpf	353	48	401	tcu	299	153	452
cenpra	349	148	497	udesc	180	110	290
cert-br	355	165	520	ufpa	0	45	45
cert-rs	335	149	484	ufrj	350	144	494
ctbc	119	165	284	ufsc-das	254	163	417
diveo	355	133	488	unb-labredes	245	141	386
diveo-2	350	165	515	unesp	349	166	515
durand-1	320	164	484	unesp-sjrp	328	142	470
ebt-rjo1	6	160	166	unicamp	345	159	504
fiocruz	316	161	477	unicamp-feec	96	79	175
fpte	95	163	258	unitau	278	160	438
hp-psc	341	112	453	upf	5	155	160
ime	137	126	263	usp	350	165	515
inpe	335	161	496	usp-ciagri	348	162	510
ita	347	158	505	usp-cirp	295	6	301
ital	85	163	248	usp-cisc	297	157	454
lncc-1	285	153	438	vivax-amr	230	164	394
mj	218	130	348	vivax-mns	180	139	319
pop-pr	0	111	111	vivax-san	0	24	24

O que a Tabela 3.2 traduz é que há necessidade de alguma forma de normalização dos dados entre os sensores uma vez que os sensores não forneceram a mesma quantidade de dados.

3.4 Situações especiais

No resumo apresentado na Seção 3.3 tem-se que 26 dias não apresentaram dados. Porém, o processamento posterior não permite que se informem dados nulos.

Portanto, aplicou-se uma formulação de média ponderada para estimar o valor de um dado não presente baseado nos dados passados. A formulação usada é apresentada na Equação 3.1.

$$x_t = 0,65 * x_{t-1} + 0,25 * x_{t-2} + 0,10 * x_{t-3} \quad (3.1)$$

onde,

x_t representa o valor de um fluxo num instante t , e

x_{t-i} representa o valor de um fluxo num instante $t - i$.

Os coeficientes utilizados foram arbitrados de tal forma que dados mais recentes tenham mais influência sobre o valor sendo estimado do que dados mais antigos. Desta forma arbitrou-se que o dado mais recente teria 65% de influência, o próximo 25% e o último 10%.

Os valores calculados para estes dias quando usadas as janelas de amostragem ano, mês e hora têm de coincidir com os valores determinados para a janela de amostragem dia.

Além do número de dias que cada sensor forneceu dados outro aspecto a ser considerado é a possibilidade de um sensor estar apresentando no seu conjunto de dados observados não o comportamento do fluxo de dados na parcela brasileira da Internet mas sim uma característica particular. Considere a Tabela 3.3 que apresenta a média de fluxos TCP, UDP e ICMP para todos os sensores e os valores coletados para um sensor.

Tabela 3.3 - Percentuais de fluxos dos diferentes protocolos de transporte em relação ao fluxo total

	TCP	UDP	ICMP
Todos os Sensores	90,5	6,7	2,8
Sensor ANSP	7,4	92,2	0,4

Verifica-se que o sensor ANSP possui um fluxo UDP consideravelmente maior do que

os demais. Embora destoe da média é um comportamento justificável para este sensor uma vez que ele está colocado dentro de uma entidade responsável por responder consultas de DNS no Brasil. Porém, não é útil para este trabalho que quer representar o comportamento médio dos fluxos.

Estes problemas conduziram ao estabelecimento de uma métrica para eliminar o conjunto de sensores cujos dados coletados se desviem muito do comportamento médio.

3.5 Seleção da amostra de trabalho

Este trabalho analisou 3 (três) métricas:

- a) número de dias que cada sensor forneceu dados;
- b) relação entre o número de fluxos TCP, UDP e ICMP; e,
- c) relação entre número de fluxos TCP com menos de três pacotes e fluxos TCP com três ou mais pacotes.

3.5.1 Métrica: número de dias

Para cada métrica é estipulada uma pontuação que penaliza o sensor que tem a tendência a ser diferente da média. No final, os sensores que tiverem menos pontos serão os utilizados, ou seja, são considerados os que melhor representam, na média, o comportamento esperado do ruído de fundo na parcela brasileira da Internet.

O critério de número de dias considera o número de dias que cada sensor coletou dados para os anos de 2005 e de 2006 isoladamente (pode ser que um sensor tenha entrado em atividade somente em 2006 não sendo uma boa referência para 2005 mas sendo muito boa para 2006).

Como métrica estabeleceu-se um limite baseado na média, no desvio padrão e em uma constante arbitrária conforme a Equação 3.2.

$$Limite = \mu \pm k * \sigma \quad (3.2)$$

onde,

μ é a média,

σ o desvio padrão; e,

k uma constante arbitrária.

Para o ano de 2005 o limite determinado foi de 125 dias e para o ano de 2006 o limite foi de 96 dias. A aplicação desta métrica encontra-se apresentada na Tabela 3.4.

Tabela 3.4 - Primeiro critério de seleção de amostra: número de dias com fluxo

Sensor	Dias 2005	Dias 2006	Sensor	Dias 2005	Dias 2006
ansp	283	83	puc-rio	350	165
brasiltelecom	296	160	rederio	349	139
cbpf-2	340	138	rederio-2	93	165
cbpf	353	48	tcu	299	153
cenpra	349	148	udesc	180	110
cert-br	355	165	ufpa	0	45
cert-rs	335	149	ufrj	350	144
ctbc	119	165	ufsc-das	254	163
diveo	355	133	unb-labredes	245	141
diveo-2	350	165	unesp	349	166
durand-1	320	164	unesp-sjrp	328	142
ebt-rj01	6	160	unicamp	345	159
fiocruz	316	161	unicamp-feec	96	79
fpte	95	163	unitau	278	160
hp-psc	341	112	upf	5	155
ime	137	126	usp	350	165
inpe	335	161	usp-ciagri	348	162
ita	347	158	usp-cirp	295	5
ital	85	163	usp-cisc	297	157
lncc-1	285	153	vivax-amr	230	164
mj	218	130	vivax-mns	180	138
pop-pr	0	111	vivax-san	0	24

Destaca-se na Tabela 3.4 os sensores que tiveram dados coletados abaixo do limite determinado. Cada destaque corresponde à penalização de 1 (um) ponto. Assim, o sensor BRASILTELECOM terá 0 pontos; ANSP terá 1 ponto; e, VIVAX-SAN terá 2 pontos.

3.5.2 Métrica: relação entre fluxos de protocolos

Se somados todos os fluxos TCP, UDP e ICMP, sem nenhum critério de seleção ou aplicação de métricas, verifica-se que aproximadamente 90% do fluxo total são fluxos TCP, aproximadamente 7% são fluxos UDP e aproximadamente 3% são fluxos ICMP.

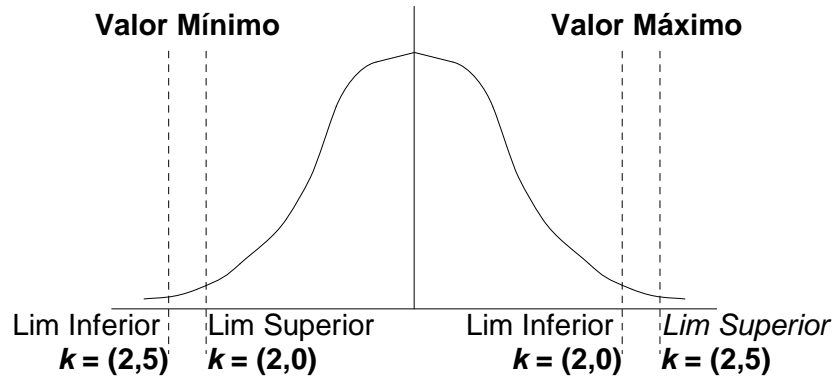


Figura 3.1 - Curva de distribuição gaussiana com os limites usados na segunda métrica para seleção da amostra de trabalho

Estes valores, a princípio, estão sendo adotados como a referência esperada de relação entre fluxos em cada sensor.

A métrica adotada determina quatro limites. Para um valor máximo determinam-se um limite superior e um inferior e para um valor mínimo outros limites superior e inferior.

Este critério se baseia no teorema de Chebyshev que diz que sendo X uma variável aleatória com média μ e variância σ^2 , para qualquer número real $k > 0$ a relação da Equação 3.3 é válida.

$$Pr(|X - \mu| \geq k \sigma) \leq \frac{1}{k^2} \quad (3.3)$$

A desigualdade de Chebyshev diz que em qualquer amostra de dados, ou em qualquer distribuição probabilística, quase todos os valores estão próximos do valor médio. Para determinar o quão afastado os dados que serão aceitos podem estar afastados da média é utilizada a Equação 3.3 com diferentes valores de k para determinar os limites máximo e mínimo de acordo com o apresentado na Figura 3.1.

A métrica privilegia, isto é, não são atribuídos pontos, os sensores cujos dados coletados se situam entre o limite superior do valor mínimo e o limite inferior do valor máximo. Para os sensores que tenham dados que se situam entre os limites inferior e superior é

atribuído 0,5 pontos. Para os sensores que tenham dados coletados que se situam acima do limite superior do valor máximo ou abaixo do limite inferior do valor mínimo, 1 ponto.

A métrica é aplicada de forma independente para os dados coletados nos anos de 2005 e de 2006 para cada os protocolos TCP, UDP e ICMP. Os percentuais adotados para aplicação da métrica consideram o fluxo por protocolo em relação ao fluxo total por sensor. Exemplificando, o sensor ANSP coletou 223.835 fluxos TCP, 2.672.226 fluxos UDP e 11.680 fluxos ICMP no ano de 2005 totalizando 2.907.741 fluxos. O percentual a ser aplicado na métrica é de 7,70% para TCP, 91,90% para UDP e 0,40% para ICMP.

Portanto, para cada protocolo, ter-se-ia de realizar duas comparações por ano: uma em relação ao valor máximo e outra em relação ao valor mínimo. Entretanto, a prática mostrou que para o protocolo TCP a distribuição das observações permanece abaixo do limite inferior do valor máximo sendo necessária somente a comparação com o valor mínimo. Para os protocolos UDP e ICMP os valores observados se localizam acima do limite superior do valor mínimo sendo necessário compará-los somente com o valor máximo.

Na Tabela 3.5 encontra-se listado os sensores ordenados em ordem decrescente de percentual de fluxo TCP. Na coluna indicada por μ encontra-se a média calculada entre linhas, isto é, a média do primeiro sensor é o próprio valor do percentual de fluxo; na segunda linha a média corresponde à média aritmética dos valores percentuais das primeira e segunda linhas; e assim por diante. Na coluna indicada por σ encontra-se o desvio padrão determinado por linha, tal como foi feito para a média.

Tabela 3.5 - Segunda métrica: fluxo TCP em 2005. Comparação com valor mínimo

Sensor	TCP	μ	σ	Limites		Sensor	TCP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
unb-labredes	98,0	98,0	0,0	98,0	98,0	hp-psc	97,8	97,9	0,2	97,5	97,4
brasiltelecom	97,7	97,8	0,2	97,5	97,4	rederio	97,1	97,7	0,4	96,9	96,7
unesp	89,0	93,4	2,9	87,6	86,2	usp-cirp	87,2	93,2	3,1	87,0	85,4
cert-br	96,9	97,5	0,5	96,6	96,3	ctbc	95,9	97,2	0,8	95,7	95,3
cbpf-2	95,7	97,0	0,9	95,2	94,7	cenpra	95,3	96,8	1,0	94,7	94,2
diveo	94,9	96,6	1,2	94,3	93,7	ita	94,2	96,4	1,3	93,7	93,0
cbpf	94,1	96,1	1,4	93,3	92,6	ebt-rjo1	93,2	95,9	1,6	92,7	91,9
unitau	92,9	95,7	1,8	92,2	91,3	inpe	92,5	95,4	1,9	91,7	90,7
mj	92,2	95,2	2,0	91,2	90,2	unesp-sjrp	91,5	95,0	2,1	90,7	89,6
cert-rs	91,2	94,8	2,3	90,2	89,1	usp	90,8	94,6	2,4	89,8	88,6
rederio-2	90,7	94,3	2,5	89,4	88,1	fiocruz	90,4	94,1	2,6	89,0	87,7
unicamp-feec	90,1	94,0	2,7	88,6	87,3	vivax-mns	90,0	93,8	2,7	88,3	86,9
ufrj	89,7	93,6	2,8	88,0	86,6	durand-1	86,8	92,9	3,3	86,4	84,7
udesc	86,2	92,7	3,5	85,7	84,0	ufsc-das	84,9	92,4	3,7	85,0	83,1
puc-rio	84,0	92,1	4,0	84,2	82,2	upf	81,2	91,7	4,4	83,0	80,8
fpte	80,4	91,4	4,8	81,9	79,5	diveo-2	76,8	90,9	5,3	80,2	77,6
tcu	75,0	90,4	5,9	78,6	75,6	vivax-amr	74,3	90,0	6,5	77,0	73,8
usp-ciagri	73,4	89,5	7,0	75,6	72,1	ime	66,9	88,9	7,8	73,2	69,3
ital	66,8	88,3	8,5	71,2	66,9	unicamp	56,1	87,4	9,9	67,6	62,7
lncc-1	52,4	86,5	11,3	64,0	58,4	usp-cisc	47,8	85,6	12,7	60,2	53,8
ansp	7,7	83,7	17,5	48,7	40,0	vivax-san	0,0	81,7	21,5	38,6	27,8
ufpa	0,0	79,8	24,7	30,4	18,1	pop-pr	0,0	77,9	27,2	23,6	10,0

As colunas Limite Superior e Inferior representam os valores determinados pela Equação 3.2 para os valores de k iguais a 2,0 e 2,5, respectivamente, em relação ao valor mínimo.

Um destaque na coluna Limite Superior na Tabela 3.5 indica que o percentual está abaixo deste valor mas não abaixo do Limite Inferior, o que penaliza o sensor com 0,5 pontos. Um destaque na coluna Limite Inferior indica que o percentual do sensor está abaixo deste valor o que penaliza o sensor com 1,0 ponto.

As colunas Limite Superior e Inferior representam os valores determinados pela Equação 3.2 para os valores de k iguais a 2,5 e 2,0, respectivamente, em relação ao valor máximo.

Um destaque na coluna Limite Inferior na Tabela 3.6 indica que o percentual está acima deste valor mas não acima do Limite Superior, o que penaliza o sensor com 0,5 pontos. Um destaque na coluna Limite Superior indica que o percentual do sensor está acima deste valor o que penaliza o sensor com 1,0 ponto.

Tabela 3.6 - Segunda métrica: fluxo UDP em 2005. Comparação com o valor máximo

Sensor	UDP	μ	σ	Limites		Sensor	UDP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
pop-pr	0,0	0,0	0,0	0,0	0,0	ufpa	0,0	0,0	0,0	0,0	0,0
vivax-san	0,1	0,0	0,0	0,0	0,0	rederio	0,9	0,2	0,5	1,4	1,2
cbpf-2	1,0	0,4	0,5	1,7	1,5	cbpf	1,1	0,5	0,6	1,9	1,6
brasiltelecom	1,4	0,6	0,6	2,1	1,8	unb-labredes	1,7	0,8	0,7	2,4	2,1
hp-psc	1,9	0,9	0,7	2,7	2,3	ctbc	2,2	1,0	0,8	3,0	2,6
cert-br	2,4	1,1	0,9	3,3	2,9	ita	3,5	1,3	1,1	4,0	3,5
cert-rs	3,6	1,5	1,2	4,5	3,9	unesp-sjrp	3,8	1,7	1,3	4,9	4,3
diveo	4,0	1,8	1,4	5,3	4,6	cenpra	4,1	2,0	1,5	5,6	4,9
unitau	4,7	2,1	1,6	6,0	5,3	rederio-2	4,8	2,3	1,6	6,4	5,6
fiocruz	5,5	2,5	1,8	6,8	6,0	inpe	5,9	2,6	1,9	7,3	6,4
ebt-rjo1	6,2	2,8	2,0	7,8	6,8	mj	6,7	3,0	2,1	8,3	7,2
diveo-2	7,8	3,2	2,3	8,9	7,8	unesp	8,3	3,4	2,5	9,6	8,3
ufrj	8,8	3,6	2,7	10,2	8,9	usp	8,8	3,8	2,8	10,8	9,4
unicamp-feec	9,0	4,0	2,9	11,3	9,8	udesc	9,2	4,2	3,0	11,7	10,2
vivax-mns	9,6	4,4	3,1	12,2	10,6	usp-cirp	9,8	4,6	3,2	12,7	11,0
durand-1	10,1	4,7	3,3	13,1	11,4	puc-rio	12,8	5,0	3,6	13,9	12,1
ufsc-das	14,8	5,3	3,9	15,1	13,1	upf	17,3	5,6	4,4	16,6	14,4
fpte	19,4	6,0	4,9	18,3	15,8	ime	20,5	6,4	5,4	19,9	17,2
tcu	24,5	6,9	6,1	22,1	19,1	vivax-amr	25,3	7,4	6,7	24,2	20,8
usp-ciagri	25,6	7,9	7,2	25,9	22,3	usp-cisc	29,6	8,4	7,9	28,2	24,2
lncc-1	30,4	9,0	8,5	30,3	26,0	ital	31,0	9,5	9,1	32,2	27,7
unicamp	40,9	10,2	10,2	35,7	30,6	ansp	91,9	12,1	15,9	51,8	43,9

A Tabela 3.7 possui interpretação semelhante à dada para a Tabela 3.6.

Tabela 3.7 - Segunda métrica: fluxo ICMP em 2005. Comparação com valor máximo

Sensor	ICMP	μ	σ	Limites		Sensor	ICMP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
pop-pr	0,0	0,0	0,0	0,0	0,0	ufpa	0,0	0,0	0,0	0,0	0,0
vivax-san	0,0	0,0	0,0	0,0	0,0	fpte	0,2	0,1	0,1	0,4	0,3
ufsc-das	0,3	0,1	0,1	0,5	0,4	unb-labredes	0,3	0,1	0,1	0,5	0,4
hp-psc	0,4	0,2	0,2	0,6	0,5	usp	0,4	0,2	0,2	0,6	0,5
vivax-mns	0,4	0,2	0,2	0,6	0,5	vivax-amr	0,4	0,2	0,2	0,7	0,6
ansp	0,4	0,2	0,2	0,7	0,6	tcu	0,5	0,3	0,2	0,7	0,6
cenpra	0,6	0,3	0,2	0,8	0,7	ebt-rjo1	0,6	0,3	0,2	0,8	0,7
cert-br	0,7	0,3	0,2	0,9	0,8	unicamp-feec	0,9	0,4	0,2	1,0	0,9
brasiltelecom	0,9	0,4	0,3	1,1	1,0	diveo	1,1	0,4	0,3	1,2	1,1
usp-ciagri	1,1	0,5	0,3	1,3	1,1	mj	1,1	0,5	0,4	1,4	1,2
ufrj	1,5	0,6	0,4	1,6	1,4	upf	1,5	0,6	0,4	1,7	1,5
inpe	1,6	0,6	0,5	1,9	1,6	ctbc	1,9	0,7	0,5	2,0	1,8
rederio	2,0	0,7	0,6	2,2	1,9	ital	2,2	0,8	0,6	2,4	2,1
ita	2,2	0,9	0,7	2,6	2,2	unitau	2,5	0,9	0,7	2,8	2,4
unesp	2,7	1,0	0,8	3,0	2,6	usp-cirp	3,0	1,0	0,9	3,2	2,8
unicamp	3,0	1,1	0,9	3,4	3,0	durand-1	3,1	1,2	1,0	3,6	3,1
puc-rio	3,2	1,2	1,0	3,8	3,3	cbpf-2	3,3	1,3	1,1	3,9	3,4
fiocruz	4,2	1,4	1,2	4,3	3,7	rederio-2	4,5	1,5	1,3	4,6	4,0
udesc	4,6	1,5	1,3	4,9	4,2	unesp-sjrp	4,7	1,6	1,4	5,2	4,5
cbpf	4,9	1,7	1,5	5,4	4,7	cert-rs	5,2	1,8	1,6	5,7	4,9
ime	12,6	2,1	2,3	7,8	6,7	diveo-2	15,4	2,4	3,1	10,0	8,5
lncc-1	17,2	2,7	3,8	12,2	10,3	usp-cisc	22,5	3,2	4,8	15,1	12,7

As próximas tabelas, 3.8, 3.9 e 3.10 apresentam as mesmas situações já vistas nas tabelas anteriores só que para o ano de 2006.

Tabela 3.8 - Segunda métrica: fluxo TCP em 2006. Comparação com valor mínimo

Sensor	TCP	μ	σ	Limites		Sensor	TCP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
unb-labredes	99,1	99,1	0,0	99,1	99,1	ebt-rjo1	97,8	98,5	0,9	96,6	96,1
ufsc-das	97,7	98,2	0,8	96,7	96,3	brasiltelecom	97,3	98,0	0,8	96,4	96,0
hp-psc	97,0	97,8	0,8	96,2	95,8	cert-br	96,4	97,6	0,9	95,7	95,3
rederio	96,3	97,4	1,0	95,5	95,0	diveo	96,2	97,2	1,0	95,3	94,8
cenpra	95,7	97,1	1,0	95,0	94,4	pop-pr	94,9	96,8	1,2	94,4	93,8
cert-rs	93,6	96,6	1,5	93,6	92,8	cbpf	93,4	96,3	1,7	92,9	92,1
inpe	92,5	96,0	1,9	92,2	91,2	usp	92,0	95,7	2,1	91,4	90,4
ctbc	90,9	95,4	2,4	90,6	89,4	fiocruz	90,9	95,1	2,6	89,9	88,6
unesp-sjrp	90,8	94,9	2,7	89,4	88,1	cbpf-2	90,4	94,6	2,8	89,0	87,5
mj	89,7	94,4	3,0	88,4	86,9	rederio-2	89,5	94,1	3,1	87,9	86,4
vivax-mns	88,8	93,9	3,2	87,4	85,8	diveo-2	88,1	93,6	3,4	86,8	85,2
unesp	87,7	93,3	3,5	86,3	84,5	unicamp-feec	86,9	93,1	3,7	85,7	83,8
unitau	86,1	92,8	3,9	85,1	83,1	puc-rio	85,1	92,5	4,1	84,3	82,3
vivax-san	81,2	92,1	4,6	83,0	80,7	fpte	81,0	91,7	4,9	81,8	79,3
ufpa	79,5	91,3	5,3	80,6	77,9	durand-1	77,8	90,8	5,8	79,2	76,3
upf	77,5	90,4	6,2	78,0	74,9	ime	74,3	89,9	6,7	76,4	73,1
unicamp	68,5	89,2	7,6	74,1	70,3	udesc	68,2	88,6	8,3	72,0	67,9
ital	67,1	88,0	8,9	70,1	65,6	usp-ciagri	66,2	87,4	9,5	68,3	63,6
vivax-amr	65,9	86,8	10,0	66,7	61,7	tcu	63,9	86,2	10,6	65,1	59,8
usp-cirp	63,8	85,6	11,0	63,6	58,0	ita	62,5	85,1	11,5	62,1	56,3
lncc-1	62,5	84,5	11,9	60,7	54,8	usp-cisc	62,0	84,0	12,2	59,5	53,4
ufrj	61,9	83,5	12,6	58,4	52,1	ansp	6,5	81,7	17,0	47,7	39,3

Tabela 3.9 - Segunda métrica: fluxo UDP em 2006. Comparação com valor máximo

Sensor	UDP	μ	σ	Limites		Sensor	UDP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
pop-pr	0,0	0,0	0,0	0,0	0,0	ufpa	0,0	0,0	0,0	0,0	0,0
vivax-san	0,0	0,0	0,0	0,0	0,0	rederio	0,9	0,2	0,5	1,4	1,2
cbpf-2	1,0	0,4	0,5	1,7	1,5	cbpf	1,1	0,5	0,6	1,9	1,6
brasiltelecom	1,4	0,6	0,6	2,1	1,8	unb-labredes	1,7	0,8	0,7	2,4	2,1
hp-psc	1,9	0,9	0,7	2,7	2,3	ctbc	2,2	1,0	0,8	3,0	2,6
cert-br	2,4	1,1	0,9	3,3	2,9	ita	3,6	1,3	1,1	4,0	3,5
cert-rs	3,6	1,5	1,2	4,5	3,9	unesp-sjrp	3,8	1,7	1,3	4,9	4,3
diveo	4,0	1,8	1,4	5,3	4,6	cenpra	4,1	2,0	1,5	5,6	4,9
unitau	4,7	2,1	1,6	6,0	5,3	rederio-2	4,8	2,3	1,6	6,4	5,6
fiocruz	5,5	2,5	1,8	6,8	6,0	inpe	5,9	2,6	1,9	7,3	6,4
ebt-rjo1	6,2	2,8	2,0	7,8	6,8	mj	6,7	3,0	2,1	8,3	7,2
diveo-2	7,8	3,2	2,3	8,9	7,8	unesp	8,3	3,4	2,5	9,6	8,3
ufrj	8,8	3,6	2,7	10,2	8,9	usp	8,8	3,8	2,8	10,8	9,4
unicamp-feec	9,0	4,0	2,9	11,3	9,8	udesc	9,2	4,2	3,0	11,7	10,2
vivax-mns	9,6	4,4	3,1	12,2	10,6	usp-cirp	9,8	4,6	3,2	12,7	11,0
durand-1	10,1	4,7	3,3	13,1	11,4	puc-rio	12,8	5,0	3,6	13,9	12,1
ufsc-das	14,8	5,3	3,9	15,1	13,1	upf	17,3	5,6	4,4	16,6	14,4
fpte	19,4	6,0	4,9	18,3	15,8	ime	20,5	6,4	5,4	19,9	17,2
tcu	24,5	6,9	6,1	22,1	19,1	vivax-amr	25,3	7,4	6,7	24,2	20,8
usp-ciagri	25,6	7,9	7,2	25,9	22,3	usp-cisc	29,6	8,4	7,9	28,2	24,2
lncc-1	30,4	9,0	8,5	30,3	26,0	ital	31,0	9,5	9,1	32,2	27,7
unicamp	40,9	10,2	10,2	35,7	30,6	ansp	91,9	12,1	15,9	51,8	43,9

Tabela 3.10 - Segunda métrica: fluxo ICMP em 2006. Comparação com valor máximo

Sensor	ICMP	μ	σ	Limites		Sensor	ICMP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
pop-pr	0,0	0,0	0,0	0,0	0,0	ufpa	0,0	0,0	0,0	0,0	0,0
vivax-san	0,0	0,0	0,0	0,0	0,0	fpte	0,2	0,1	0,1	0,4	0,3
ufsc-das	0,3	0,1	0,1	0,5	0,4	unb-labredes	0,3	0,1	0,1	0,5	0,4
hp-psc	0,4	0,2	0,2	0,6	0,5	usp	0,4	0,2	0,2	0,6	0,5
vivax-mns	0,4	0,2	0,2	0,6	0,5	vivax-amr	0,4	0,2	0,2	0,7	0,6
ansp	0,4	0,2	0,2	0,7	0,6	tcu	0,5	0,3	0,2	0,7	0,6
cenpra	0,6	0,3	0,2	0,8	0,7	ebt-rjo1	0,6	0,3	0,2	0,8	0,7
cert-br	0,7	0,3	0,2	0,9	0,8	unicamp-feec	0,9	0,4	0,2	1,0	0,9
brasiltelecom	0,9	0,4	0,3	1,1	1,0	diveo	1,1	0,4	0,3	1,2	1,1
usp-ciagri	1,1	0,5	0,3	1,3	1,1	mj	1,1	0,5	0,4	1,4	1,2
ufrj	1,5	0,6	0,4	1,6	1,4	upf	1,5	0,6	0,4	1,7	1,5
inpe	1,6	0,6	0,5	1,9	1,6	ctbc	1,9	0,7	0,5	2,0	1,8
rederio	2,0	0,7	0,6	2,2	1,9	ital	2,2	0,8	0,6	2,4	2,1
ita	2,2	0,9	0,7	2,6	2,2	unitau	2,5	0,9	0,7	2,8	2,4
unesp	2,7	1,0	0,8	3,0	2,6	usp-cirp	3,0	1,0	0,9	3,2	2,8
unicamp	3,0	1,1	0,9	3,4	3,0	durand-1	3,1	1,2	1,0	3,6	3,1
puc-rio	3,2	1,2	1,0	3,8	3,3	cbpf-2	3,3	1,3	1,1	3,9	3,4
fiocruz	4,2	1,4	1,2	4,3	3,7	rederio-2	4,5	1,5	1,3	4,6	4,0
udesc	4,6	1,5	1,3	4,9	4,2	unesp-sjrp	4,7	1,6	1,4	5,2	4,5
cbpf	4,9	1,7	1,5	5,4	4,7	cert-rs	5,2	1,8	1,6	5,7	4,9
ime	12,6	2,1	2,3	7,8	6,7	diveo-2	15,4	2,4	3,1	10,0	8,5
lncc-1	17,2	2,7	3,8	12,2	10,3	usp-cisc	22,5	3,2	4,8	15,1	12,7

3.5.3 Métrica: relação entre fluxos TCP

A métrica privilegia, tal como na anterior, os sensores cujos dados coletados se situam entre o limite superior do valor mínimo e o limite inferior do valor máximo. Para os sensores que tenham dados que se situam entre os limites inferior e superior é atribuído 0,5 pontos. Para os sensores que tenham dados coletados que se situam acima do limite superior do valor máximo ou abaixo do limite inferior do valor mínimo, 1 ponto.

A métrica é aplicada de forma independente para os dados coletados nos anos de 2005 e de 2006 para os fluxos TCP com menos de 3 pacotes e os com 3 ou mais pacotes.

Portanto, para cada tipo de fluxo TCP, teria-se de realizar duas comparações por ano: uma em relação ao valor máximo e outra em relação ao valor mínimo. Entretanto, a prática mostrou que para o o fluxo com menos de 3 pacotes a distribuição das observações permanece abaixo do limite inferior do valor máximo sendo necessária somente a comparação com o valor mínimo.

Tabela 3.11 - Terceira métrica: fluxo TCP < 3 em 2005. Comparação com valor mínimo

Sensor	TCP	μ	σ	Limites		Sensor	TCP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
durand-1	100,0	100,0	0,0	100,0	100,0	udesc	99,8	99,9	0,1	99,6	99,6
ital	99,7	99,8	0,1	99,6	99,5	ita	98,7	99,5	0,6	98,4	98,1
puc-rio	97,8	99,2	0,9	97,4	96,9	ebt-rjo1	97,4	98,9	1,1	96,7	96,1
usp-cisc	95,7	98,4	1,6	95,3	94,5	usp-ciagri	94,8	98,0	1,9	94,1	93,1
rederio-2	94,7	97,6	2,1	93,3	92,3	unitau	94,0	97,2	2,3	92,6	91,5
fpte	91,3	96,7	2,8	91,1	89,6	ufsc-das	91,0	96,2	3,2	89,9	88,3
upf	90,5	95,8	3,4	88,9	87,2	unicamp-feec	90,4	95,4	3,6	88,2	86,4
brasiltelecom	90,3	95,1	3,7	87,6	85,8	fiocruz	89,4	94,7	3,8	87,0	85,1
cbpf	88,7	94,4	4,0	86,4	84,4	tcu	88,7	94,0	4,1	85,8	83,8
vivax-mns	88,4	93,7	4,2	85,4	83,3	lncc-1	88,4	93,5	4,2	85,0	82,9
vivax-amr	88,4	93,2	4,3	84,7	82,5	ctbc	88,3	93,0	4,3	84,4	82,2
mj	87,1	92,8	4,4	84,0	81,8	diveo-2	86,9	92,5	4,5	83,6	81,4
diveo	86,2	92,3	4,5	83,2	80,9	unicamp	84,1	91,9	4,7	82,5	80,1
ufrj	83,3	91,6	4,9	81,8	79,3	ansp	82,6	91,3	5,1	81,1	78,5
unesp	78,2	90,9	5,6	79,7	76,9	cbpf-2	77,9	90,4	6,0	78,5	75,5
usp-cirp	77,2	90,0	6,3	77,3	74,2	cert-rs	76,8	89,6	6,7	76,3	73,0
cert-br	75,4	89,2	7,0	75,2	71,7	unesp-sjrp	72,6	88,7	7,5	73,8	70,0
usp	69,6	88,1	8,0	72,1	68,1	unb-labredes	62,8	87,4	9,0	69,5	65,0
hp-psc	62,3	86,7	9,8	67,2	62,4	inpe	56,0	85,9	10,8	64,3	58,8
ime	52,2	85,1	12,0	61,1	55,1	rederio	45,6	84,1	13,4	57,3	50,7
cenpra	40,7	83,0	14,8	53,3	45,9	vivax-san	0,0	81,0	19,5	42,1	32,4
ufpa	0,0	79,2	22,9	33,4	22,0	pop-pr	0,0	77,4	25,6	26,3	13,5

Para o fluxo TCP com 3 ou mais pacotes os valores observados se localizam acima do limite superior do valor mínimo sendo necessário compará-los somente com o valor máximo.

Tabela 3.12 - Terceira métrica: fluxo TCP ≥ 3 em 2005. Comparação com valor máximo

Sensor	TCP	μ	σ	Limites		Sensor	TCP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
pop-pr	0,0	0,0	0,0	0,0	0,0	ufpa	0,0	0,0	0,0	0,0	0,0
vivax-san	0,0	0,0	0,0	0,0	0,0	durand-1	0,0	0,0	0,0	0,1	0,1
udesc	0,2	0,1	0,1	0,3	0,2	ital	0,3	0,1	0,1	0,4	0,4
ita	1,3	0,3	0,5	1,5	1,2	puc-rio	2,2	0,5	0,8	2,5	2,1
ebt-rjo1	2,6	0,7	1,0	3,3	2,8	usp-cisc	4,3	1,1	1,5	4,8	4,1
usp-ciagri	5,2	1,5	1,9	6,2	5,2	rederio-2	5,3	1,8	2,1	7,1	6,0
unitau	6,0	2,1	2,3	7,9	6,8	fpte	8,7	2,6	2,8	9,7	8,3
ufsc-das	9,0	3,0	3,2	11,0	9,4	upf	9,5	3,4	3,5	12,2	10,4
unicamp-feec	9,6	3,8	3,7	13,1	11,2	brasiltelecom	9,7	4,1	3,9	13,8	11,8
fiocruz	10,6	4,5	4,0	14,5	12,5	cbpf	11,3	4,8	4,2	15,3	13,2
tcu	11,3	5,1	4,3	16,0	13,8	vivax-mns	11,6	5,4	4,5	16,5	14,3
lncc-1	11,6	5,7	4,5	17,0	14,8	vivax-amr	11,6	5,9	4,6	17,4	15,1
ctbc	11,7	6,1	4,7	17,8	15,5	mj	12,9	6,4	4,7	18,3	15,9
diveo-2	13,1	6,7	4,8	18,7	16,3	diveo	13,8	6,9	4,9	19,2	16,8
unicamp	15,9	7,2	5,1	20,0	17,4	ufrj	16,7	7,5	5,3	20,8	18,2
ansp	17,4	7,9	5,5	21,6	18,9	unesp	21,8	8,3	6,0	23,2	20,2
cbpf-2	22,1	8,7	6,3	24,6	21,4	usp-cirp	22,8	9,1	6,7	25,9	22,5
cert-rs	23,2	9,5	7,0	27,0	23,5	cert-br	24,6	9,9	7,3	28,3	24,6
unesp-sjrp	27,4	10,4	7,8	29,9	26,0	usp	30,4	10,9	8,3	31,8	27,6
unb-labredes	37,2	11,6	9,2	34,7	30,1	hp-psc	37,7	12,3	10,0	37,3	32,3
inpe	44,0	13,0	11,1	40,7	35,2	ime	47,8	13,9	12,2	44,3	38,2
rederio	54,4	14,8	13,5	48,6	41,8	cenpra	59,3	15,8	15,0	53,2	45,7

Tabela 3.13 - Terceira métrica: fluxo TCP< 3 em 2006. Comparação com valor mínimo

Sensor	TCP	μ	σ	Limites		Sensor	TCP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
ufpa	100	100	0,0	100	100	vivax-san	100	100	0,0	100	100
durand-1	100,0	100,0	0,0	99,9	99,9	ctbc	99,4	99,8	0,3	99,3	99,1
pop-pr	99,0	99,7	0,5	98,7	98,5	puc-rio	97,7	99,3	0,9	97,6	97,1
usp-cirp	97,5	99,1	1,1	97,0	96,4	ansp	96,8	98,8	1,3	96,2	95,6
ita	95,5	98,4	1,6	95,2	94,4	hp-psc	95,3	98,1	1,8	94,5	93,6
cbpf-2	94,5	97,8	2,0	93,7	92,7	usp-cisc	94,2	97,5	2,2	93,0	91,9
unicamp-feec	93,8	97,2	2,4	92,5	91,3	rederio-2	92,9	96,9	2,5	91,8	90,6
fpte	91,6	96,5	2,8	90,9	89,5	lncc-1	90,9	96,2	3,1	90,1	88,6
udesc	89,6	95,8	3,4	89,1	87,4	unicamp	89,5	95,4	3,6	88,3	86,5
vivax-mns	89,3	95,1	3,8	87,6	85,7	usp-ciagri	89,2	94,8	3,9	87,0	85,1
upf	89,1	94,6	4,0	86,6	84,6	vivax-amr	88,5	94,3	4,1	86,1	84,0
ufsc-das	87,2	94,0	4,3	85,4	83,3	diveo-2	85,8	93,6	4,5	84,6	82,4
mj	84,8	93,3	4,7	83,8	81,4	cbpf	83,2	92,9	5,1	82,8	80,3
ital	82,9	92,5	5,3	81,9	79,2	unitau	82,8	92,2	5,5	81,1	78,4
tcu	82,7	91,8	5,7	80,4	77,6	ufrj	82,6	91,5	5,9	79,8	76,9
unesp-sjrp	73,2	91,0	6,6	77,7	74,4	brasiltelecom	70,7	90,3	7,4	75,4	71,7
fiocruz	70,3	89,7	8,1	73,5	69,4	unb-labredes	67,4	89,1	8,9	71,3	66,9
diveo	67,3	88,4	9,5	69,5	64,8	ime	67,2	87,8	10,0	67,9	62,9
ebt-rjo1	66,4	87,3	10,5	66,4	61,1	usp	64,3	86,7	11,0	64,7	59,3
cert-rs	59,2	86,0	11,7	62,6	56,8	unesp	54,4	85,2	12,6	60,0	53,8
inpe	51,7	84,4	13,5	57,4	50,7	cert-br	50,0	83,5	14,3	54,9	47,8
cenpra	48,2	82,7	15,1	52,5	44,9	rederio	28,6	81,5	17,0	47,4	38,9

Tabela 3.14 - Terceira métrica: fluxo TCP ≥ 3 em 2006. Comparação com valor máximo

Sensor	TCP	μ	σ	Limites		Sensor	TCP	μ	σ	Limites	
				Sup	Inf					Sup	Inf
ufpa	0,0	0,0	0,0	0,0	0,0	vivax-san	0,0	0,0	0,0	0,0	0,0
durand-1	0,0	0,0	0,0	0,1	0,1	ctbc	0,6	0,2	0,3	0,9	0,7
pop-pr	1,0	0,3	0,5	1,5	1,3	puc-rio	2,3	0,7	0,9	2,9	2,4
usp-cirp	2,5	0,9	1,1	3,6	3,0	ansp	3,2	1,2	1,3	4,4	3,8
ita	4,5	1,6	1,6	5,6	4,8	hp-psc	4,7	1,9	1,8	6,4	5,5
cbpf-2	5,5	2,2	2,0	7,3	6,3	usp-cisc	5,8	2,5	2,2	8,1	7,0
unicamp-feec	6,2	2,8	2,4	8,7	7,5	rederio-2	7,1	3,1	2,5	9,4	8,2
fpte	8,4	3,5	2,8	10,5	9,1	lncc-1	9,1	3,8	3,1	11,4	9,9
udesc	10,4	4,2	3,4	12,6	10,9	unicamp	10,5	4,6	3,6	13,5	11,7
vivax-mns	10,7	4,9	3,8	14,3	12,4	usp-ciagri	10,8	5,2	3,9	14,9	13,0
upf	10,9	5,4	4,0	15,4	13,4	vivax-amr	11,5	5,7	4,1	16,0	13,9
ufsc-das	12,8	6,0	4,3	16,7	14,6	diveo-2	14,2	6,4	4,5	17,6	15,4
mj	15,2	6,7	4,7	18,6	16,2	cbpf	16,8	7,1	5,1	19,7	17,2
ital	17,1	7,5	5,3	20,8	18,1	unitau	17,2	7,8	5,5	21,6	18,9
tcu	17,3	8,2	5,7	22,4	19,6	ufrj	17,4	8,5	5,9	23,1	20,2
unesp-sjrp	26,8	9,0	6,6	25,6	22,3	brasiltelecom	29,3	9,7	7,4	28,3	24,6
fiocruz	29,7	10,3	8,1	30,6	26,5	unb-labredes	32,6	10,9	8,9	33,1	28,7
diveo	32,7	11,6	9,5	35,2	30,5	ime	32,8	12,2	10,0	37,1	32,1
ebt-rjo1	33,6	12,7	10,5	38,9	33,6	usp	35,7	13,3	11,0	40,7	35,3
cert-rs	40,8	14,0	11,7	43,2	37,4	unesp	45,6	14,8	12,6	46,2	40,0
inpe	48,3	15,6	13,5	49,3	42,6	cert-br	50,0	16,5	14,3	52,2	45,1
cenpra	51,8	17,3	15,1	55,1	47,5	rederio	71,4	18,5	17,0	61,1	52,6

3.5.4 Conjunto de sensores selecionados pela aplicação da métrica

A pontuação resultante da aplicação das métricas encontra-se resumida na Tabela 3.15. Para melhor visualização a tabela está ordenada, em ordem crescente, do número de pontos dados a cada sensor pela aplicação das métricas. Ela mostra que o sensor VIVAX-MNS é o mais aderente ao conjunto de critérios estabelecidos. A partir deste resultado irá se tentar buscar os sensores que possuam observações que se comportem o mais próximo possível das métricas estabelecidas. Para tal adotou-se o critério de adotar

o conjunto de sensores cuja pontuação se situe abaixo da mediana (valor calculado para mediana 2).

Tabela 3.15 - Pontuação obtida pelos sensores após aplicação das métricas

Sensor	Critério			Total	Sensor	Critério			Total
	1	2	3			1	2	3	
vivax-mns	0,0	0,0	0,0	0,0	mj	0,0	0,0	0,0	0,0
ita	0,0	0,0	0,5	0,5	durand-1	0,0	0,5	0,0	0,5
diveo	0,0	0,0	1,0	1,0	unitau	0,0	1,0	0,0	1,0
ufrj	0,0	1,0	0,0	1,0	cbpf	0,0	1,0	0,0	1,0
cert-br	0,0	0,0	1,0	1,0	ufsc-das	0,0	1,5	0,0	1,5
udesc	0,0	1,5	0,0	1,5	puc-rio	0,0	1,5	0,5	2,0
usp-ciagri	0,0	2,0	0,0	2,0	rederio-2	1,0	1,0	0,0	2,0
unicamp-feec	2,0	0,0	0,0	2,0	brasiltelecom	0,0	0,0	2,0	2,0
usp	0,0	0,0	2,0	2,0	cbpf-2	1,0	0,0	1,0	2,0
fiocruz	0,0	1,0	1,0	2,0	ebt-rjo1	1,0	0,0	1,0	2,0
cert-rs	0,0	1,0	1,0	2,0	ctbc	1,0	1,0	0,0	2,0
hp-psc	0,0	0,0	2,0	2,0	usp-cirp	1,0	1,0	1,0	3,0
inpe	0,0	0,0	3,0	3,0	diveo-2	0,0	3,0	0,0	3,0
unesp	0,0	1,0	2,0	3,0	unb-labredes	0,0	0,0	3,0	3,0
cenpra	0,0	0,0	3,0	3,0	pop-pr	1,0	1,0	1,0	3,0
vivax-amr	0,0	3,0	0,0	3,0	tcu	0,0	3,5	0,0	3,5
rederio	0,0	0,0	4,0	4,0	unesp-sjrp	0,0	1,0	3,0	4,0
upf	1,0	3,0	0,0	4,0	vivax-san	2,0	1,5	1,0	4,5
fpte	1,0	3,0	0,5	4,5	lncc-1	0,0	4,5	0,0	4,5
ufpa	2,0	1,5	1,0	4,5	ansp	1,0	4,0	0,0	5,0
unicamp	0,0	5,0	0,0	5,0	ital	1,0	4,0	0,0	5,0
usp-cisc	0,0	5,0	0,5	5,5	ime	0,0	5,5	3,0	8,5

Para permitir uma melhor visualização do critério de seleção adotado é gerado o histograma da distribuição de pontos dos sensores na Figura 3.2.

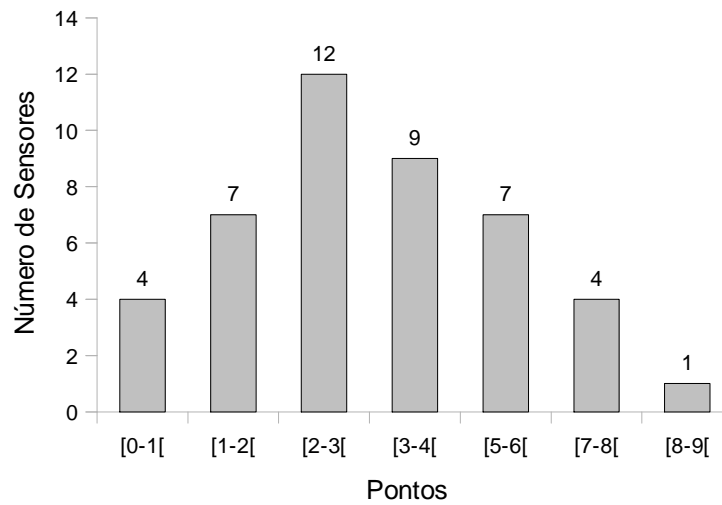


Figura 3.2 - Histograma de distribuição dos pontos obtidos após aplicação dos critérios de seleção.

Da análise do histograma verifica-se que 23 sensores, de uma população de 44 sensores, o que corresponde a mais de 50% da população, possui menos de 3 pontos.

Assim, passam a constituir a amostra utilizada, a partir deste ponto, para todas os demais cálculos os seguintes sensores: brasilelecom, cbpf, cbpf-2, cert-br, cert-rs, ctbc, diveo, durand-1, ebt-rjo1, fiocruz, hp-psc, ita, mj, puc-rio, rederio-2, udesc, ufrj, ufsc-das, unicamp-feec, unitau, usp, usp-ciagri e vivax-mns.

4 CARACTERÍSTICAS DO RUÍDO DE FUNDO DA INTERNET BRASILEIRA

O ruído de fundo está presente na Internet independentemente da vontade dos usuários. Ao se mensurar e caracterizar este tráfego são fornecidas ferramentas para os administradores de sistemas que lhes permita verificar a ocorrência de variações neste comportamento o que estaria indicando a ocorrência de atividades maliciosas de grande espectro. Barros (2007) cita critérios para caracterizar este tráfego.

Quando se pensa em infraestrutura crítica de uma nação este conhecimento é altamente relevante uma vez que ataques a estas estruturas tendem a ter uma assinatura de grande quantidade de tráfego. Não só para a detecção mas, também, para a prevenção uma vez que, se detectada as fases iniciais da infecção é possível a geração de alertas precoces.

O conjunto de sensores determinados no Capítulo 3 será usado como representativo do comportamento médio esperado do ruído de fundo na parcela brasileira da Internet.

4.1 Percentual de fluxos TCP, UDP e ICMP no ruído de fundo

A Figura 4.1 apresenta o valor percentual de cada protocolo, em relação ao fluxo total, obtido do conjunto amostral, usando a janela de amostragem ano.

Embora haja variações entre os anos estipulou-se, pela média, os seguintes valores esperados para cada um dos protocolos: TCP 93,70%, UDP 4,50% e ICMP 1,80%.

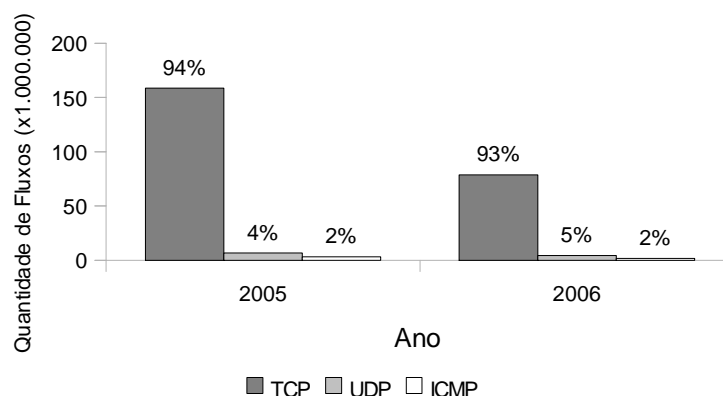


Figura 4.1 - Percentual de fluxos dos protocolos TCP, UDP e ICMP no ruído de fundo da parcela brasileira da Internet.

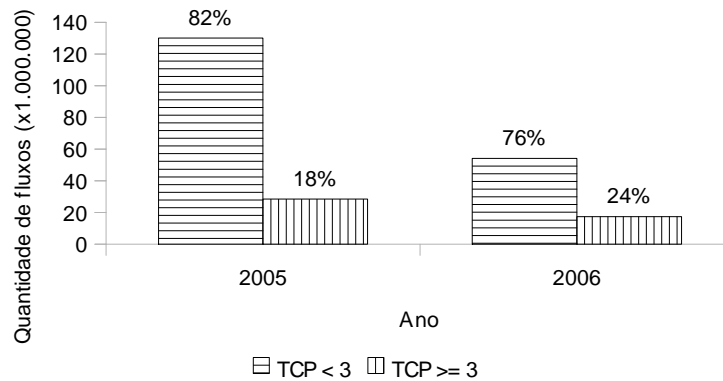


Figura 4.2 - Percentual de fluxos TCP com menos de 3 e com 3 ou mais pacotes no ruído de fundo da parcela brasileira da Internet.

4.2 Percentual de fluxos TCP

A Figura 4.2 apresenta o valor percentual dos fluxos com menos de 3 pacotes e os com 3 ou mais pacotes em relação ao fluxo total do protocolo TCP, obtido do conjunto amostral, usando a janela de amostragem ano.

Verifica-se que, em média, tem-se a seguinte distribuição dos fluxos TCP na parcela brasileira da Internet: com menos de 3 pacotes 78,9% e com 3 ou mais pacotes 21,1%.

Esta característica quantifica o tráfego eminentemente de varredura e o tráfego malicioso que estabelece conexão.

4.3 Portas mais acessadas

Caracteriza o relacionamento entre fluxo e portas acessadas. Para estes processo de caracterização será utilizada a representação descrita por Grizzard et al (2005). Neste trabalho o relacionamento entre as portas acessadas e o fluxo é representado através de um gráfico bi-dimensional onde o eixo horizontal representa as portas e o eixo vertical a escala de tempo considerada. Cada posição dentro do gráfico recebe um valor a partir da Equação 4.1.

$$I(\text{intensidade}) = \begin{cases} 0, & c=0 \\ 0,75(\frac{c}{c_{max}}) + 0,25, & c \neq 0 \end{cases} \quad (4.1)$$

onde,

I é a intensidade do sinal a ser representado no gráfico,

c é o valor do fluxo destinado a uma porta considerada a janela de amostragem, e

c_{max} é o total de fluxos destinado a uma determinada porta.

Na representação final é gerada uma imagem em tons de cinza em que cada pixel, ou conjunto de pixels, contém a intensidade. $i = 0$, ou falta de intensidade, é representado pela cor preta. Qualquer outro valor é representado a partir de uma tonalidade com 25% de cinza ($i = 0,25$) até o branco ($i = 1$). Optou-se por uma representação com menos regiões escuras porque ela fornece um melhor contraste com as áreas de atividade.

Como a maioria das varreduras ocorre de forma automatizada sem considerar critérios técnicos há ocorrências de acessos a portas não relevantes e muitos acessos esporádicos. Esta característica não traz nenhum significado para o trabalho. É simplesmente uma varredura automatizada varrendo endereços sem maiores consequências. O que nos importa são as portas que efetivamente estão sendo procuradas e sondadas.

Segundo Barros (2010a), para eliminar as informações sem significado estabeleceu-se que os fluxos de interesse são os que se propagam ao longo de pelo menos três janelas de amostragem consecutivas e os com alta taxa de amostragem.

Com a aplicação do critério reduz-se a quantidade de portas, do fluxo TCP com menos de 3 pacotes, para 4.651 e 72 se consideradas as janelas de amostragem dia e hora, respectivamente.

Para os fluxos TCP com 3 ou mais pacotes não foi aplicado o critério uma vez que so-

mente 744 portas foram acessadas e este número foi considerado aceitável.

Tabela 4.1 - Percentual de fluxos TCP com menos e com mais de 3 pacotes, por porta, para as janelas hora e dia

Fluxo TCP < 3 Pacotes						Fluxo TCP >= 3 Pacotes		
Hora			Dia			Hora E/ou Dia		
Porta	%	% Acum.	Porta	%	% Acum.	Porta	%	% Acum.
445	37,57%	37,57%	445	35,10%	35,10%	1080	26,18%	26,18%
1080	9,45%	47,03%	1080	8,86%	43,96%	135	20,44%	46,62%
1433	9,12%	56,15%	1433	8,52%	52,48%	1433	15,43%	62,05%
139	8,79%	64,94%	139	8,23%	60,71%	139	14,31%	76,36%
135	8,35%	73,29%	135	7,82%	68,53%	445	11,15%	87,52%
80	5,93%	79,22%	80	5,54%	74,06%	4899	3,00%	90,51%
4899	2,81%	82,03%	4899	2,65%	76,71%	1025	2,16%	92,67%
3306	1,79%	83,81%	3306	1,68%	78,39%	80	2,15%	94,82%
1023	1,57%	85,38%	1023	1,49%	79,89%	9898	1,41%	96,23%
22	1,17%	86,56%	22	1,11%	80,99%	22	1,13%	97,36%
15118	1,17%	87,72%	15118	1,09%	82,09%	4444	0,73%	98,09%
137	1,17%	88,89%	137	1,09%	83,18%	4000	0,34%	98,43%
5554	1,08%	89,97%	5554	1,05%	84,23%	15118	0,31%	98,74%
1025	0,78%	90,75%	2100	0,92%	85,15%	3306	0,30%	99,04%
42	0,77%	91,52%	10000	0,77%	85,92%	5000	0,29%	99,33%
2100	0,74%	92,26%	1025	0,75%	86,66%	11768	0,14%	99,47%
10000	0,71%	92,96%	42	0,74%	87,40%	6101	0,09%	99,56%
25	0,66%	93,62%	25	0,65%	88,05%	6129	0,05%	99,61%
6129	0,64%	94,26%	5900	0,64%	88,70%	554	0,04%	99,65%
5900	0,64%	94,90%	6129	0,63%	89,33%	5554	0,04%	99,69%
6101	0,56%	95,46%	6101	0,56%	89,88%	3128	0,04%	99,73%
5000	0,54%	96,00%	5000	0,53%	90,41%	2745	0,03%	99,76%
21	0,47%	96,48%	9898	0,52%	90,94%	8080	0,03%	99,79%
11768	0,39%	96,87%	21	0,43%	91,36%	25	0,02%	99,80%
9898	0,37%	97,24%	143	0,38%	91,74%	41523	0,02%	99,82%
143	0,35%	97,59%	8080	0,37%	92,12%	3127	0,02%	99,84%

A Tabela 4.1 mostra que aproximadamente 97,5% do fluxo TCP com menos de 3 paco-

tes e janela de amostragem hora é direcionado para somente 26 portas. Entretanto, para a janela de amostragem dia, são acessadas 951 portas para obter esta mesma percentagem acumulada. Para fluxos TCP com 3 ou mais pacotes somente 7 portas representam 97,5% dos fluxos.

A porta 445 foi a que mais recebeu fluxos com menos de 3 pacotes com quase quatro vezes mais fluxos do que a segunda, a 1080. A porta que recebeu mais fluxos com 3 ou mais pacotes recebeu foi a 1080 seguida, de perto, da 135 e, um pouco mais distante, das portas 1433, 139 e 445.

A Tabela 4.2 apresenta uma listagem das 12 portas mais acessadas para os fluxos com 3 ou mais pacotes, qual seu uso correto e qual possível vulnerabilidade pode ser explorada,

Tabela 4.2 - Portas mais acessadas, seus usos corretos e possíveis vulnerabilidades/ataques

Porta	Vulnerabilidades/Ataques	Registro IANA
22	Busca por aplicações SSH com vulnerabilidades conhecidas. <i>InCommand, Shaft, Skun, Adore sshd, accidental hit</i>	SSH Remote Login Protocol
80	Intercâmbio de páginas web, e serviços associados. <i>AckCmd, Back Orifice 2000, BackDoor.Gaster, Blaster.D Worm, CGI Backdoor, Code Red, Code Red II, Code Red.F, IISworm, MyDoom.B, Trojan.AnyMail, ...</i>	World Wide Web HTTP
135	Mapeamento dos serviços RPC para atribuição dinâmica de portas. <i>DCOM/MSBlast, Netbios RPC, W32.Blaster, W32/Lovsan.worm</i>	epmap - DCE endpoint resolution
139	Compartilhamento de arquivos e de impressoras nos sistemas Windows. <i>Chode, Fire HackeR, Msinit, Nimda, Opaserv, Qaz, God Message worm, SMB Relay, Fire HackeR, ...</i>	netbios-ssn - NETBIOS Session Service

(continua)

Tabela 4.2 - Conclusão

Porta	Vulnerabilidades/Ataques	Registro IANA
445	A partir do Windows 2000 o protocolo SMB, passa a usar, através dos protocolos de transporte TCP e UDP, a porta 445. <i>Backdoor.rkit.b, Lioten, Randon, Sasser, Nimda, Trojan.Netdepix.b, W32.HLLW.Deloder, W32.Scane, ...</i>	microsoft-ds - Microsoft-DS
1025	É a primeira porta que é atribuída dinamicamente. Logo, praticamente qualquer programa que requeira uma porta pode ser atribuído a esta porta. <i>AcidkoR, DataSpy Network X, KiLo, MuSka52, NetSpy, Optix, Remote Anything, Remote Explorer Y2K, Remote Storm, ...</i>	network blackjack
1080	Geralmente associada ao aplicativo Wingate – firewall/proxy para Windows. <i>MyDoom.F, Seeking Win32:BugBear-B, SubSeven 2.2, WinHole</i>	socks
1433	Porta principal do Microsoft SQL Server <i>SQL Snake, w32.spybot.ofn, Voyager Alpha Force</i>	ms-sql-s - Microsoft-SQL-Server
4000	Porta de controle do ICP <i>Connect-Back Backdoor, Psyber StreamingServer, RemoteAnything, Skydance trojan, WityWorm, ...</i>	terabase
4444	Uma vulnerabilidade na interface DCOM RPC da Microsoft, permite ataques na porta 135/TCP e a criação, nos hosts comprometidos, de um "backdoor" com uma "shell" privilegiada de comandos na porta 4444/TCP <i>AlexTrojan, CrackDown, Oracle, MS Blaster listening port, W32.Blaster Worm, W32.Hllw.Donk.M, W32.rei-dana.a, ...</i>	kbr524
4899	Porta principal do serviço Radmin-- um serviço que permite acesso remoto a um sistema Windows. <i>W32.RaHack</i>	radmin-port - RAdmin Port
9898	"Backdoor" deixado por outros "malwares" como o worm Sasser que estabelece um servidor FTP nesta porta. <i>Dabber, W32.dabber.a</i>	monkeycom - MonkeyCom

Fonte: Adaptada de Suspect Ports, Latham, J. L., [online], <<http://www.jlathamsite.com/dslr/suspectports.htm>>, Out 2010; Port Numbers [online], <<http://www.iana.org/assignments/port-numbers>>, Mar 2010 e Network ICE, Port Knowledgebase, [online], <http://www.iss.net/security_center/advice/Ex>

ploits/Ports/default.htm>, Mar 2010.

A Figura 4.3 ilustra o acesso às portas representadas na Tabela 4.1 para o período de amostragem hora durante uma semana (de 01/01/2005 00:00 até 07/01/2005 23:00). A representação por somente uma semana é uma questão puramente de conveniência podendo, se necessário, ser representado qualquer intervalo e qualquer número de portas.

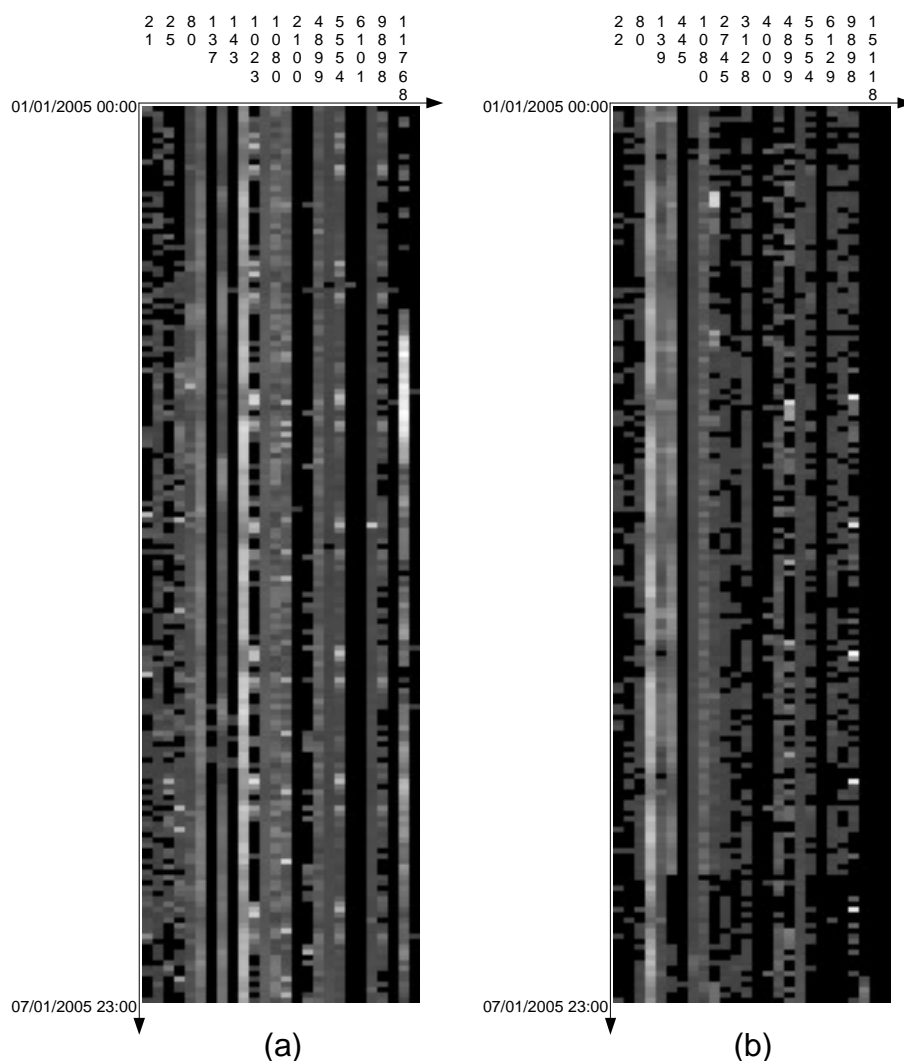


Figura 4.3 - (a) Fluxo TCP com menos de 3 pacotes na primeira semana de janeiro de 2005, usando janela de amostragem hora, nas 26 portas mais significativas; (b) Fluxo TCP com 3 ou mais pacotes na mesma situação.

A análise da imagem é feita através da observação da variação das tonalidades das linhas horizontais e verticais. Na Figura 4.3 (a) as linhas horizontais representam varredu-

ras, isto é, diferentes portas sendo acessadas dentro de uma mesma janela de amostragem e, as linhas verticais, se uma porta é muito varrida ou não. Já, na Figura 4.3 (b) as linhas horizontais representam conexões a diferentes portas dentro de uma mesma janela de amostragem enquanto que, as verticais, indicam conexões a uma mesma porta ao longo do tempo.

Verifica-se que algumas portas, como as 135 e 445 são continuamente varridas uma vez que, para estas portas, as linhas verticais se destacam. Que no período amostrado, a princípio, não houve uma ferramenta automatizada que tenha varrido todas as portas uma vez que não há linhas horizontais que se destaquem.

4.4 Endereçamento IP

Os sensores do CBH são constituídos de sub redes com endereços válidos da organização que os gere. São, portanto, máquinas com endereços IP válidos! O espaço do endereçamento IP à disposição do CBH encontra-se listado na Tabela 4.3.

Tabela 4.3 - Espaço de endereçamento IP, por sensor, do CBH

Sensor	Número De Endereços IP	Sensor	Número De Endereços IP
brasiltelecom	128	mj	32
cbpf	256	puc.rio	256
cbpf.2	2.048	rederio.2	256
cert.br	16	udesc	32
cert.rs	32	ufrj	32
ctbc	256	ufsc.das	32
diveo	64	unicamp.feec	16
durand.1	16	unitau	16
ebt.rjo1	64	usp	64
fiocruz	256	usp.ciagri	16
hp.psc	16	vivax.mns	16
ita	256	Total	4.176

Dos registros sanitizados pode-se, ainda, em relação a endereçamento IP, verificar quan-

tos endereços IP foram usados para gerar os fluxos capturados. Segundo Barros (2010a), não se pode assegurar que o endereço listado no arquivo do sensor corresponda a uma máquina única – o uso de um tradutor de endereços (NAT) permite o uso de um mesmo endereço por várias máquinas. Assim, a Tabela 4.4 lista a quantidade de endereços únicos que acessaram cada sensor e não a quantidade de máquinas que geraram fluxo.

Na Tabela 4.4 ocorrem sobreposições de endereços entre sensores, isto é, um determinado endereço IP cujo fluxo tenha sido capturado pelo sensor cbpf e pelo sensor mj é considerado duas vezes, uma para cada sensor. Retiradas as sobreposições tem-se que 2.614.825 endereços IP únicos geraram fluxos com menos de 3 pacotes e 1.001.935 geraram fluxos com 3 ou mais pacotes.

Tabela 4.4 - Espaço de endereçamento IP remoto, por sensor

Sensor	TCP < 3	TCP >= 3	Sensor	TCP < 3	TCP >= 3
brasiltelecom	1.269.478	295.335	mj	145.304	40.777
cbpf	259.718	36.749	puc.rio	77.636	136
cbpf.2	180.222	74.187	rederio.2	136.756	265
cert.br	399.263	140.772	udesc	51.062	3.188
cert.rs	153.925	119.211	ufrj	12.355	1.824
ctbc	222.720	37.544	ufsc.das	8.149	414
diveo	887.506	273.776	unicamp.feec	2.529	217
durand.1	92.426	39	unitau	115.878	29.881
ebt.rjo1	151.245	50.661	usp	36.789	35.391
fiocruz	65.695	19.768	usp.ciagri	37.002	5.547
hp.psc	347.981	145.710	vivax.mns	166.265	52.905
ita	58.875	1.040	Total	4.878.779	1.365.337

De posse desta informação o que se deseja é determinar que tipo de endereço gera mais fluxos. Para tal faz-se uma representação gráfica da quantidade de fluxos gerados por cada endereço IP único, isto é, se um certo endereço *a.b.c.d* aparece na listagem dos diversos sensores 10 vezes ele terá 10 fluxos. Esta abordagem permite verificar a quantidade de fluxos a partir do endereço de origem.

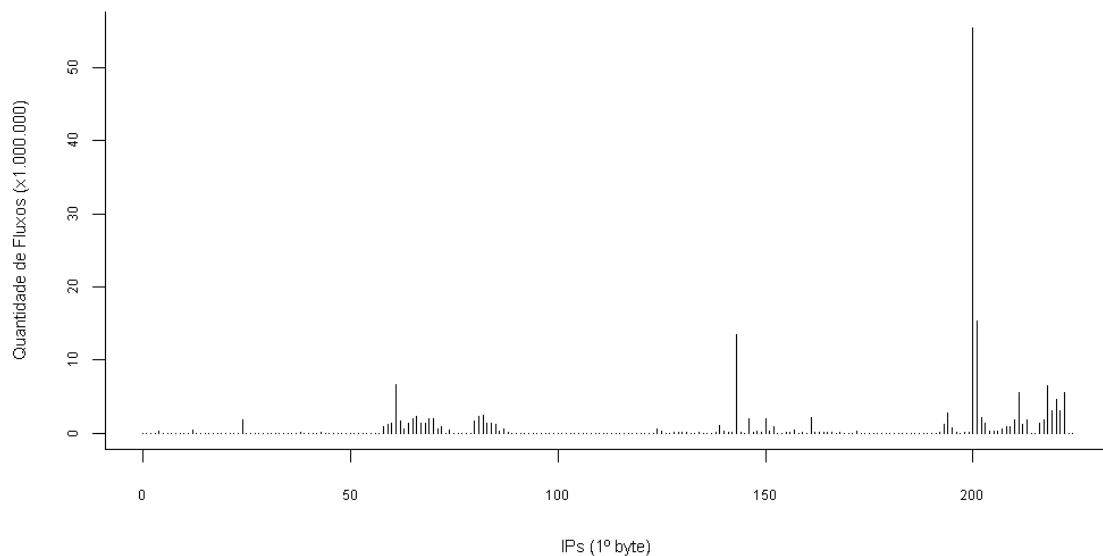


Figura 4.4 - Quantidade de fluxos TCP com menos de 3 pacotes agrupados por endereço IP de origem.

Os endereços IP são agrupados pelo primeiro byte de sua representação decimal independente da classe da rede ao qual pertence. Como o endereçamento IP válido trata somente até o endereço 223.255.255.255 (classes A, B e C; classes D e E são não válidas para tráfego na Internet) todos os endereços superiores a 224 (inclusive) serão agrupados no endereço 224.

A representação gráfica do número de fluxos gerados por cada endereço IP pode ser vista na Figura 4.4. Verifica-se que há um endereço em particular, o que possui o primeiro octeto 200, que gerou muito mais fluxos que do que todos os demais.

Gráfico semelhante é apresentado na Figura 4.5 para os fluxos TCP com 3 ou mais pacotes.

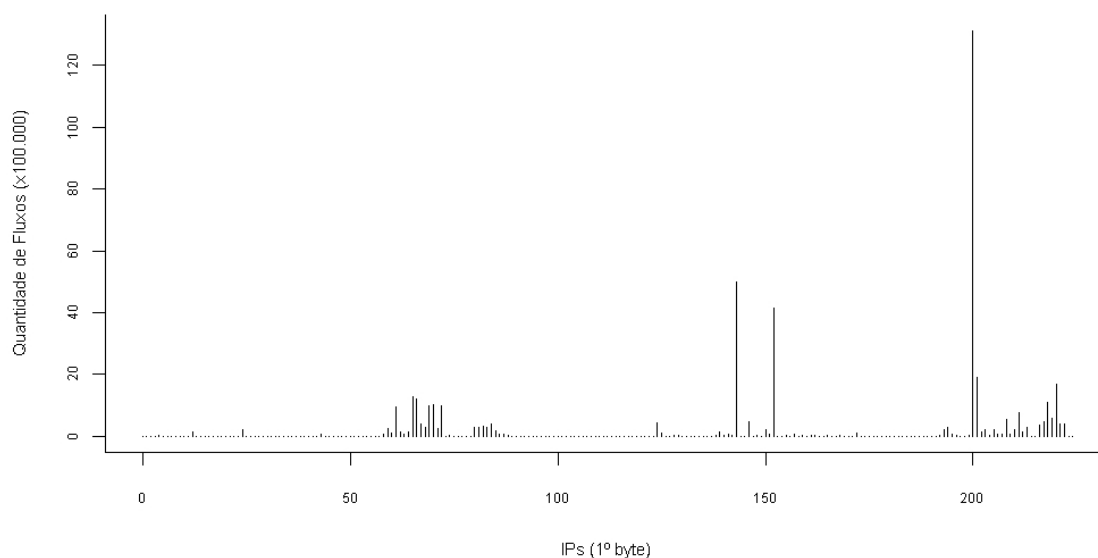


Figura 4.5 - Quantidade média de fluxos TCP com 3 ou mais pacotes agrupados por endereço IP de origem.

Assim como no fluxo TCP com menos de 3 pacotes, os endereços IP cujo primeiro octeto é 200 gera mais fluxos TCP com 3 ou mais pacotes do que todos os demais.

Outra análise viável apresenta o relacionamento entre os endereços IP com as portas que tentaram acessar. Esta abordagem já foi usada por outros autores como se pode ver em Grizard et al. (2005).

A Figura 4.6 apresenta esta relação para fluxos TCP com menos de 3 pacotes. São representadas no eixo vertical as portas listadas na Tabela 4.1. Na horizontal representa-se o endereçamento IP a partir do seu primeiro octeto, de 0 à 224.

A análise desta figura é feita buscando-se por linhas horizontais e verticais. Uma faixa horizontal indica uma porta que todos os endereços IP tentam acessar. Uma faixa vertical um endereço que acessa todas as portas representadas.

Nenhuma das portas representadas na Figura 4.6 possui fluxo estatisticamente significativo ao longo de todos os endereços IP, isto é, não há uma porta específica que seja varrida por todos os endereços. Entretanto, alguns endereços IP varrem todas as portas. A faixa mais clara à direita da figura representa os endereços IP iniciados por 200.

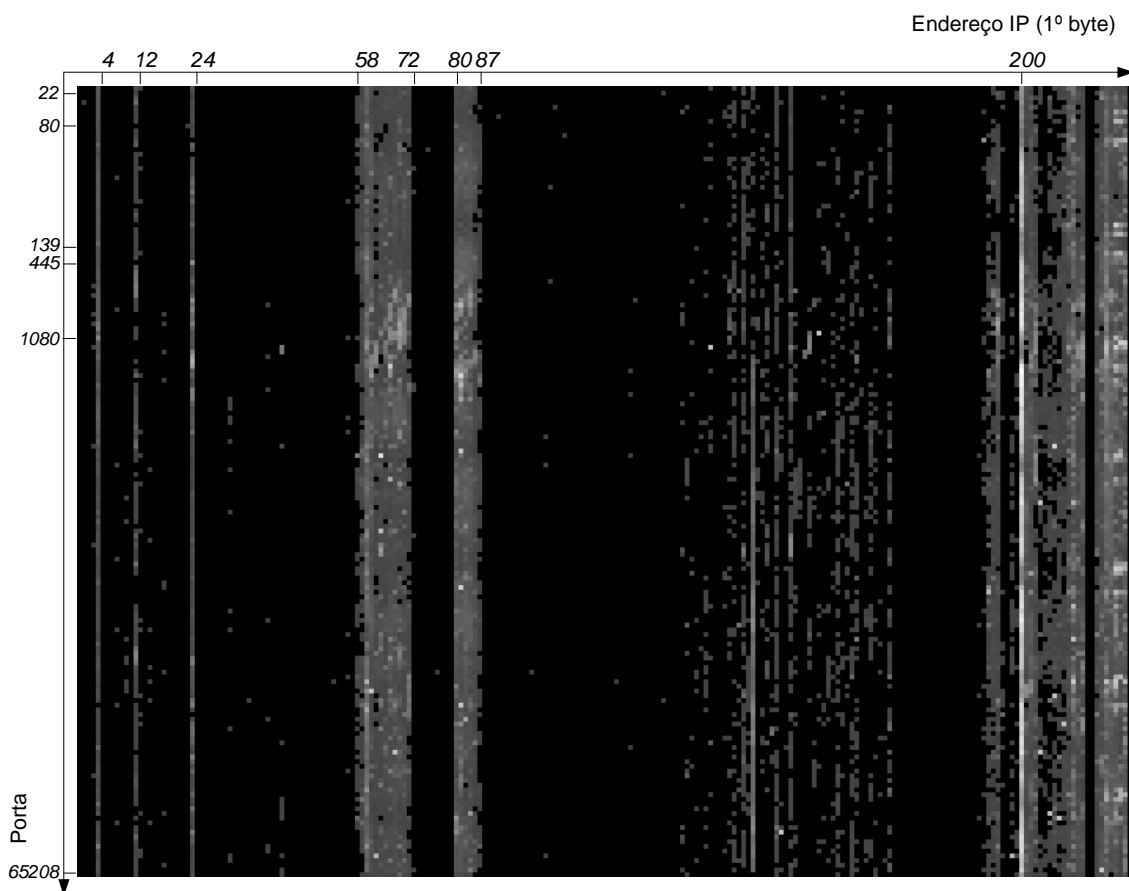


Figura 4.6 - Relacionamento entre endereços IP de origem e as portas de destino, janela de amostragem hora, fluxo TCP com menos de 3 pacotes, durante a primeira semana de janeiro de 2005

4.5 Comportamento dos fluxos TCP

A partir dos dados sanitizados foram somados todos os fluxos TCP com menos de 3 pacotes, de todos os sensores, em intervalos de uma hora para todos os dias de dados. Exemplificando: para o intervalo de 0:00 e 0:59 foram somados todos os fluxos, dos 23 sensores, dos 520 dias de dados e obtida a média.

A representação obtida permite verificar os horários em que as atividades maliciosas mais ocorrem. A representação dos fluxos TCP encontra-se ilustrada na Figura 4.7. Da análise da Figura verifica-se que, para os dados coletados, as atividades maliciosas caracterizadas pelos fluxos com menos de 3 pacotes se iniciam no início da tarde, por volta das 13:00, e vão aumentando de intensidade até atingir seu máximo às 19:00.

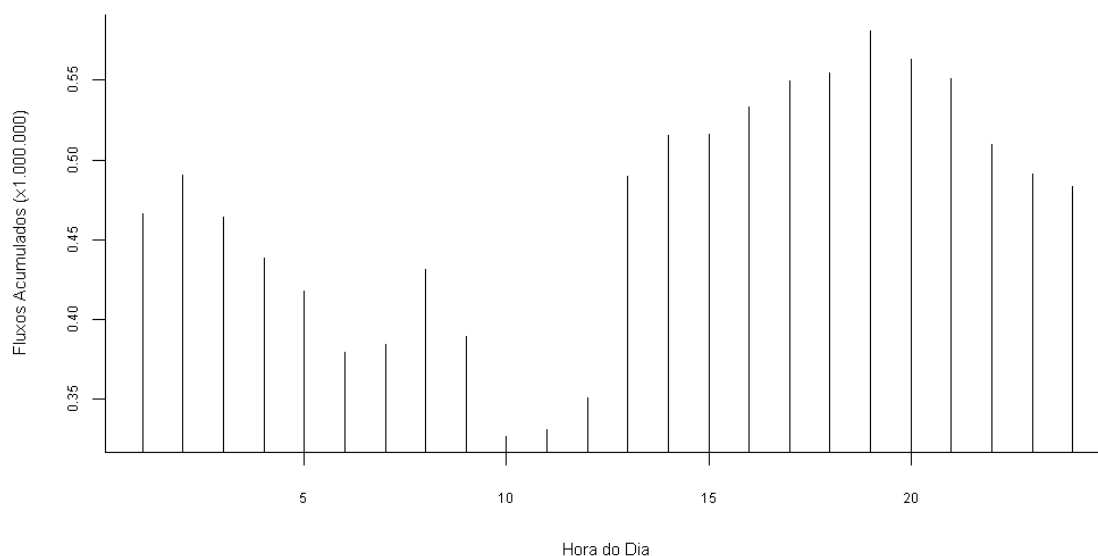


Figura 4.7 - Distribuição dos fluxos TCP com menos de 3 pacotes em relação aos horários do dia.

Na Figura 4.8 encontra-se análise para fluxos TCP com 3 ou mais pacotes. A análise da indica que o comportamento dos fluxos TCP com 3 ou mais pacotes é muito similar ao dos fluxos com menos de 3 pacotes. As atividades maliciosas se iniciam no início da tarde e tem um pico por volta das 19:00.

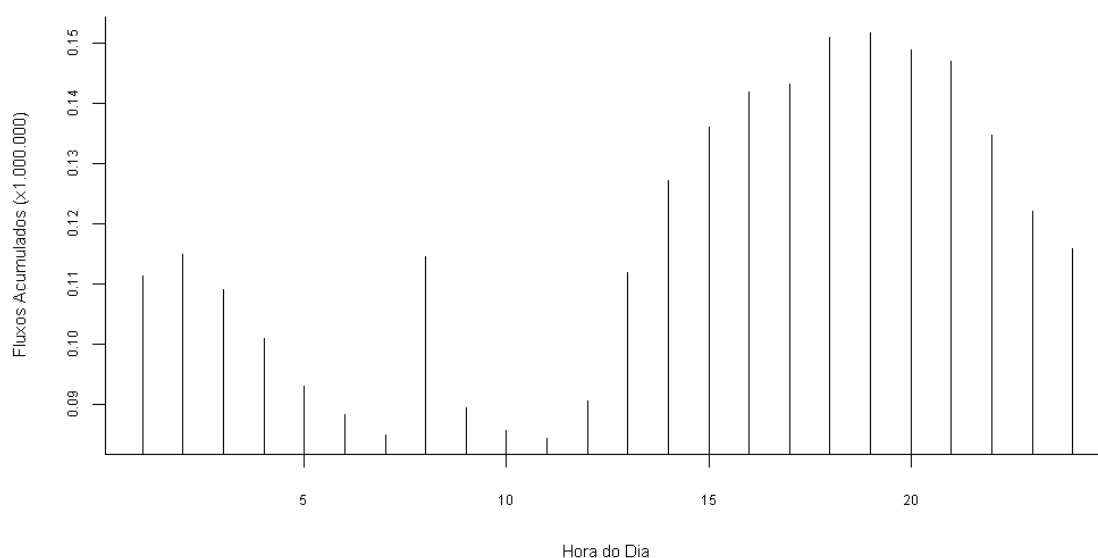


Figura 4.8 - Distribuição dos fluxos TCP com 3 ou mais pacotes em relação aos horários do dia.

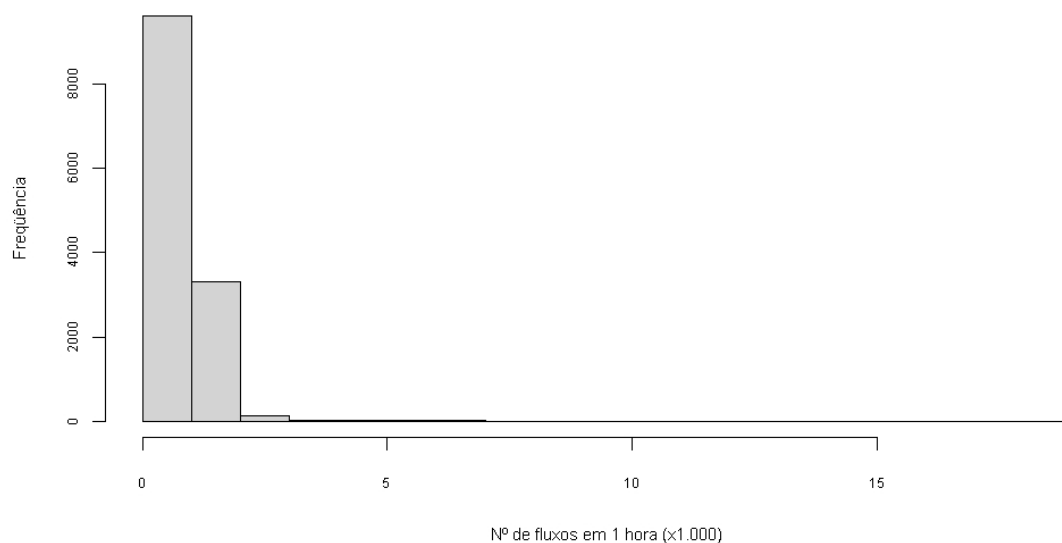


Figura 4.9 - Distribuição de frequência dos fluxos TCP com menos de 3 pacotes.

Outra análise que os dados coletados permite é olhar para a quantidade de fluxos de cada tipo que ocorre dentro da janela de amostragem hora. Esta representação, para fluxos TCP com menos de 3 pacotes encontra-se na Figura 4.9 e para fluxos TCP com 3 ou mais pacotes na Figura 4.10.

A Figura 4.9 mostra que, em uma hora, é mais comum que circulem até 1.000 fluxos TCP com menos de 3 pacotes. Já a Figura 4.10 mostra que, em uma hora, a quantidade de fluxos TCP com 3 ou mais pacotes se situa entre 2.000 e 4.000 fluxos. Porém, diferentemente do observado para os fluxos com menos de 3 pacotes, há diversas frequências com comportamento estatístico significativo.

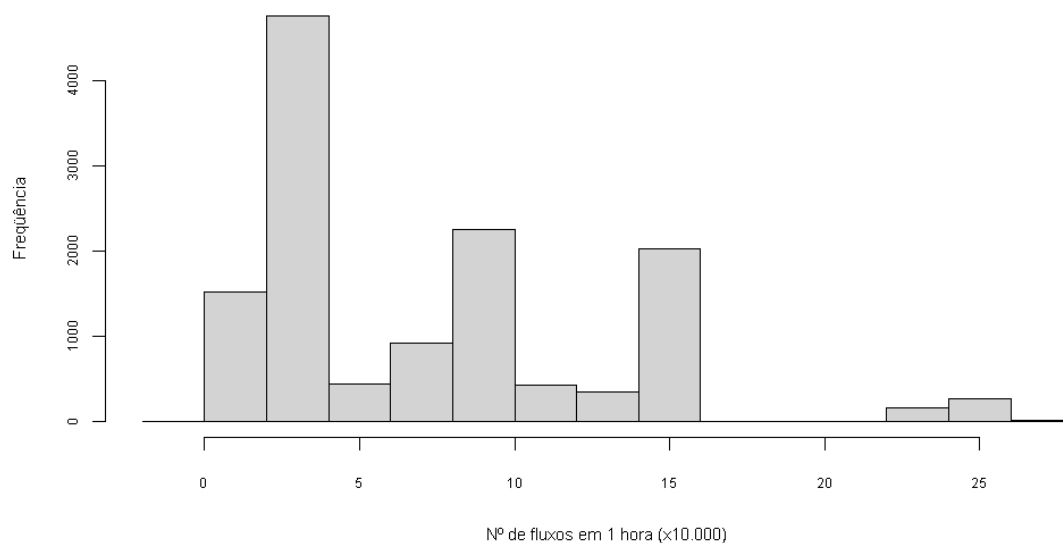


Figura 4.10 - Distribuição de frequência dos fluxos TCP com 3 ou mais pacotes.

4.6 Modelagem do ruído de fundo

Segundo Richardson et al (2005) um sensor não consegue, a princípio, distinguir dentro dos fluxos do ruído de fundo, qual é um pacote de ruído e o de uma atividade maliciosa. Para detectar atividades maliciosas o sensor deve observar um aumento do fluxo de pacotes para determinar que um código malicioso está presente e em atividade. Na sua experiência ele obteve uma média de 42 pacotes de varredura, a cada hora, com um desvio padrão de 57. A Tabela 4.5 sumariza os resultados obtidos nesta pesquisa com os dados do CBH.

Tabela 4.5 - Médias e desvios padrões para os fluxos TCP com menos e com mais de 3 pacotes em relação ao número de sensores

Média Em Relação A(o)	Janela			
	Dia		Hora	
	TCP < 3	TCP >= 3	TCP < 3	TCP >= 3
Sensor	19.580 ± 9.113	4.897 ± 4.328	855 ± 616	214 ± 245

A comparação entre os resultados obtidos por este trabalho e o de Richardson et al (2005) é inevitável até mesmo por causa da grande diferença entre os dois valores.

Como a experiência produzida por Richardson foi feita sobre redes telescópio verificando somente as varreduras que ocorreram ela é fundamentalmente diferente da realizada neste trabalho que usa honeypots em redes de produção com endereçamento reais.

Ainda de Richardson et all (2005) há uma proposta de modelagem matemática, em função do fluxo de entrada, para o ruído de fundo na Internet. O modelo considera que os pacotes que constituem o ruído são emitidos a uma taxa constante Φ , a cada janela de amostragem, e enviados para um dos possíveis endereços disponíveis na Internet de forma aleatória. Assim, a probabilidade de um sensor detectar um pacote destes é de $1/2^{32}$.

Para um sistema com k sensores a probabilidade de que um pacote seja detectado é $p_k = k/2^{32}$. Para uma determinada janela de amostragem o número de pacotes detectados é representado pela variável aleatória N . Esta variável pode ser representada por uma distribuição binomial de média $\mu_{ruído,k}$ e variância $(\sigma_{ruído,k})^2$, expressas, matematicamente pela Equação 4.2.

$$\begin{aligned}\mu_{ruído,k} &= \Phi * p_k = \frac{\Phi * k}{2^{32}} \\ (\sigma_{ruído,k})^2 &= \Phi * p_k * (1 - p_k) = \mu_{ruído,k} \left(1 - \frac{k}{2^{32}}\right)\end{aligned}\tag{4.2}$$

Para determinar o valor da taxa de emissão de pacotes Φ , na parcela brasileira da Internet usando dados do CBH deve-se, inicialmente, determinara a média de fluxos detectados por cada sensor, independentemente de protocolo uma vez que o ruído é constituído de todos os protocolos. Realizados os cálculos determina-se que a média horária detectada no CBH é de 19.286 com um desvio padrão de 11.999 fluxos. Deve-se lembrar de que estão sendo monitorados 4.176 endereços IP únicos. Assim, substituindo-se estes valores na Equação 4.2 tem-se a Equação numérica 4.3.

$$\Phi = \frac{\mu_{ruído,k} * 2^{32}}{k} = \frac{19.286 * 2^{32}}{4.176} * \frac{1}{3.600} = 5.509.840 \text{ fluxos por segundo}\tag{4.3}$$

Segundo Pang et all (2004) espera-se para a Internet uma taxa de emissão de 46.6 mi-

lhões de pacotes por segundo. Segundo Richardson et al (2005) este número é de 50,1 milhões de pacotes por segundo.

O valor determinado neste trabalho não é de pacotes mas sim de fluxos de acordo com a definição adotada para este trabalho. Se considerada a pior hipótese em que cada fluxo é constituído por 3 pacotes ter-se-ia uma taxa de 16,5 milhões de pacotes por segundo. Mesmo nesta situação o valor calculado ainda é muito inferior aos encontrados pelos demais pesquisadores.

Esta diferença pode ser atribuída a vários motivos como características de roteamento, de topologia da parcela brasileira, mas, fundamentalmente, da diferença de metodologia empregada.

5 ANÁLISE DE SÉRIES TEMPORAIS

Uma série temporal é um conjunto de observações, de um mesmo atributo de determinada população, efetuadas sequencialmente no tempo. Os atributos observados podem ser números $\{x_1, x_2, \dots, x_n\}$, que geram as séries temporais unidimensionais, ou vetores numéricos $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, que geram as séries temporais multidimensionais.

Uma série temporal é dita contínua caso suas observações sejam efetuadas ao longo do tempo sem interrupções. Se o conjunto de instantes em que o fenômeno é observado é discreto diz-se que a série temporal é discreta – se as observações são realizadas em espaços de tempo regulares (diário, mensal, trimestral, anual, etc.), mesmo que o tempo seja contínuo, as observações do fenômeno em intervalos de tempo, usualmente equidistantes, conduz a uma discretização desse conjunto.

Uma série temporal pode ser determinística ou estocástica. É determinística se as observações futuras podem ser previstas com exatidão a partir das observações passadas. É estocástica quando a previsão exata é impossível devido à aleatoriedade das observações; neste caso o futuro é probabilisticamente determinado pelo passado.

Para que se possa trabalhar com séries temporais há necessidade de se modelar o fenômeno sendo observado. Para tal é necessário reconhecer características especiais da série temporal que descreve o fenômeno em estudo.

O primeiro passo no reconhecimento destas características especiais é, normalmente, a representação gráfica. A observação do gráfico permite identificar estruturas representadas por diversos conjuntos de observações ao longo do tempo. Permite verificar a variação do fenômeno ao longo do tempo, a eventual existência de:

- a) tendência: a variação a longo prazo no nível médio das observações.

Segundo Statsoft (2010), a tendência representa um componente sistemático, linear ou não linear, que varia ao longo do tempo e não se repete, ao menos não se repete dentro do intervalo de tempo dos dados capturados;

- b) sazonalidade: conjuntos de observações que apresentam um padrão que se repete no tempo, isto é, a sazonalidade tem natureza similar à da tendência mas se repete a intervalos regulares ao longo do tempo.

Segundo Statsoft (2010), sazonalidade é a tendência de uma série temporal exibir um comportamento que se repete a cada L períodos. O termo ciclo é usado para representar o intervalo de tempo antes que o comportamento se repita. L é, portanto, o tamanho do ciclo em períodos.

A sazonalidade pode, ainda, ser classificada como aditiva ou multiplicativa. Aditiva quando no final de um período há um valor constante que deve ser acrescentado às observações. Multiplicativo quando o valor não é constante mas uma função do comportamento da série no último período; e,

- c) variabilidade: comportamento das observações ao longo do tempo e em torno do nível médio. Esta variação pode ser constante ou, principalmente em séries com tendência acentuada, variável.

Após uma primeira análise visual, empírica, inicia-se o processo analítico de análise de séries temporais.

Segundo Pollock (1999) a análise de séries temporais pode ser realizada a partir de duas abordagens distintas, mas não necessariamente excludentes:

- a) a abordagem no domínio do tempo; e,
- b) a abordagem no domínio das frequências.

A abordagem no domínio do tempo assume que a correlação entre pontos adjacentes é melhor explicada em termos da dependência do valor atual com valores passados. Procura modelar valores futuros como uma função paramétrica dos valores atual e passados.

A análise de séries temporais no domínio do tempo pode ser feita através de uma classificação sistemática de modelos chamada de média móvel auto-regressiva integrada (*autoregressive integrated moving average – ARIMA*) ou através da decomposição da série temporal transformando-a em uma soma, ou multiplicação, de outras séries.

A abordagem no domínio das frequências assume que a principal característica de interesse na análise da série temporal refere-se às variações periódicas e/ou senoidais, achadas na maioria dos dados reais, causadas por fenômenos físicos, biológicos ou ambientais.

5.1 Conceitos básicos

Segundo Shumway e Stoffer (2006) e Tsay (2002), uma série temporal estritamente estacionária é aquela em que o comportamento probabilístico de cada coleção de valores $\{x_{t1}, x_{t2}, \dots, x_{tk}\}$ é idêntico ao comportamento da coleção deslocada no tempo $\{x_{t1+h}, x_{t2+h}, \dots, x_{tk+h}\}$. Isto é matematicamente expresso pela Equação 5.1.

$$P(x_{t1} \leq c_1, \dots, x_{tk} \leq c_k) = P(x_{t1+h} \leq c_1, \dots, x_{tk+h} \leq c_k) \quad (5.1)$$

para todo $k = 1, 2, \dots$, para todos os tempos t_1, t_2, \dots, t_k , todos os números c_1, c_2, \dots, c_k , e todos os deslocamentos de tempo $h = 0, \pm 1, \pm 2, \dots$

Uma série temporal é dita fracamente estacionária se é um processo de variância finita e, ainda, atende:

- a) a função do valor médio, μ_t , definida pela Equação 5.2, é constante e não depende do tempo:

$$\mu_{xt} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx \quad (5.2)$$

- b) a função de covariância $\gamma(s, t)$, definida pela Equação 5.3, depende somente de s e t através da diferença $|s - t|$:

$$\gamma_{xt} = E[(x - \mu_s)(x_t - \mu_t)] \quad (5.3)$$

Segundo Shumway e Stoffer (2006), para todos os efeitos práticos as séries temporais fracamente estacionárias são consideradas estacionárias.

A função de autocorrelação (ACF) mede a previsibilidade linear existente entre o dado de uma série num instante t (x_t) a partir do dado num instante s (x_s). É definida através da Equação 5.4.

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}} \quad (5.4)$$

A ACF fornece um perfil das correlações lineares em todos os intervalos e mostra quais valores de t conduzem a uma melhor previsibilidade.

Se x_t pode ser perfeitamente previsto a partir de x_s através de uma relação linear do tipo $x_t = \beta_0 + \beta_1 x_s$, a correlação é 1.

A maioria das análises é realizada sobre dados reais ou seja, sobre pequenas amostras. Este número limitado de observações impede o cálculo da função de autocorrelação. Na prática o que se faz é a estimação do valor.

Se uma série temporal é estacionária sua média, que é um valor constante, pode ser estimada a partir da Equação 5.5.

$$\hat{x} = \frac{1}{n} \sum_{t=1}^n x_t \quad (5.5)$$

A autocorrelação, a partir da amostra, é estimada a partir da Equação 5.6.

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \quad (5.6)$$

A autocorrelação permite avaliar se os dados vêm de uma série aleatória ou se há corre-

lações estatisticamente significativas em determinados intervalos.

A função de autocorrelação parcial (PACF) de um processo estacionário, representado por Φ_{hh} , $h = 1, 2, \dots$, é definida pela Equação 5.7.

$$\begin{aligned}\phi_{11} &= \text{corr}(x_1, x_0) = \rho(1) \\ \phi_{hh} &= \text{corr}(x_h - x_h^{h-1}, x_0 - x_0^{h-1}), h \geq 2\end{aligned}\tag{5.7}$$

Na prática são poucas as séries temporais estacionárias. Na sua grande maioria as séries temporais apresentam componentes de tendência, de sazonalidade e componentes aleatórios. Isto é, o dado é constituído de padrões sistemáticos e de ruído aleatório.

5.2 Componentes de séries não estacionárias

Como não há um tratamento estatístico convencional para séries não estacionárias o que se pretende, nos casos em que as séries são não estacionárias, são métodos para salientar os padrões escondidos numa. Com os efeitos não estacionários minimizados pode-se aplicar as técnicas de análise das séries temporais estacionárias.

Segundo Statsoft (2010), a maioria dos padrões que compõem as séries temporais não-estacionárias podem ser descritos em termos de duas classes básicas de componentes: tendência e sazonalidade. Podem ocorrer, também, comportamentos não previsíveis. Elas contribuem para que a série seja não-estacionária e tenha um comportamento não linear. Nestes casos aplicam-se transformações matemáticas na tentativa de equalizar a variabilidade das observações ao longo da série.

5.2.1 Componente de tendência

Segundo Shumway e Stoffer (2006), o modelo de tendência estacionária pode ser, matematicamente, descrito pela Equação 5.8.

$$x_t = \mu_t + y_t\tag{5.8}$$

onde

x_t são as observações,

μ_t a tendência, e

y_t o processo estacionário.

É comum que, numa série que apresente um componente de tendência este venha a esconder o comportamento estacionário da série. Assim, é interessante que se remova o componente de tendência antes de analisar a série.

Para removê-lo é necessário que se estime um valor para a tendência e que se remova este componente através da formulação descrita na Equação 5.9.

$$\hat{y}_t = x_t - \hat{\mu}_t \quad (5.9)$$

A estimação do componente de tendência pode ser feito a partir de uma das seguintes técnicas:

- a) regressão (*function fitting*): adequação dos dados temporais a funções lineares. Eventualmente funções logarítmicas, exponenciais ou polinomiais podem ser usadas;
- b) diferenciação: subtração de elementos de uma série temporal afastados de um intervalo (*lag*); ou,
- c) suavização (*smoothing*): aplicação de alguma forma de média local aos dados de tal forma que os componentes não sistemáticos das observações individuais se cancelem.

5.2.1.1 Regressão

No modelo de regressão clássica, se uma série temporal x_t , $t = 1, 2, \dots, n$, sofre influência de uma coleção de séries independentes $z_{t1}, z_{t2}, \dots, z_{tq}$, a relação entre ambas pode ser

expressa através da Equação 5.10.

$$x_t = \beta_1 z_{t_1} + \beta_2 z_{t_2} + \dots + \beta_q z_{t_q} + w_t \quad (5.10)$$

onde

cada β_j é um coeficiente da regressão, e

w_t é um erro aleatório ou um ruído.

5.2.1.2 Diferenciação

Se o modelo para tendência usado é o descrito pela Equação 5.8 pode-se obter um processo estacionário se executada a operação descrita na Equação 5.11.

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \mu_t - \mu_{t-1} + y_t - y_{t-1} = \delta + w_t + y_t - y_{t-1} \quad (5.11)$$

No processo da diferenciação não há necessidade de estimar parâmetros como no método da regressão. Entretanto, não se consegue estimar o processo estacionário. Se há necessidade de conhecer o processo estacionário a regressão é a melhor alternativa; se o que se deseja é a transformação de uma série num processo estacionário a diferenciação é mais indicada.

Como diferenciação desempenha um papel muito importante nas séries temporais ela recebe seu próprio operador, ∇ . A primeira diferença é representada pela Equação 5.12.

$$\nabla x_t = x_t - x_{t-1} \quad (5.12)$$

A primeira diferença é usada para eliminar tendências lineares. Outras tendências podem ser eliminadas com a aplicação de diferenças de ordens mais altas.

A segunda diferença é a diferença da primeira diferença: $\nabla(\nabla x_t)$. Algebricamente esta operação encontra-se expressa na Equação 5.13.

$$\begin{aligned}\nabla(\nabla x_t) &= \nabla(x_t - x_{t-1}) = \nabla(x_t) - \nabla(x_{t-1}) \\ (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) &= x_t - 2x_{t-1} + x_{t-2}\end{aligned}\quad (5.13)$$

Para facilitar a representação das diferenças de ordens mais altas usa-se um operador chamado de operador de retardo que é definido por $Bx_t = x_{t-1}$. A característica algébrica deste operador encontra-se descrita na Equação 5.14.

$$B^2 x_t = B(Bx_t) = Bx_{t-1} = x_{t-2} \Rightarrow B^k x_t = x_{t-k} \quad (5.14)$$

A relação entre os operadores ∇ e B é dada por $\nabla^k = (1 - B)^k$. Se desenvolvida a expressão encontra-se a formulação apresentada na Equação 5.15.

$$\begin{aligned}\nabla x_t &= (1 - B)x_t = x_t - Bx_t = x_t - x_{t-1} \\ \nabla^2 x_t &= (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}\end{aligned}\quad (5.15)$$

5.2.1.3 Suavização/Filtragem

Esta técnica também pode ser aplicada na modelagem de comportamentos sazonais e/ou cíclicos.

Está sendo aplicado o nome de suavização juntamente com o de filtragem uma vez que, a formulação base é a mesma. Entretanto, há uma diferença conceitual no uso de cada um dos termos. A suavização tem como objetivo achar o valor num determinado ponto usando, se necessário, toda a amostra incluindo o que ocorre após o instante desejado. A filtragem, por sua vez, se propõe a determinar um valor suavizado para um determinado instante usando somente informações passadas e é melhor adaptada para previsões.

Assim, se o interesse for somente o formato da curva formada pela série temporal o mais indicado é usar suavização ao passo que, se o esperado é entender o que está acontecendo no final da série para que projeções futuras possam ser realizadas a filtragem é mais adequada.

Existem dois grandes grupos de funções que podem ser usadas para suavização, as fun-

ções de suavização pela média e as funções de suavização exponencial. Nas funções pela média as observações passadas são igualmente ponderadas ao passo que nas exponenciais os pesos são decrescentes à medida que as observações são mais antigas.

Estes dois grandes grupos estão ilustrados abaixo por algumas de suas funções:

a) Suavização pela média:

- Média móvel simples (filtragem);

$$\hat{x}_t = \frac{1}{N} \sum_{i=0}^{N-1} x_{t-i} \quad (5.16)$$

- Média móvel central ponderada (suavização);

$$\hat{x}_t = \sum_{j=-k}^k a_j x_{t-j}, \quad a_j = a_{-j} \geq 0 \quad e \quad \sum_{j=-k}^k a_j = 1 \quad (5.17)$$

b) Suavização exponencial:

- Suavização exponencial dupla;

$$\begin{aligned} \hat{x}_t &= \alpha x_{t-1} + (1-\alpha)(\hat{x}_{t-1} + b_{t-1}), \quad 0 \leq \alpha \leq 1 \\ b_t &= \beta(\hat{x}_t - \hat{x}_{t-1}) + (1-\beta)b_{t-1}, \quad 0 \leq \beta \leq 1 \end{aligned} \quad (5.18)$$

5.2.2 Componente de sazonalidade

Segundo Pollock (1999), sazonalidade, ou dependência sazonal, é medida pela autocorrelação e representa a dependência, ou correlação, de ordem k – intervalo entre as amostras da série, também chamado de *lag* – entre cada i -ésimo elemento e o $(i-k)$ -ésimo elemento da série.

O comportamento periódico pode ser modelado, dentre outras possibilidades, através

de:

- a) diferenças de ordem superior a 2; ou,
- b) suavização.

5.2.2.1 Diferenciação

Seja $X_t = S_t + \varepsilon_t$, onde S_t é uma série de período d . Portanto, X_t depende de X_{t-d} , talvez de X_{t-2d} , e assim por diante.

Devido ao componente sazonal a primeira diferença não é suficiente para tornar a série estacionária. Para tal tomam-se diferenças da ordem do período sazonal. Uma primeira diferença sazonal é denotada por ∇_d , que pode ser expressa pela Equação 5.19.

$$\nabla_d X_t = X_t - X_{t-d} \quad (5.19)$$

onde d é o período sazonal.

A técnica de remoção de sazonalidades pode ser combinada com a de remoção de tendências. Se, X_t é uma série temporal com componente sazonal de período d e tendência polinomial de grau k , aplica-se sucessivamente as duas transformações para remover a tendência e sazonalidade tal qual exemplificado pela Equação 5.20.

$$\nabla^k(\nabla_d X_t) \quad (5.20)$$

5.2.2.2 Suavização

A suavização exponencial tripla é usada quando os dados apresentam tendência e sazonalidade. Existem dois modelos possíveis que dependem do tipo de sazonalidade:

- a) modelo sazonal multiplicativo, ou,

b) modelo sazonal aditivo.

No modelo sazonal aditivo, o adotado por este trabalho, a série temporal é representada pela Equação 5.21.

$$X_t = \hat{a} + \hat{b}t + S_{t+1+(h-1) \bmod p} + \epsilon_t \quad (5.21)$$

onde,

\hat{a} é o sinal base, de nível ou constante;

\hat{b} é o componente de tendência;

S_t é componente de sazonalidade; e,

ϵ_t é o componente de erro aleatório.

As equações usadas para estimação dos parâmetros estão apresentadas na Equação 5.22:

$$\begin{aligned} \hat{a}_t &= \alpha(x_t - S_{t-L}) + (1-\alpha)(\hat{a}_{t-1} + \hat{b}_{t-1}) \\ \hat{b}_t &= \beta(\hat{a}_t - \hat{a}_{t-1}) + (1-\beta)\hat{b}_{t-1} \\ S_t &= \gamma(x_t - \hat{a}_t) + (1-\gamma)S_{t-p} \end{aligned} \quad (5.22)$$

5.2.3 Componentes estruturados

5.2.3.1 Processos auto-regressivos

Ainda segundo Shumway e Stoffer (2006), modelos de auto-regressão baseiam-se na ideia de que valores atuais podem ser explicados em função de p valores passados, isto é, p determina o número de passos que se precisa andar para o passado para prever o valor atual.

O modelo auto-regressivo de ordem p (AR(p)) é expresso pela Equação 5.23.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t \quad (5.23)$$

onde

x_t é uma série estacionária,

ϕ_1, \dots, ϕ_p são constantes, e

w_t é um ruído branco ou gaussiano.

O modelo da média móvel de ordem q (MA(q)) combina, linearmente, o ruído para obter o dado observado como expresso pela Equação 5.24.

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q} \quad (5.24)$$

Uma série temporal é uma média móvel auto-regressiva (ARMA(p, q)) se é estacionária e pode ser expressa pela Equação 5.25.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q} \quad (5.25)$$

Segundo Guimarães (2008) a condição necessária para a aplicação do modelo de auto-regressão é que a série temporal seja estacionária, ou seja, que sua média e variância sejam constantes no tempo.

Nas séries temporais que não possuem suas médias e variâncias constantes ao longo do tempo há necessidade de transformá-las em séries estacionárias. Uma possível transformação consiste em determinar diferenças sucessivas da série original até obter uma série estacionária.

O número d de diferenciações para tornar a série estacionária é chamado de ordem de integração. A inclusão deste processo, permite a utilização do modelo auto-regressivo integrado a média móvel (ARIMA(p, d, q)).

O modelo $ARIMA(p, d, q)$, processo integrado de médias móveis, corresponde ao modelo $ARMA(p, q)$ quando a série temporal original é substituída pela sua d -ésima diferença. Assim, quando $d = 0$ o modelo $ARIMA(p, 0, q)$ é equivalente ao modelo $ARMA(p, q)$.

Ainda segundo Guimarães (2008) são necessários três processos para se obter um modelo adequado:

- a) Identificação: consiste em determinar qual dos modelos de Box e Jenkins é o mais adequado. Duas ferramentas são usadas para medir a correlação entre as observações dentro de uma série temporal: são as funções de autocorrelação (ACF) e autocorrelação parcial (PACF).
 - A função de autocorrelação (ACF) descreve a correlação entre duas observações adjacentes da mesma série temporal, ocorridos em diferentes períodos;
 - A função de autocorrelação parcial (PACF) mede o grau de associação entre as observações de uma série temporal quando o efeito de defasagem é retirado;
- b) Estimação: consiste em estimar os parâmetros auto-regressivo (ϕ) e média móvel (θ);
- c) Verificação: avaliar a consistência dos modelos através da análise dos resíduos. As análises das funções de autocorrelação (ACF) e autocorrelação parcial (PACF) dos resíduos devem apresentar um comportamento aleatório.

5.2.3.2 SARIMA

São processos auto-regressivos integrados sazonais, isto é, tanto as partes $AR(p)$ como $MA(q)$ podem ser sazonais. São representados por $(p, d, q) \times (P, D, Q)_s$ e descritos pelo modelo apresentado na Equação 5.26.

$$\phi(B)\phi(B^s)(1-B)^d(1-B^s)^D X(t) = \theta(B)\theta(B^s)Z(t) \quad (5.26)$$

onde s é o período,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \quad \text{é o polinômio AR}(p),$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q \quad \text{é o polinômio MA}(q),$$

$$\phi(B^s) = 1 - B_1 B^s - B_2 B^{2s} - \dots - B_p B^{ps} \quad \text{é o polinômio sazonal AR},$$

$$\theta(B^s) = 1 + A_1 B^s + A_2 B^{2s} + \dots + A_Q B^{Qs} \quad \text{é o polinômio sazonal MA, e}$$

$Z(t)$ é um ruído branco.

6 PREDIÇÃO DE EVENTOS USANDO SÉRIES TEMPORAIS

Segundo Pollock (1999), há dois ramos da teoria de predição usando séries temporais:

- a) a predição pura cujo objetivo é prever o valor de uma série temporal em um instante futuro com base na série já observada; e
- b) a extração do sinal cujo objetivo é inferir o valor de um sinal a partir de um registro ao qual é sobreposto um ruído. Neste caso pode-se querer a estimação do sinal a qualquer instante, isto é, passado, presente ou futuro.

Ainda segundo Pollock (1999) a base de toda a teoria de previsão vem dos trabalhos de Wiener e Kolmogorov cuja base teórica se restringe aos processos estacionários clássicos. Entretanto, a maioria dos problemas práticos tratam séries que não são estacionárias. A teoria de Wiener e Kolmogorov pode ser adaptada para tratar com estas séries reduzindo-as para um estado estacionário através do uso de operadores e filtros.

Segundo Shumway (2006), o objetivo mais importantes da análise de séries temporais é a previsão de valores futuros com base nos valores passados, usando ou não um modelo ajustado.

Se T é o tempo presente pretende-se determinar os valores de X_{t+1} , X_{t+2} , ..., X_{t+k} , com base no conhecimento de X_t , X_{t-1} , X_{t-2} ; Se $\hat{x}_T(\tau)$ é a previsão no instante $X_{T+\tau}$ seu valor é conhecido por valor médio condicional e é função dos valores passados como representado na Equação 6.1.

$$\hat{x}_t(\tau) = E(x_{T+\tau} | x_t, x_{t-1}, \dots) \quad (6.1)$$

A função representada na Equação 6.1 é chamada de função de previsão. O instante T é a origem da previsão e o inteiro τ o horizonte da previsão. Vários modelos podem ser usados para previsão. A escolha do modelo mais adequado se dá pela análise dos erros de previsão.

Os modelos clássicos mais divulgados na literatura para realização de previsões a partir de séries temporais são: média móvel, suavização exponencial dupla, suavização exponencial com tendência, suavização exponencial com tendência e sazonalidade e os modelos de Box-Jenkins.

6.1 Erros

Sendo o tempo atual T e conhecidos os dados dos tempos 1 até T , prever um resultado futuro, τ instantes à frente do tempo atual, é tentar achar o valor do dado no instante $T + \tau$. O que se espera é realizar uma boa estimação da variável aleatória neste instante de tempo porque, por mais aderente que uma série temporal seja ao modelo que se está usando para descrevê-la, o ruído é aleatório e, portanto, não predizível.

Para analisar se uma previsão é boa analisam-se os erros associados à previsão, isto é, a diferença entre a observação e o valor estimado tal como representado na Equação 6.2.

$$\varepsilon_{\tau} = x_{T+\tau} - \hat{x}_{T+\tau} \quad (6.2)$$

Um possível critério para analisar o desempenho de um estimador ou preditor $\hat{x}_{T+\tau}$ de uma variável aleatória x é o erro médio quadrático (*mean square error* – MSE). Ele pode ser matematicamente descrito através da Equação 6.3.

$$MSE = \frac{\sum_{j=1}^n \varepsilon_j^2}{n} \quad (6.3)$$

onde n são o número de observações.

Outro possível critério é através do erro médio absoluto (*mean absolute deviation* – MAD) que é expresso através da Equação 6.4.

$$MAD = \frac{\sum_{j=1}^n |\varepsilon_j|}{n} \quad (6.4)$$

6.2 Médias móveis

A utilização do modelo de média móvel ocorre quando a série temporal não apresenta tendência ou sazonalidade, fato muito raro nos dados reais.

A previsão com médias móveis é feita usando-se o modelo constante apresentado na Equação 6.5.

$$\hat{x}_{T+\tau} = \hat{b}_T \text{ para } \tau = 1, 2, \dots \quad (6.5)$$

O parâmetro b pode ser estimado como a média das últimas m observações. Se usada a média simples seu valor é obtido a partir da Equação 6.6.

$$\hat{b}_T(m) = \frac{1}{m} \sum_{i=T-m+1}^T x_i \quad (6.6)$$

Se, para estimação do parâmetro b for usada a média ponderada seu valor é obtido pela Equação 6.7.

$$\hat{b}_T(m) = \frac{1}{\sum_{i=1}^m w_i} \sum_{i=T-m+1}^T w_i x_i \quad (6.7)$$

O número de observações utilizadas determina o grau de sensibilidade deste modelo em relação aos dados mais recentes.

6.3 Suavização exponencial

O modelo de suavização exponencial dupla é adequado para séries temporais que apresentam tendência, mas não apresentam sazonalidade. Há mais de um modelo de suavização exponencial dupla. Será apresentado o Modelo de Holt.

A predição é feita usando-se a Equação 6.8.

$$\hat{x}_{T+\tau} = a_T + \tau b_T \quad (6.8)$$

O coeficiente linear (a_T) representa a estimativa inicial da média, ou nível, e o coeficiente angular (b_T) é estimativa inicial da tendência. As Equações utilizadas para determinação destes coeficientes estão apresentadas na Equação 6.9.

$$\begin{aligned} a_T &= \alpha x_T + (1 - \alpha)(a_{T-1} + b_{T-1}), \quad 0 \leq \alpha \leq 1 \\ b_T &= \beta(a_T - a_{T-1}) + (1 - \beta)b_{T-1}, \quad 0 \leq \beta \leq 1 \end{aligned} \quad (6.9)$$

Como se vê nas formulações apresentadas em 6.9 os coeficientes linear e angular dependem dos parâmetros α e β , que devem ser estimados, e do valor da observação no instante (x_T).

Para simplificar o processo de estimação dos parâmetros α e β usa-se um único parâmetro de estimação, δ , que se relaciona com α e β através da formulação apresentada na Equação 6.10.

$$\alpha = 1 - (1 - \delta)^2, \quad \beta = \frac{\delta^2}{1 - (1 - \delta)^2} \quad (6.10)$$

A suavização exponencial com tendência e sazonalidade, também chamada de suavização tripla, é adequado às séries temporais que apresentem os componentes de nível, de tendência e de sazonalidade. Será apresentado o modelo aditivo de Holt-Winter.

A predição é feita usando-se a Equação 6.11.

$$\hat{x}_{T+\tau} = a_T + \tau b_T + S_{T-L+\tau} \quad (6.11)$$

onde a é o sinal base, de nível ou constante, b é o componente de tendência e S_T é componente de sazonalidade. L representa o tamanho do ciclo.

As Equações utilizadas para determinação destes coeficientes estão apresentadas na Equação 6.9.

$$\begin{aligned}
a_T &= \alpha(x_T - S_{T-L}) + (1-\alpha)(a_{T-1} + b_{T-1}) \\
b_T &= \beta(a_T - a_{T-1}) + (1-\beta)b_{T-1} \\
S_T &= \gamma(x_T - a_T) + (1-\gamma)S_{T-L}
\end{aligned} \tag{6.12}$$

6.4 Modelos auto-regressivos

Segundo Box et al. (2008) os modelos de Box-Jenkins, também conhecidos como ARIMA (Autoregressive Integrated Moving Average) tem a finalidade de encontrar uma função que descreva uma série temporal e que permita fazer previsões.

Os modelos ARIMA são formados por três componentes: o auto-regressivo (AR), o de integração (I) e o de média móvel (MA). A série pode ser modelada para cada um dos componentes ou por combinações entre eles, resultando em vários modelos, como já mostrado em 5.2.3.1 .

Sendo a equação do modelo ajustado conhecida, a previsão do valor de $\hat{x}_T(\tau)$ é obtida substituindo valores futuros dos erros por zero e valores futuros da série X_{T+1} , X_{T+2} , ... pelo valor médio condicional. Os valores anteriores a X_T e os erros são substituídos pelas observações ajustadas.

Para exemplificar considere o modelo SARIMA(1; 0; 0) x (0; 1; 1)₁₂. Substituindo-se os parâmetros em 5.26 tem-se a Equação 6.13.

$$(1 - \phi B)(1 - B^{12})X_t = (1 + AB^{12})Z_t \tag{6.13}$$

Se desenvolvida algebricamente tem-se a formulação expressa na Equação 6.14.

$$\begin{aligned}
(1 - B^{12} - \phi B + \phi B^{13})X_t &= Z_t + AB^{12}Z_t \\
X_T &= X_{T-12} + \phi(X_{T-1} - X_{T-13}) + Z_T + AZ_{T-12}
\end{aligned} \tag{6.14}$$

As previsões são dadas pelas Equações 6.15.

$$\begin{aligned}\hat{x}_T(1) &= x_{T-11} + \phi(x_T - x_{T-12}) + Az_{T-11} \\ \hat{x}_T(2) &= x_{T-10} + \phi(\hat{x}_T(1) - x_{T-11}) + Az_{T-10}\end{aligned}\tag{6.15}$$

No caso dos modelos AR(p) o algebrismo apresentado nas Equações 6.16

$$\begin{aligned}(1 - \phi B - \phi_2 B^2 - \dots - \phi_p B^p) X_t &= Z_t \\ X_T - \phi X_{T-1} - \phi_2 X_{T-2} - \dots - \phi_p X_{T-p} &= Z_t\end{aligned}\tag{6.16}$$

gera as funções de previsão descritas nas Equações 6.17.

$$\begin{aligned}\hat{x}_T(1) &= \phi_1 x_T + \phi_2 x_{T-1} + \dots + \phi_p x_{T-p+1} \\ \hat{x}_T(2) &= \phi_1 \hat{x}_T(1) + \phi_2 x_{T-1} + \dots + \phi_p x_{T-p+2} \\ &\dots \\ \hat{x}_T(p+1) &= \phi_1 \hat{x}_T(p) + \phi_2 \hat{x}_T(p-1) + \dots + \phi_p \hat{x}_T(1)\end{aligned}\tag{6.17}$$

Realizando-se um algebrismo parecido com o feito em 6.16, para o caso de modelos MA(q), tem-se as funções de previsão conforme as Equações 6.18.

$$\begin{aligned}\hat{x}_T(1) &= \theta_1 z_T + \theta_2 z_{T-1} + \dots + \theta_q z_{T-q+1} \\ \hat{x}_T(2) &= \theta_2 z_T + \dots + \theta_q z_{T-q+2} \\ &\dots \\ \hat{x}_T(q) &= \theta_q z_T \\ \hat{x}_T(q+j) &= 0 \quad , \quad j=1, 2, 3, \dots\end{aligned}\tag{6.18}$$

7 ANÁLISE DE SÉRIES TEMPORAIS APLICADA AOS DADOS DO CBH

As técnicas de análise de séries temporais foram aplicadas aos dados do CBH usando-se janelas de amostragem dia e hora através do pacote R¹ para análise de dados estatísticos, na sua versão 2.11.1.

7.1 Visualização dos dados do CBH

Os dados recebidos dos sensores são colocados em um arquivo que contém, para a janela de amostragem selecionada, a quantidade de fluxos TCP, UDP e ICMP de todos os sensores.

A Figura 7.1 mostra, como já afirmado anteriormente, que os sensores capturam uma quantidade muito maior de fluxos TCP do que os demais protocolos. A partir deste instante, neste trabalho, salvo quando explicitamente comentado, todo tratamento será feito sobre dados do protocolo de transporte TCP.

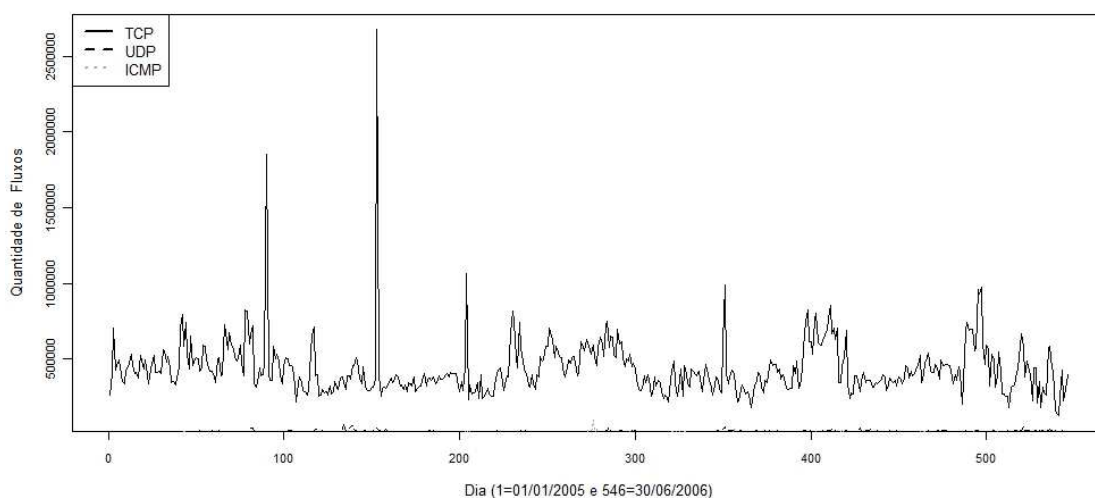


Figura 7.1 - Fluxo dos protocolos TCP, UDP e ICMP, janela de amostragem dia, como série temporal.

1 É um ambiente para realização de cálculos estatísticos e geração de gráficos que pode ser complicado em várias plataformas Unix, Windows and MacOS. O sítio principal é <http://www.r-project.org/>

Na análise de séries temporais deve-se conhecer, e preferencialmente tratar, a sazonalidade. Será testada a hipótese de que há sazonalidade nos dados de fluxo médio com janela de amostragem dia. Para tal é feita a representação dos dados num gráfico do tipo *boxplot* com os dados agrupados a cada 7 dias (1 semana).

Boxplot ou diagrama de *Box* e *Whisker* permite a representação de um sumário completo dos dados apresentando os 1º, 2º (mediana) e 3º quartis¹ da distribuição através de um retângulo que corresponde aos 50% valores centrais da distribuição. Usando o valor do intervalo inter-quartil (IIQ) – que corresponde à distância entre os 1º e 3º quartis – traça dois segmentos externos ao retângulo localizados a $Q3 + (1,5) \text{ IIQ}$ e $Q1 - (1,5) \text{ IIQ}$. Observações acima, ou abaixo, destes limites são chamadas de pontos exteriores (*outliers*) e representadas isoladamente.

A análise intuitiva da Figura 7.2 parece corroborar com a hipótese de que há uma sazonalidade e que o ciclo de 7 dias parece conduzir a uma senoidal, com diferentes amplitudes, porém, com a mesma frequência.

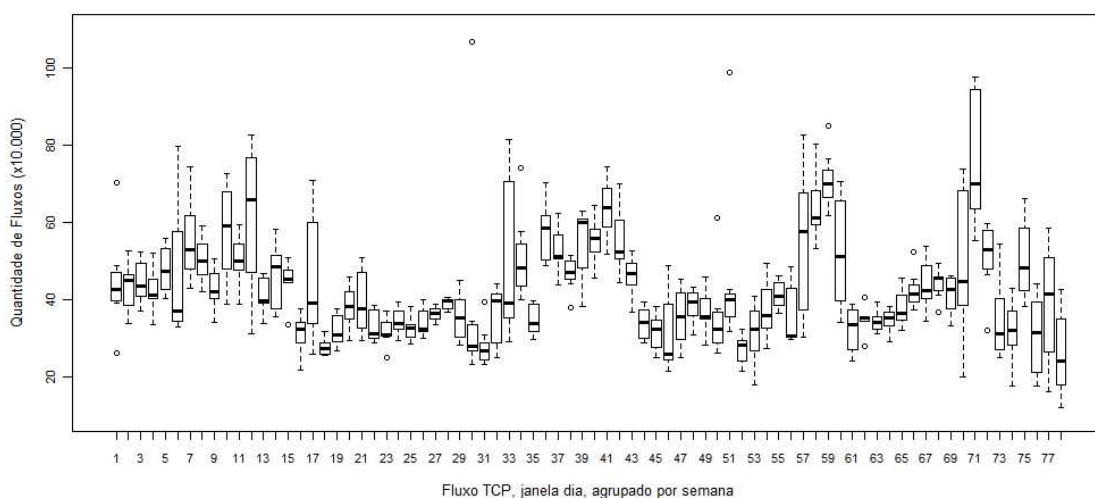


Figura 7.2 - *Boxplot* do fluxo do protocolo TCP, janela de amostragem dia, agrupado a cada 7 dias.

1 Quartil é qualquer um dos três valores que divide o conjunto ordenado de dados em quatro partes iguais. Cada parte representa 1/4 da amostra ou da população.

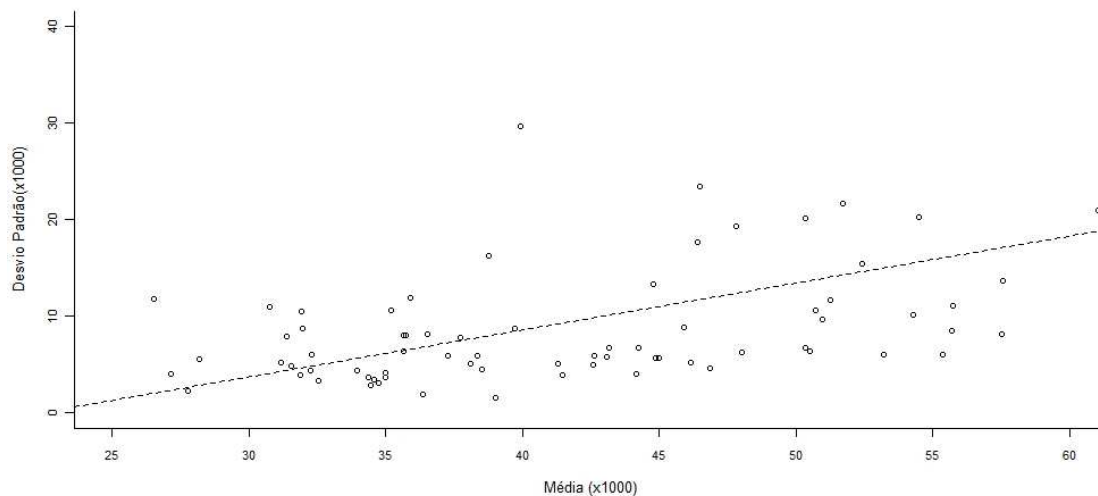


Figura 7.3 - Média versus desvio padrão para agrupamentos de fluxos médios TCP, janela de amostragem dia, a cada 7 dias

Pode-se, ainda, verificar se os dados carregam, de forma intrínseca, alguma característica que nos permita aplicar alguma transformação nos mesmos de tal forma que nossa representação e consequente análise fique favorecida.

Uma opção para divisão do conjunto de dados é dada por Barros (2008) calculando-se as médias e os desvios padrão a cada 7 observações. Os valores das médias e dos desvios padrão são representados em um gráfico.

A Figura 7.3 mostra uma associação linear entre a média e o desvio padrão. Quando ocorre este tipo de associação, isto é, sempre que o desvio-padrão dos dados varia linearmente com a média, pode-se usar a transformação logarítmica para estabilizar variância.

Ao se aplicar a transformação logarítmica a nova representação dos dados, para a janela de amostragem dia, é dada pela Figura 7.4.

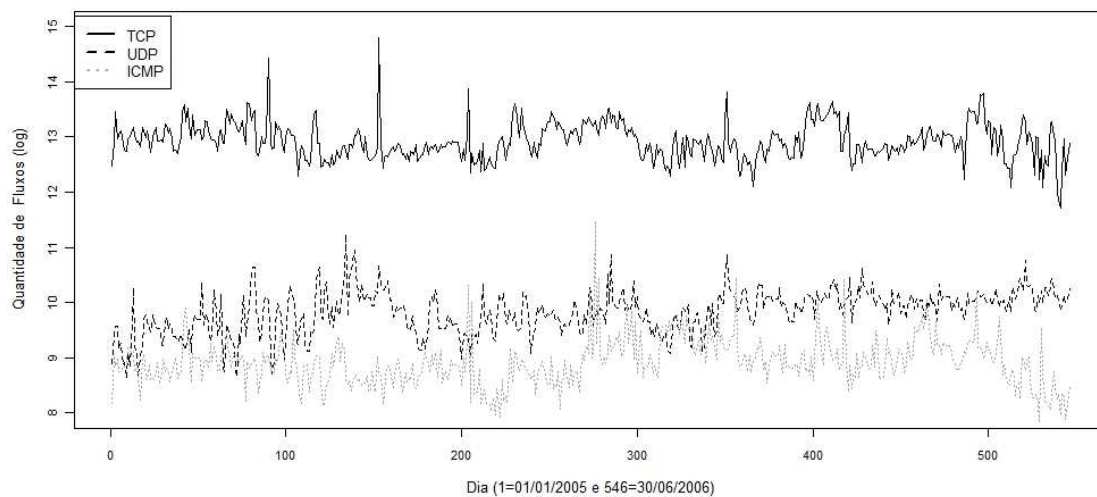


Figura 7.4 - Logaritmo dos fluxos dos protocolos TCP, UDP e ICMP, janela de amostragem dia

A representação dos dados para a janela de amostragem hora, já aplicada a transformação logarítmica é dada pela Figura 7.5.

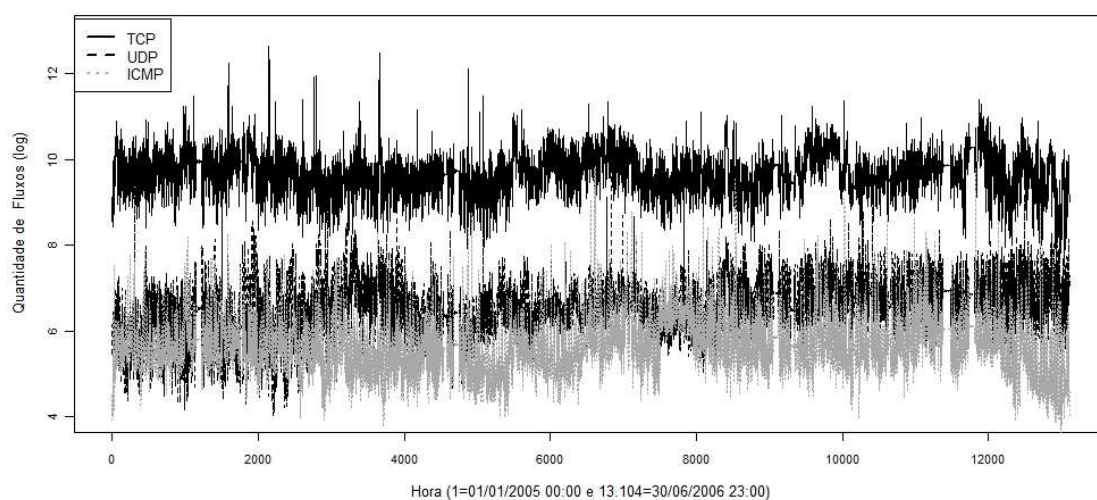


Figura 7.5 - Logaritmo dos fluxos dos protocolos TCP, UDP e ICMP, janela de amostragem hora

7.2 Análise do Fluxo TCP

O que se espera, ao aplicar as técnicas de análise de séries temporais nos fluxos TCP, é encontrar um modelo matemático que descreva os dados como a soma, ou multiplicação, de um conjunto de séries bem estruturadas e bem conhecidas mais um ruído. E, ainda, se for possível a aplicação de um modelo estatístico, a realização de predição de resultados.

Pelo que já foi discutido anteriormente o conjunto de dados de entrada será aquele que já sofreu a aplicação de uma transformação logarítmica.

Segundo Barros (2008) a análise é feita através de:

- a) correlograma: exhibe a ACF e mede a dependência linear existente entre o dado de uma série num instante t (x_t) e outro de um instante s (x_s). Se x_t pode ser previsto a partir de x_s através de uma relação linear do tipo $x_t = \beta_0 + \beta_1 x_s$, a correlação é 1.

Verifica a dinâmica linear dos dados. Pode ser usado, ainda, para diagnosticar se uma determinada série é um ruído branco. Nesta situação todos os ACFs são zero ou muito próximos de zero;

- b) correlograma parcial: exhibe a PACF e determina se a série é um processo auto-regressivo genuíno ($AR(p)$). Ela é dita um $AR(p)$ tem-se que todos os PACFs são próximos de zero para todo $k > p$;
- c) densidade espectral, ou *spectrum*: captura o conteúdo de frequência de um processo estocástico e ajuda a identificar periodicidades. Estas aparecem destacadas no gráfico após a estimação da densidade espectral através da exibição de picos nas frequências correspondentes a estes períodos;
- d) p-valores: são obtidos a partir do teste de Ljung-Box caso se comprove que a sé-

rie possui ruído gaussiano e somente nesta situação são exibidos. Lung-Box é um teste de aleatoriedade que se baseia no gráfico da autocorrelação entretanto, ao invés de testar a aleatoriedade a cada intervalo em particular (*lag*) ele testa a aleatoriedade global baseado em um certo número de intervalos.

No campo dos testes de hipóteses, o p-valor, é a probabilidade de que uma amostra retirada de uma população sendo testada comprove a hipótese nula: a população sendo testada é uma distribuição normal. Um valor de 0,05 por exemplo, indica que há uma probabilidade de 5% de que a amostra sendo testada comprove a hipótese nula. O resultado pode ser interpretado como:

- Valor p próximo de 0 – a hipótese nula é falsa;
 - Valor p próximo de 1 - não há evidência suficiente para rejeitar a hipótese nula;
 - Normalmente considera-se um valor p de 0,05 como o patamar para avaliar a hipótese nula. Se o valor p for inferior a 0,05 rejeita-se a hipótese nula; caso contrário, não há evidências para rejeitar a hipótese nula (o que não significa automaticamente que seja verdadeira). Em situações de maior exigência é usado um valor p inferior a 0,05.
- e) histograma: é uma representação gráfica da distribuição de frequências dos dados observados, normalmente um gráfico de barras verticais. A construção de histogramas tem caráter preliminar em qualquer estudo e é um importante indicador da distribuição de dados;
- f) gráfico de probabilidades ou Q-Q: é um método gráfico para o diagnóstico de diferenças entre a distribuição de probabilidade de uma população e uma distribuição usada para a comparação. Uma forma básica de gráfico surge quando a distribuição para a comparação é uma distribuição teórica. Um exemplo do tipo de di-

ferências que podem se comprovar é a não-normalidade da distribuição de uma variável em uma população. Se a distribuição da variável é a mesma que a distribuição de comparação obter-se-á, aproximadamente, uma linha reta.

Quantis são pontos obtidos a intervalos regulares a partir da função de distribuição acumulada. Uma vez dividido os dados ordenados em q sub-conjuntos de mesmo tamanho; os quantis são os dados que fazem fronteira entre os subconjuntos subsequentes. O k -ésimo q -quantil é o valor x que tal que $P(X \leq x) \geq p$ e $P(X \geq x) \geq 1 - p$, $p = k/q$.

A exibição das informações foi dividida em dois conjuntos. O primeiro exhibe a série, o correlograma, o correlograma parcial e a densidade espectral além do gráfico de p-valores caso a série seja gaussiana. Para tal, em R, foi criado o Código 7.1:

```
eda.ts = function (x, LAG = 0) {    # Recebe a série como parâmetro

op = par(no.readonly = TRUE); par(mar=c(0,0,0,0), oma=c(1,4,2,1))

p.min = .05      # Critério de aceitação para gaussiana

k = 15; p = rep(NA, k)

for (i in 1:k) { p[i] = Box.test(x, i, type = "Ljung-Box")$p.value # Teste Ljung-Box
}

# Se o maior valor do teste de Ljung-Box for superior a 0.05 é gaussiana

if( max(p)>p.min ) { par(mfrow=c(5,1))

} else { par(mfrow=c(4,1))

}

if(!is.ts(x)) x = ts(x)    # Se não do tipo série temporal converte

plot(x, axes=FALSE)    # Desenha a série

abline (h = mean(x))
```

```

axis(2); axis(3); box() # Coloca os eixos à esquerda e superior

if (LAG == 0) { acf(x, axes=FALSE)      # Desenha o correlograma
} else { acf(x, axes=FALSE, lag.max=LAG)
}

axis(2, las=2); box(); mtext("ACF", side=2, line=2.5)

if (LAG == 0) { pacf(x, axes=FALSE)    # Desenha o correlograma parcial
} else { pacf(x, axes=FALSE, lag.max=LAG)
}

axis(2, las=2); box(); mtext("PACF", side=2, line=2.5)

xx=spectrum(ts.hw.residuals, col=par("fg"), log="dB",main="", axes=FALSE )

axis(2, las=2); mtext("Spectrum", side=2, line=2.5); box()

abline(v=xx$freq[length(xx$freq)]/1:10, lty=3)

# Desenha os p-valores caso seja gaussiana

if( max(p)>p.min ) {

    main =plot(p, type="h", ylim=c(0,1),lwd=3, main="", axes=F)

    axis(2, las=2); box(); mtext("Ljung-Box p-value", side=2, line=2.5)

    abline(h=c(0,.05),lty=3)

}

par(op)

}

```

Código 7.1 - Criação da função `eda.ts()` para apresentação do correlograma, do correlograma parcial e da densidade espectral além do gráfico de p-valores caso a série seja gaussiana

O segundo conjunto de gráficos exibe o histograma e o gráfico Q-Q. Como tem parâme-

tros que são variáveis a cada exibição não foi criada nenhuma função. São emitidos os comandos do Código 7.2.

```
eha.ts = function (x) {  
  op=par(mfrow=c(1,2),mar=c(4,4,1,0)+.1,oma=c(0,0,0,1))  
  hist(x,col="lightgray",xlab="Log(#fluxos)",main="",ylab="Frequência",xaxt="n",  
  cex.axis=.9, cex.lab=1.1)  
  M= round(max(x),1); m=round(min(x),1)  
  axis(1,c(seq(m,M,0.5)),labels=F)  
  mtext(c(seq(m,M,0.5)),side=1,las=1,at=c(seq(m,M,0.5)),cex=.6,line=0)  
  qqnorm(x,main="",xlab="Quantidades Teóricas", ylab="Quantidades  
  Amostradas", xaxt="n", cex.axis=.9, cex.lab=1.1,bty="l")  
  axis(1,c(seq(-3,3,1)),labels=F)  
  mtext(c(seq(-3,3,1)),side=1,las=1,at=c(seq(-3,3,1)),cex=.6,line=0)  
  qqline(x)  
  par(op)  
}
```

Código 7.2 - Criação da função eha.ts() para apresentação do histograma e do gráfico Q-Q

A aplicação da função eda.ts() aos dados do fluxo TCP, janela de amostragem dia, já aplicada a função de transformação logarítmica está apresentada na Figura 7.6.

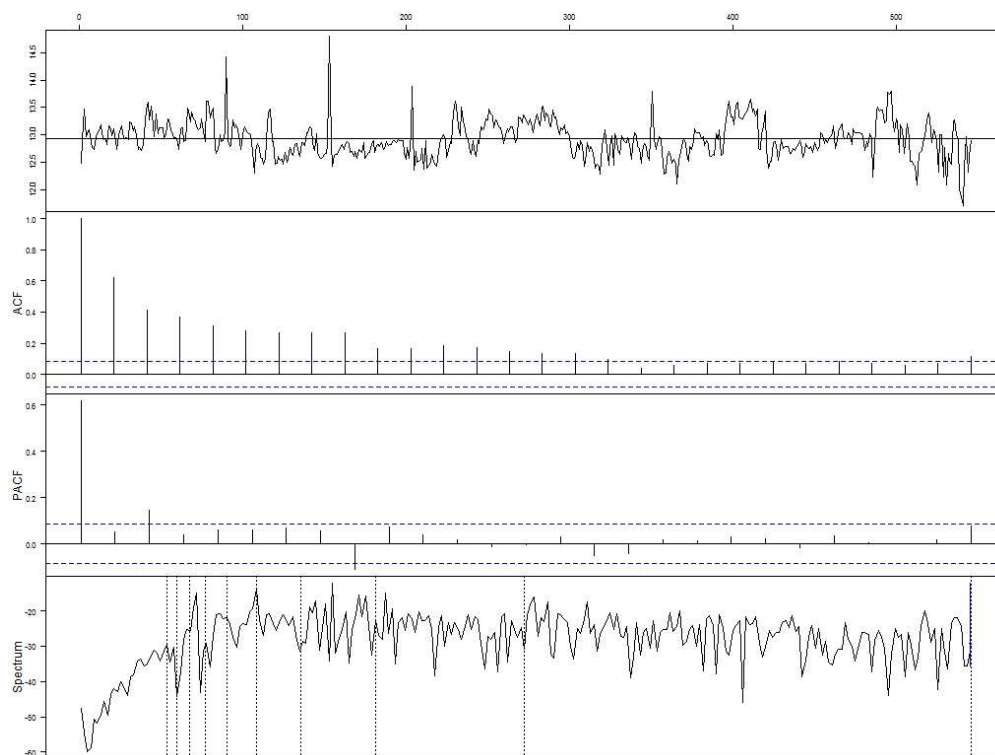


Figura 7.6 - Análise da série temporal do fluxo do protocolo TCP já aplicada a função de transformação logarítmica, janela de amostragem dia: série, correlograma, correlograma parcial e densidade espectral

A Figura 7.6, e todas as demais que serão apresentadas com este formato daqui para frente, apresenta, de cima para baixo, os seguintes gráficos:

- a) representação da série temporal sendo analisada num gráfico do tipo observação versus tempo. Na horizontal encontram-se os tempos. Na Figura 7.6 a janela de é dia, logo, tem-se o intervalo de tempo 1=01/01/2005 à 546=30/06/2006. Na vertical representa-se a quantidade de fluxos para cada tempo;
- b) correlograma. Na horizontal representam-se intervalos. A quantidade total de intervalos (*lags*) sendo representados é dado por $10 \cdot \log_{10} N$, onde N é o número de observações. Portanto, para a janela de amostragem dia são exibidos 28 intervalos. Na vertical encontra-se os valores da autocorrelação para cada um dos intervalos. As duas linhas horizontais correspondem ao intervalo de confiança de 95% da distribuição normal ($\pm 0,16$);

- c) correlograma parcial. Os eixos possuem os mesmos valores e definições do gráfico de autocorrelação;
- d) gráfico de densidade espectral. No eixo horizontal estão representadas as frequências. Considera-se somente metade do número de observações. No eixo vertical estão representadas os logaritmos na base 10 (decibel) das densidades espectrais para cada frequência. As linhas verticais são ilustrativas para representar, da esquerda para a direita, a posição de frequências $1/10$, $1/9$ até $1/2$ na lateral direita.

A análise estatística da Figura 7.6 ainda não foi esclarecedor. Aparentemente há componentes de tendência e de sazonalidade que devem ser explicitados. Porém, do gráfico de autocorrelação parece haver uma dependência linear com intervalo 1. A PACF será analisada mais tarde quando se tratar de séries auto-regressivas. O gráfico de densidade espectral ainda está com muito ruído embora pareça ter alguns picos bem destacados. Como o gráfico de p-valores não está desenhado isto significa que o componente aleatório da série não é um ruído branco.

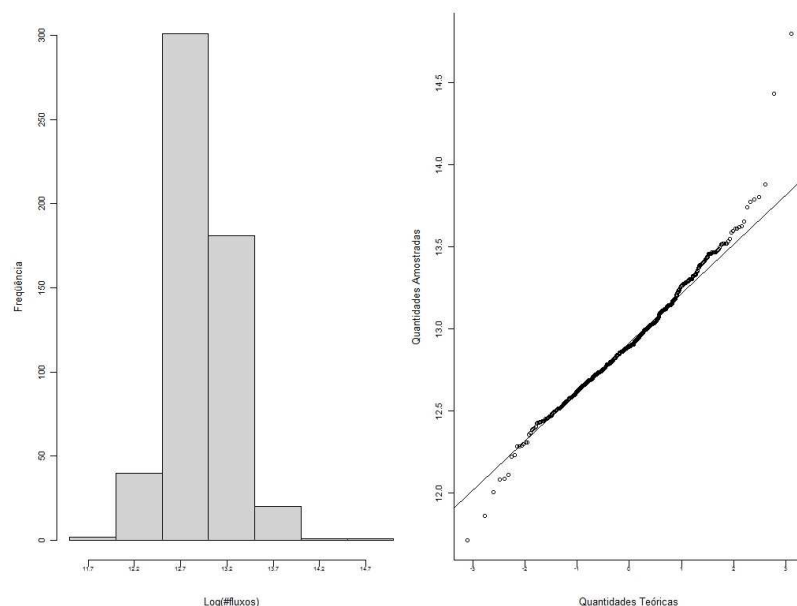


Figura 7.7 - Análise da série temporal do protocolo TCP com janela de amostragem dia: histograma e gráfico Q-Q

As demais representações gráficas da análise estatística encontram-se, para a janela de amostragem dia, representadas na Figura 7.7 nos passa uma informação muito interessante: os fluxos TCP seguem uma distribuição normal. Esta assertiva é corroborada pelo ajuste dos pontos do gráfico Q-Q gerando uma linha reta e pela forma do histograma que lembra uma curva gaussiana.

Para a janela de amostragem hora é feita a mesma análise anterior que encontra-se representada nas Figuras 7.8 e 7.9. Para esta janela de amostragem o intervalo de tempo varia de 1=01/01/2005 00:00 à 13104=30/06/2006 23:00.

A análise da Figura 7.8, particularmente do gráfico da autocorrelação, mostra que há uma repetição de picos a cada 24 intervalos. Esta repetição confirma a suspeita levantada quando analisada a série com janela de amostragem dia. Nesta análise verificou-se repetição com intervalo 1 o que corresponde a 24 intervalos nessa análise. A densidade espectral mostra alguns picos que serão analisados com mais detalhes adiante.

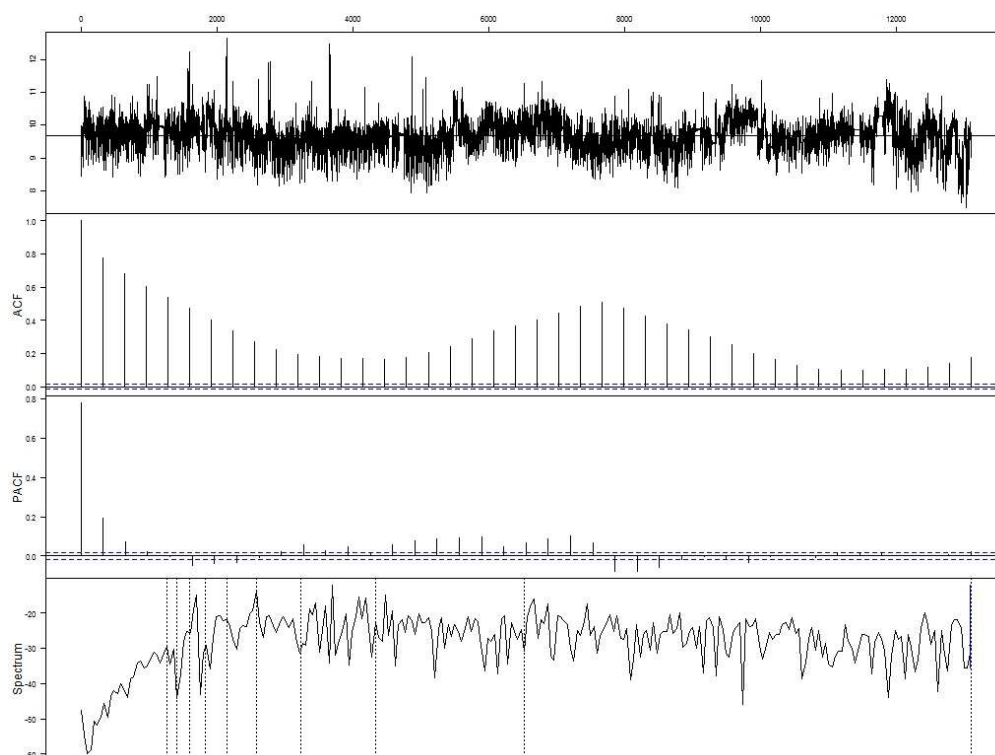


Figura 7.8 - Análise da série temporal do fluxo do protocolo TCP já aplicada a função de transformação logarítmica, janela de amostragem hora: série, correlograma, correlograma parcial e densidade espectral

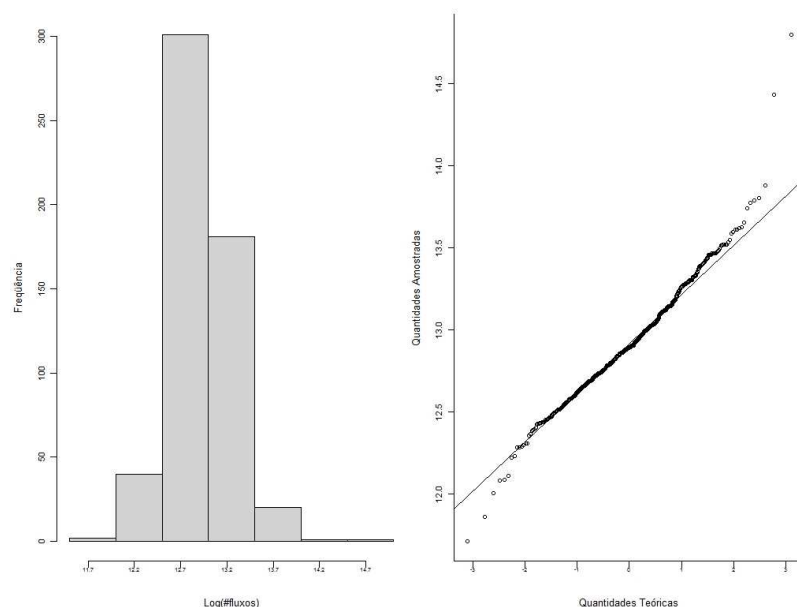


Figura 7.9 - Análise da série temporal do protocolo TCP com janela de amostragem hora: histograma e gráfico Q-Q.

A Figura 7.9 mostra que o fluxo TCP, com janela de amostragem hora, segue, como já esperado uma distribuição gaussiana.

A análise das séries temporais, com janelas de amostragem dia e hora, através do correlograma parece indicar a existência de um relacionamento linear entre os dados com intervalo de 1 (24 horas). Porém, não é possível, com estas informações, verificar a existência de um relacionamento não linear.

Para verificar a existência de dependências não lineares traçam-se gráficos de dispersão que confrontam valores de um determinado instante de tempo (t) com valores de outro instante ($t - h$); h varia de 1 até n , onde n pode ser expresso em dias ou horas dependendo da janela de amostragem usada. Num gráfico de dispersão exibem-se os valores atuais no eixo vertical e os valores deslocados de h unidades de tempo na horizontal.

As Figuras 7.10 e 7.11 mostram que a concentração de valores ao longo da reta $y = x$, isto é, onde há menor dispersão em relação a reta, para intervalo 1, com janela de amostragem dia, e 24, para janela de amostragem hora, que a relação existente entre os dados é linear.

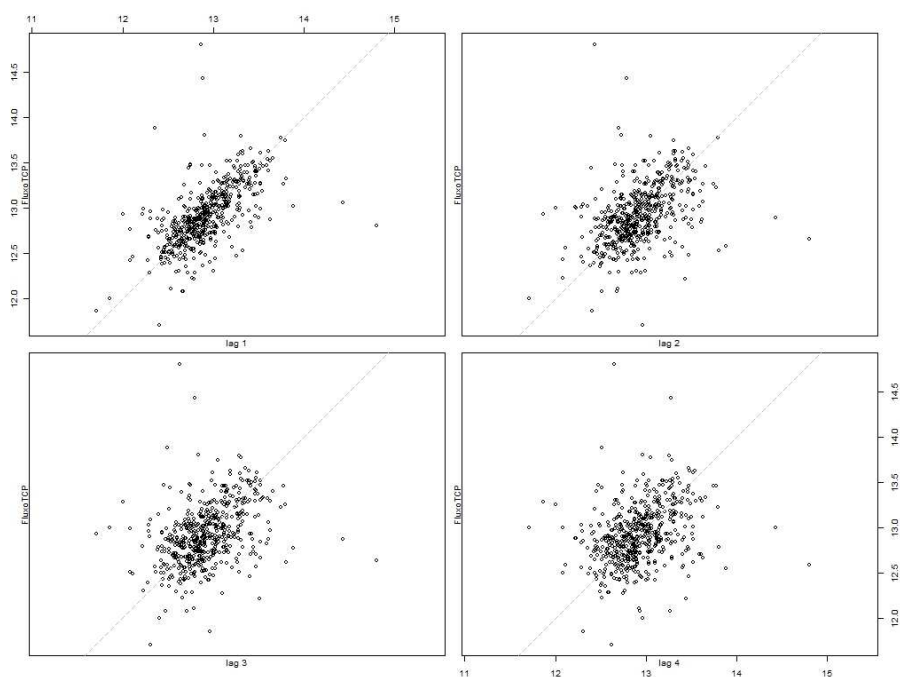


Figura 7.10 - Gráfico de dispersão do fluxo TCP, com janela de amostragem dia, para intervalos de 1 a 4 dias

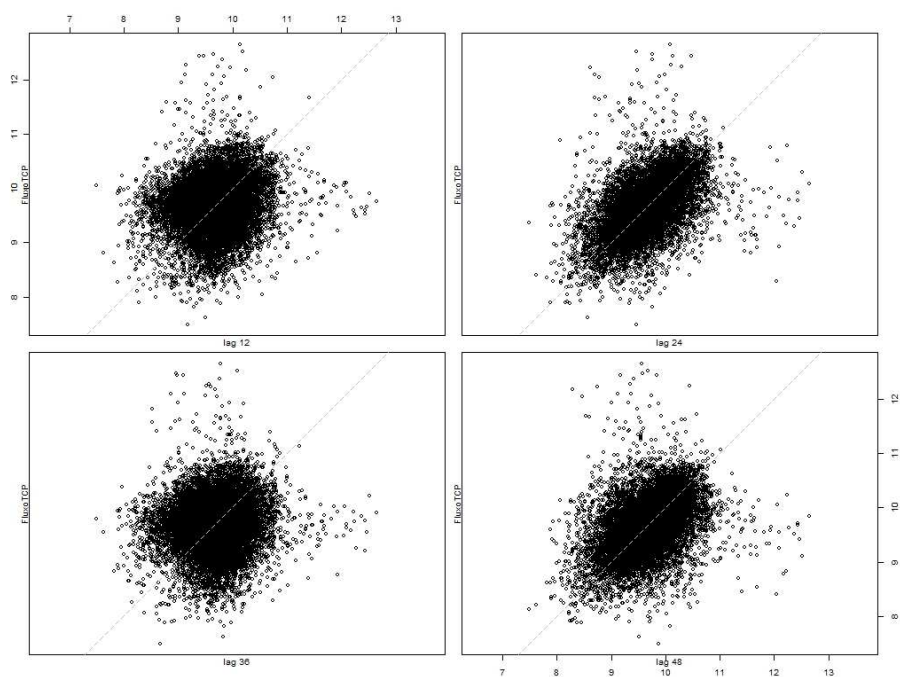


Figura 7.11 - Gráfico de dispersão do fluxo TCP, com janela de amostragem hora, para intervalos de 12, 24, 36 e 48 horas

7.3 Remoção do Componente de Tendência

Várias técnicas podem ser usadas para realizar a análise da existência de um componente de tendência e sua remoção. Neste trabalho serão apresentadas as técnicas de diferenciação e de suavização.

7.3.1 Diferenciação

Para remoção da tendência deve-se construir sucessivamente as séries das primeiras diferenças, BX_t segundas diferenças, B^2X_t , ... , k-ésimas diferenças, B^kX_t , até que a série obtida não revele tendência. A série resultante é analisada na Figura 7.12.

Observa-se na Figura 7.12 que a série resultante, isto é, o ruído resultante da supressão do componente de tendência da série original, pela aplicação da primeira diferença não é gaussiana uma vez que o gráfico dos p-valores não foi exibido.

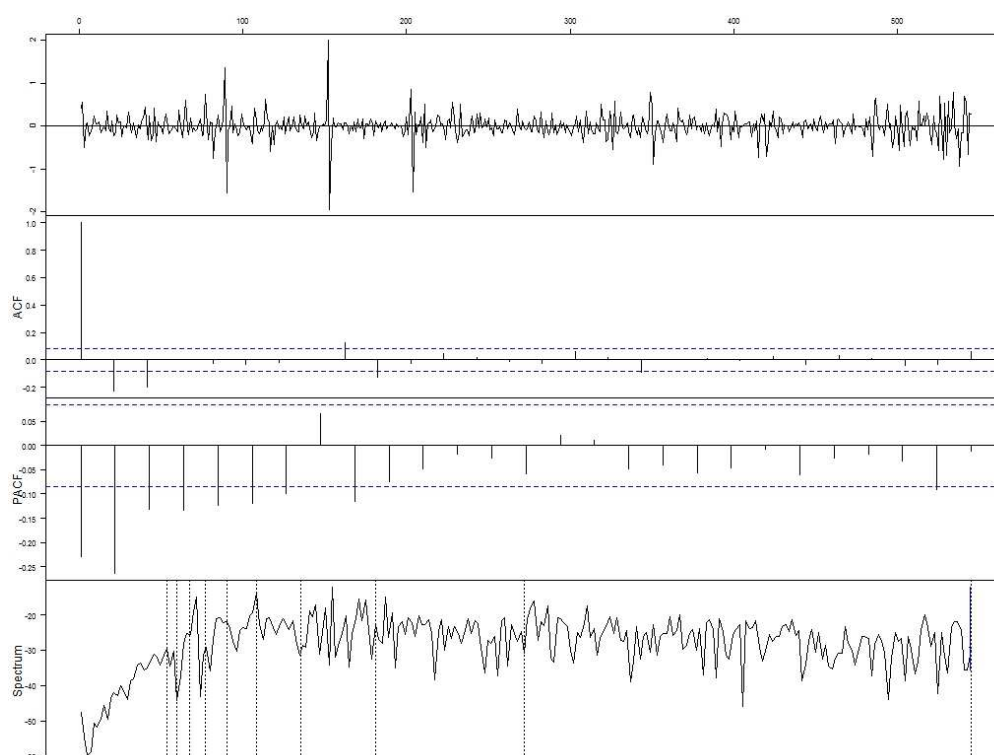


Figura 7.12 - Análise da série temporal resultante da aplicação da primeira diferença com intervalo 1, janela de amostragem dia: série, correlograma, correlograma parcial e densidade espectral

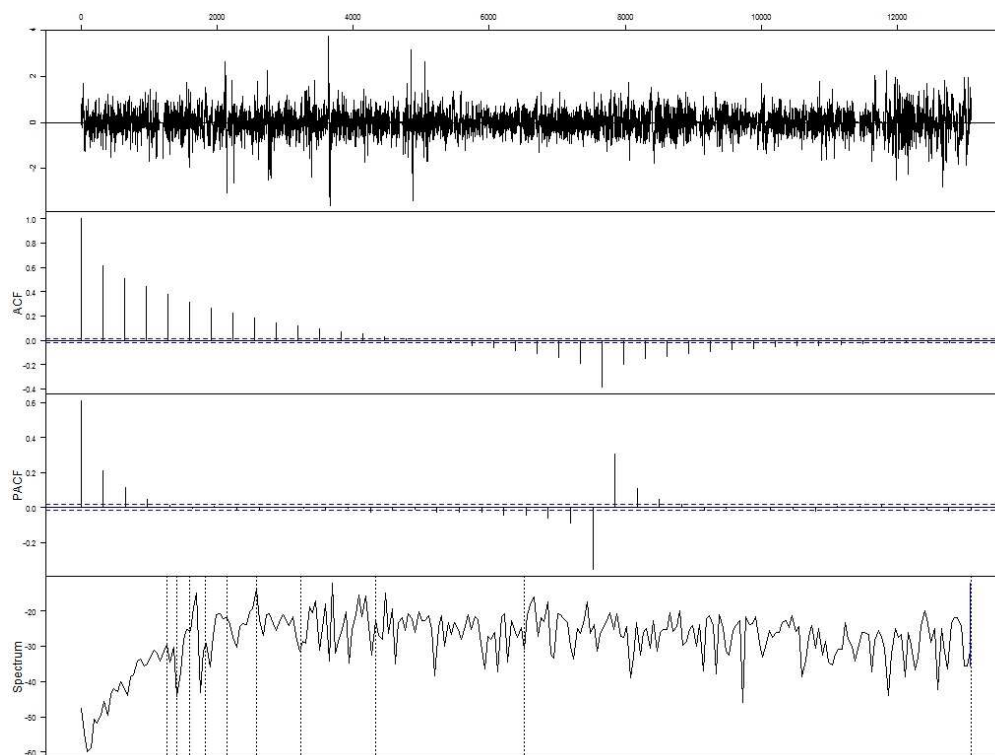


Figura 7.13 - Análise da série temporal resultante da aplicação da primeira diferença com intervalo 24, janela de amostragem hora: série, correlograma, correlograma parcial e densidade espectral

A série oriunda da aplicação da diferenciação na série com janela de amostragem hora também não é gaussiana como se pode ver na Figura 7.13.

A aplicação da primeira diferença pode ser interpretada como a aplicação da 1ª derivada. Isto implica em dizer que o resultado da aplicação da primeira diferença pode ser vista como a velocidade de fluxos por janela de amostragem. Para as duas janelas consideradas a média das velocidades é praticamente zero.

Estar próximo de zero significa que as velocidades variam sobre uma distribuição contínua no tempo alternando velocidades positivas e negativas que no final geram uma curva com média nula (desvios de 0.28 e 0.48 para dia e hora respectivamente). Isto quer dizer que, ao longo do tempo não há uma tendência nítida de crescimento ou de decréscimo dos fluxos constituintes do ruído de fundo.

O uso deste modelo para predição de eventos futuros não possui modelagem adequada e, portanto, não será mais usado deste ponto em diante.

7.3.2 Suavização/Filtragem

Neste trabalho usa-se somente filtragem uma vez que o objetivo final é a predição de eventos. Para aplicá-la é preciso estimar o tamanho da janela onde a média é calculada e os pesos a serem aplicados. Num critério de tentativa-e-erro comparou-se os erros médios quadráticos dos valores obtidos a partir de janelas com tamanho variando de 3 até 6 e com pesos variando de 0,6, para o elemento mais próximo, até 0,05, o mais afastado.

Para janelas de amostragem dia o uso de janelas variando de 3 a 6 significa que se acredita que dados de até 6 dias atrás influem no comportamento do dado atual. Como o menor valor usado é de 3 dias assumiu-se que o comportamento atual é dependente de, pelo menos, os três dias anteriores. Análise semelhante é válida quando usando-se janela de amostragem hora sendo que, neste caso, teriam-se variações entre 3 e 6 horas.

Os pesos representam o quanto cada dado vai influenciar o comportamento do dado atual, isto é, a aplicação de um peso 0,6 e outro 0,4 indica que o primeiro tem uma influência de 60% e o segundo uma influência de 40%.

O erro médio quadrático apresentou valor mínimo para uma janela de tamanho 3 com pesos 0,85, 0,10 e 0,05, tanto para janela de amostragem dia e hora. Aplica-se a filtragem com a janela e os pesos obtendo-se uma nova série. Diminui-se, então, esta nova série da série original obtendo uma série sem o componente de tendência. A representação da análise desta nova série é vista na Figura 7.14.

A Figura 7.14 apresenta, pela primeira vez, o gráfico dos p-valores. A aplicação de filtragem para remoção do componente de tendência conduziu a um ruído gaussiano. Este é um resultado importante que pode ser usado para predição uma vez que é ruído branco e há modelo matemático bem conhecido para predição de valores da série original.

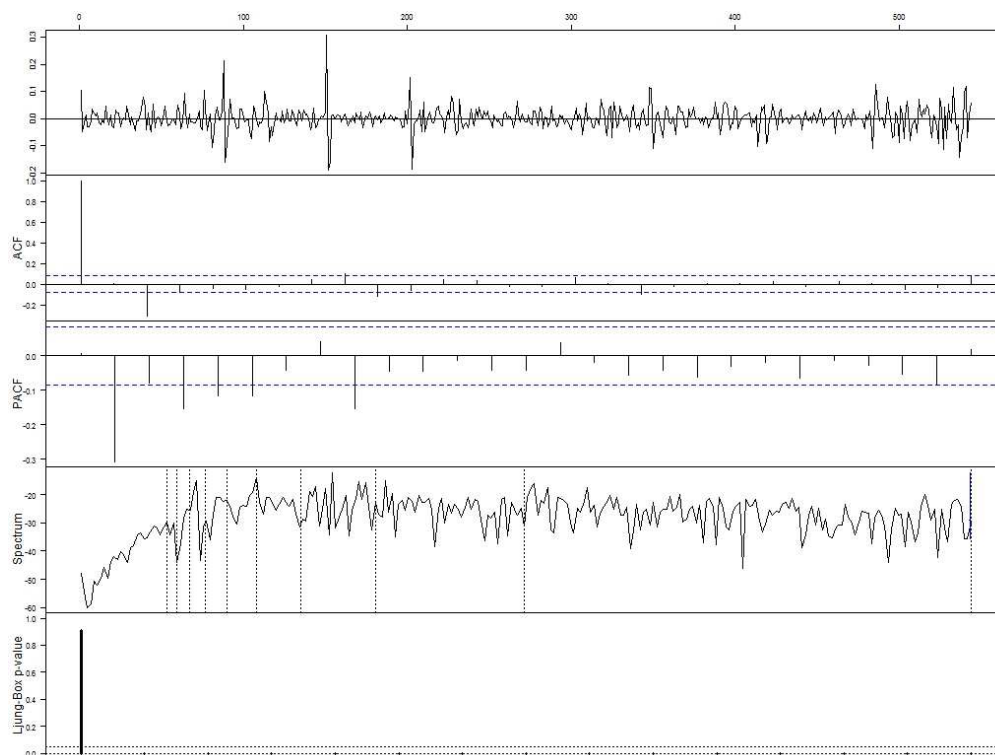


Figura 7.14 - Análise da série temporal resultante da aplicação de filtragem com janela de tamanho 3, pesos 0,85, 0,1 e 0,05, janela de amostragem dia

A mesma técnica, quando aplicada aos dados com janela de amostragem hora gera uma série cuja análise encontra-se na Figura 7.15.

Tal qual na série resultante para a janela de amostragem dia a série dos resíduos da janela de amostragem hora possui ruído gaussiano.

A presença de ruído gaussiano já permitiria o uso deste modelo para predição sem a necessidade de avaliação do componente sazonal. Entretanto, este trabalho tenta esgotar todas as possibilidades.

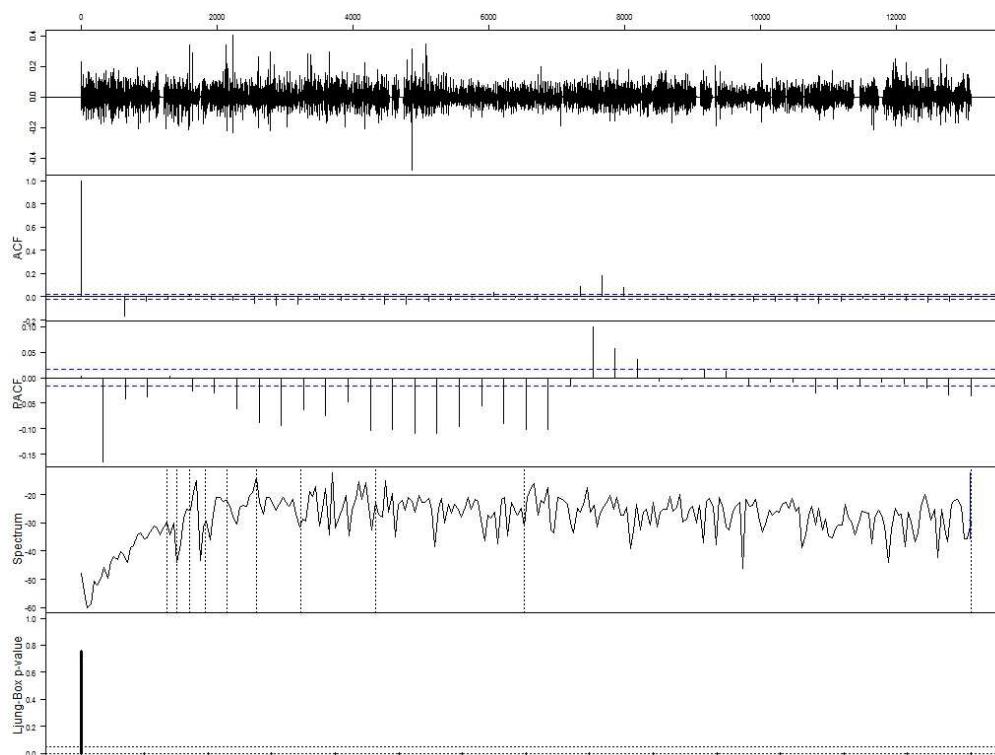


Figura 7.15 - Análise da série temporal resultante da aplicação de filtragem com janela de tamanho 3, pesos 0,85, 0,1 e 0,05, janela de amostragem hora

7.3.2.1 Suavização exponencial dupla

O modelo encontra-se explicado em 5.2.1.3 e os parâmetros α e β , já apresentados na Fórmula 5.18, tem de ser estimados. Um possível relacionamento entre estes dois parâmetros pode ser expresso um através da Fórmula 7.1.

$$\begin{aligned}\alpha &= 1 - (1 - \delta)^2 \\ \beta &= \frac{\delta^2}{1 - (1 - \delta)^2}\end{aligned}\tag{7.1}$$

Para determiná-los usou-se função *HoltWinters()* para que ela, a partir de diversos valores para os parâmetros de α e β calculasse o erro médio quadrático. O conjunto de parâmetros que conduz ao menor erro quadrático é o de interesse.

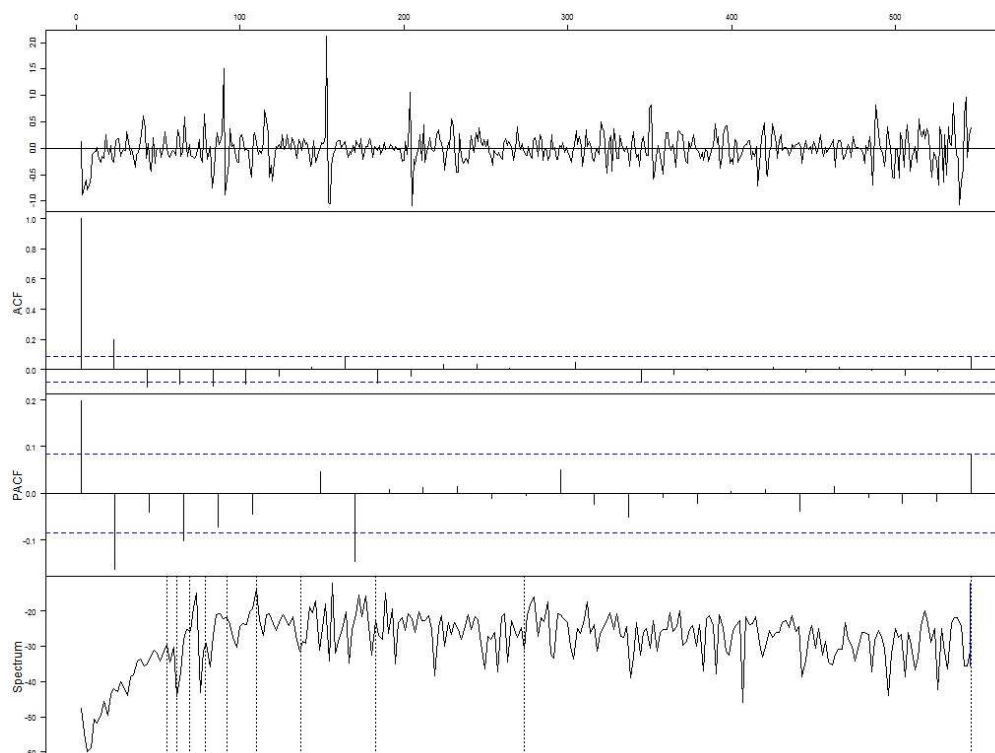


Figura 7.16 - Análise da série temporal resultante da aplicação de suavização exponencial dupla com $\alpha = 0.493056$ e $\beta = 0.1682243$, janela de amostragem dia

Os valores de δ passados variam de $[0,1, 0,99]$ intervalados de 0,001. O menor erro quadrático ocorreu para os valores $\alpha = 0.493056$ e $\beta = 0.1682243$, para janela de amostragem dia e $\alpha = 0.557775$ e $\beta = 0.2012012$, para janela de amostragem hora.

Adotado o modelo aditivo tem-se que a série correspondente ao resíduo é obtida pela subtração das série original da ajustada (a sem o componente de tendência). A análise dos resíduos são apresentados nas Figura 7.16 e 7.17.

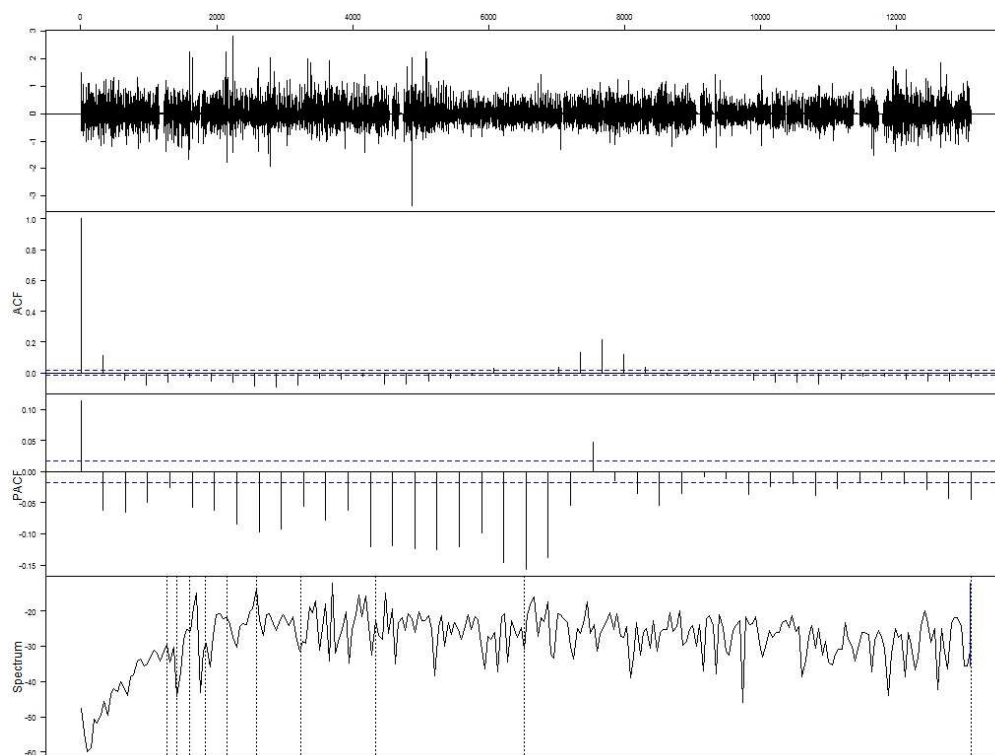


Figura 7.17 - Análise da série temporal resultante da aplicação de suavização exponencial dupla com $\alpha = 0.557775$ e $\beta = 0.2012012$, janela de amostragem hora

As duas figuras anteriores mostram que o ruído resultante não é gaussiano.

7.4 Remoção do componente de sazonalidade

Sazonalidade são flutuações que ocorrem repetidamente a intervalos regulares dentro do período de análise. Para que seja possível retirar o componente de sazonalidade da série original é necessário o conhecimento da frequência ou do período com que ocorrem as flutuações.

A princípio, nem a frequência nem o período são conhecidos. Para determiná-los pode-se usar, por exemplo, os métodos do periodograma escalado e/ou a densidade espectral. Ambos os métodos são representações do sinal no domínio das frequências.

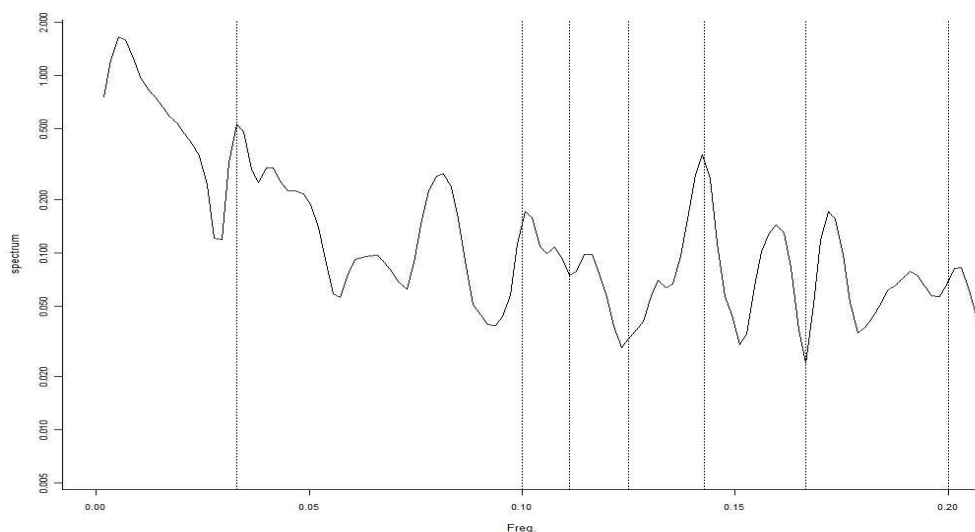


Figura 7.18 - Gráfico da densidade espectral, suavizado e limitado ao intervalo [0, 0.2] das frequências, para fluxo TCP com janela de amostragem dia

A representação gráfica da análise das séries temporais apresenta o gráfico da densidade espectral o qual não foi analisado até o presente momento. Ele será o foco daqui para frente. Entretanto, todas as figuras até este ponto mostram tanta informação que é muito difícil distinguir um pico de destaque nestas condições.

Para melhorar a informação disponível no gráfico da densidade espectral pode-se introduzir um parâmetro de suavização e, desta representação, identificar um intervalo onde ocorre um pico e limitar a representação para um intervalo que contenha o pico.

A Figura 7.18 apresenta a densidade espectral, suavizada e limitada, para a série temporal, já aplicada a transformação logarítmica, dos fluxos TCP com janela de amostragem dia. Nela é possível notar a presença de um pico que se sobrepõe aos demais na posição aproximada de 0,03. Consultando os valores da densidade espectral verifica-se um máximo no ponto (0.03298611, 0.5333935).

Para a frequência de 0,03298611, são realizados aproximadamente 18 ciclos durante as 546 observações disponíveis ou seja, um ciclo a cada 30 observações (ocorrência de ciclos mensais).

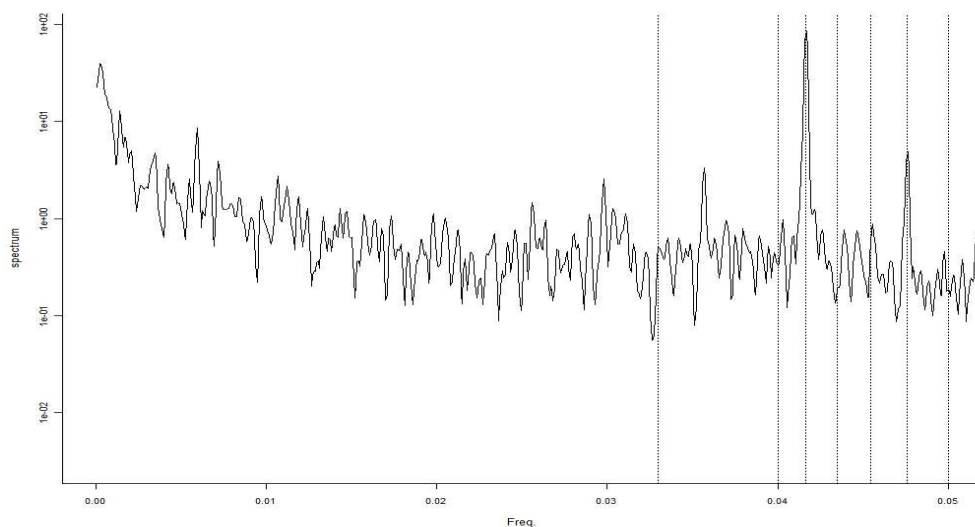


Figura 7.19 - Gráfico da densidade espectral, suavizado e limitado ao intervalo [0, 0.05] das frequências, para fluxo TCP com janela de amostragem hora.

Na Figura 7.19 verifica-se a presença de um pico que se sobrepõe aos demais na posição aproximada de 0,04. Consultando os valores da densidade espectral verifica-se um máximo no ponto (0.04168572, 85.44535). Logo, verifica-se que para a frequência de 0,04168572, são realizados aproximadamente 547 ciclos nas 13.104 observações ou, um ciclo a cada 24 observações (ciclos diários).

Determinado que há uma sazonalidade vamos apresentar a suavização exponencial tripla que é uma das possíveis formulações capazes de eliminar o componente de sazonalidade da série original resultando em um resíduo que é um ruído gaussiano.

7.4.1 Suavização exponencial tripla

Tal como em 7.3.2.1, é usada a função *HoltWinters()*, com os parâmetros já levantados. Entretanto, como se subentende a existência de um período é necessária a criação de um objeto do tipo “série temporal” (*ts*) onde a frequência é informada. Em R, embora o nome do parâmetro usado pela função de criação do objeto seja *frequency* o requerido é o período.

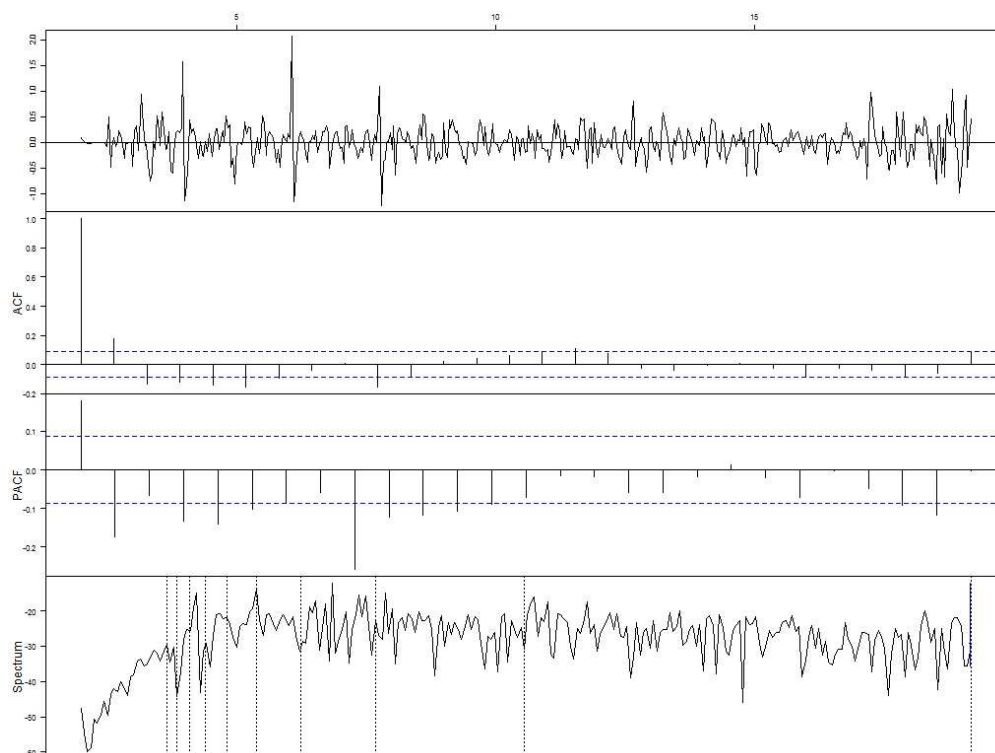


Figura 7.20 - Análise da série temporal resultante da aplicação de suavização exponencial tripla com $\alpha = 0.493056$, $\beta = 0.1682243$ e $\gamma = 0.3850921$, janela de amostragem dia

A aplicação do modelo é feita sobre série original aplicada a transformação logarítmica. A análise do resíduo resultante da aplicação deste modelo, com os parâmetros descritos, encontra-se na Figura 7.20, para a janela de amostragem dia, e na Figura 7.21 para a janela de amostragem hora.

Após a aplicação deste método obtém-se, para a janela de amostragem dia ruído não gaussianos que não podem ser usados para predição. Entretanto, para a janela de amostragem hora o ruído gerado foi, pelo teste, considerado gaussiano embora, como se pode ver pela Figura 7.21 sua condição de aceitabilidade foi a menor possível. Sendo assim, nenhum dos ruídos gerados será usado para geração de eventos futuros.

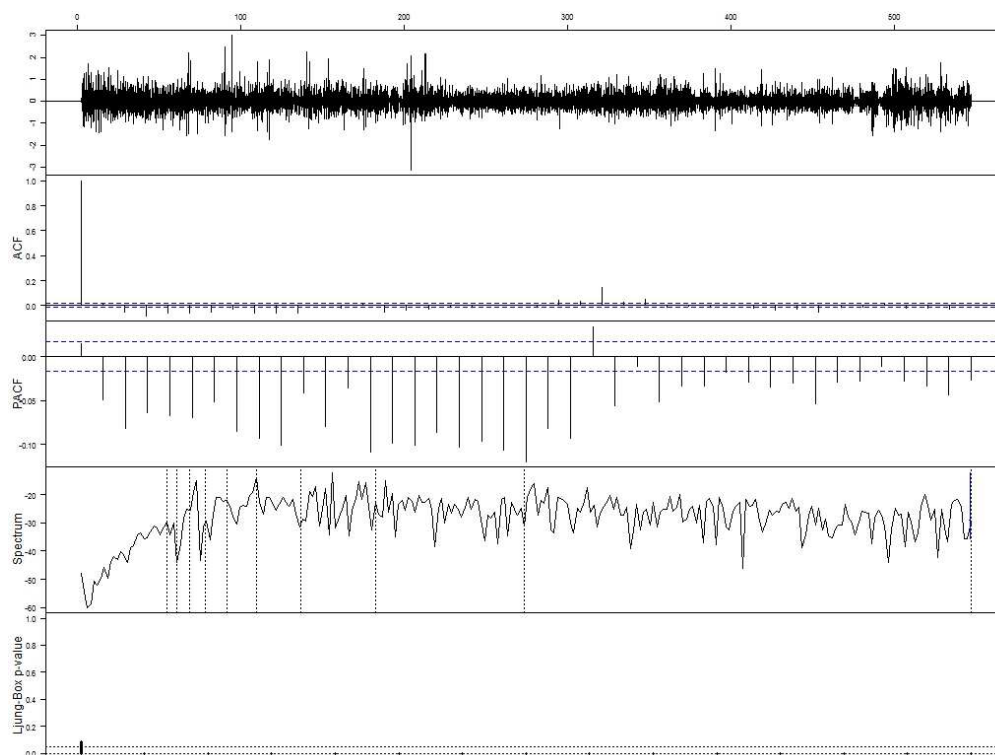


Figura 7.21 - Análise da série temporal resultante da aplicação de suavização exponencial tripla com $\alpha = 0.557775$, $\beta = 0.2012012$ e $\gamma = 0.03205071$, janela de amostragem hora

7.5 Remoção de estruturas auto-regressivas

Nesta análise não se está considerando qualquer análise de sazonalidade e dando como entrada, no modelo de remoção de estruturas auto-regressivas, a série original já aplicada a transformação logarítmica. Lembrando que o modelo ARIMA pode ser constituído de uma série de combinações entre os componentes auto-regressivo (AR), de integração (I) e de média móvel (MA). Desta forma é necessário determinar qual o conjunto de parâmetros que é melhor para as séries sendo tratadas.

Foi usado o critério Akaike para determinação do melhor conjunto de parâmetros. Passa-se série original como parâmetro de entrada para a função *arima()* e são testadas todas as combinações, entre $[0,2]$, para os parâmetros p , q e d .

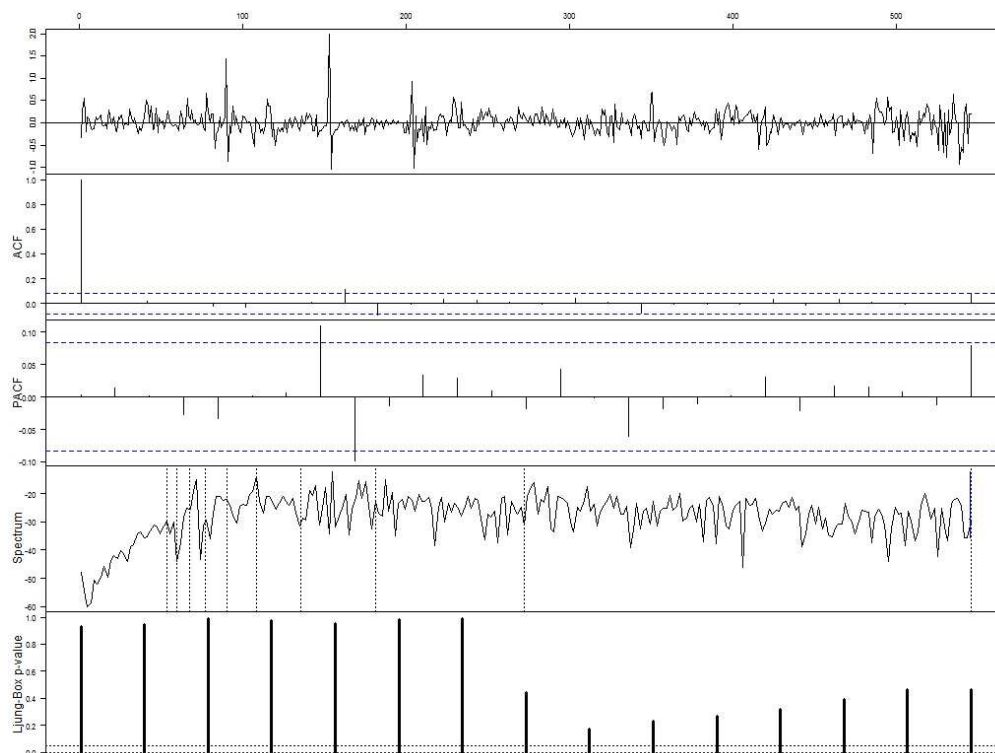


Figura 7.22 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 1$, $d = 0$ e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem dia

Para a série com janela de amostragem dia o menor valor AIC obtido foi para o conjunto de parâmetros $p = 1$, $d = 0$ e $q = 2$; para a janela de amostragem hora os parâmetros foram $p = 1$, $d = 1$ e $q = 2$.

É possível verificar que para a janela de amostragem dia a análise da série determinou que a mesma era estacionária para a janela de amostragem dia e que necessitava do cálculo da primeira diferença, para torná-la estacionária, para a janela de amostragem hora.

A análise do ruído, para janela de amostragem dia, está apresentada na Figura 7.22. Ela mostra que o ruído resultante é um gaussiano. Não há mais correlações aparentes nos ACF e PACF e o gráfico de densidade espectral não apresenta uma frequência dominante e o gráfico de p-valores foi exibido. De acordo com a análise este modelo é um candidato a ser utilizado em predição.

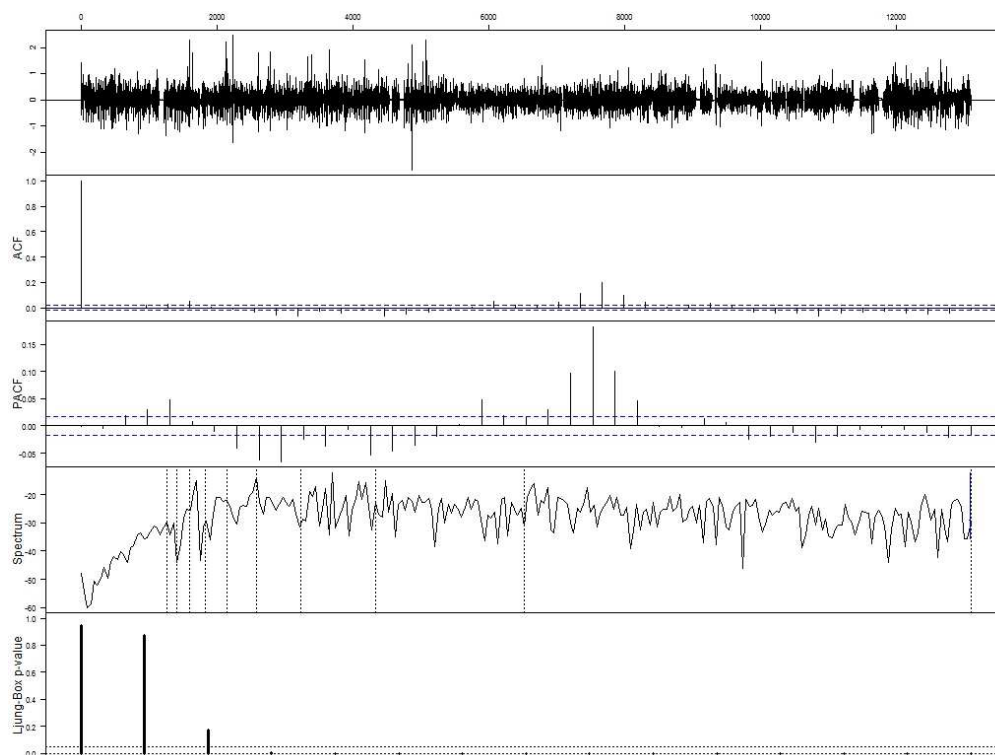


Figura 7.23 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 1$, $d = 1$ e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem hora

A Figura 7.23 mostra que o ruído resultante, para a janela de amostragem hora, é um ruído gaussiano. Entretanto, parece haver um componente sazonal ou alguma forma de correlação nos gráficos ACF e PACF.

7.5.1 SARIMA

A seleção do melhor conjunto de parâmetros, dentre todos os possíveis, é uma tarefa difícil uma vez que são 6 parâmetros cada um com 3 valores possíveis (assumindo que o conjunto possível de valores se situe entre $[0, 2]$).

Os dados de entrada são os fluxos TCP originais, já aplicada a transformação logarítmica. É necessário que se informe o período da sazonalidade de cada série e, para tal, serão usados os valores já determinados na Seção 7.4, isto é, para o fluxo TCP com janela de amostragem dia é utilizado 30 e, para janela de amostragem hora, 24.

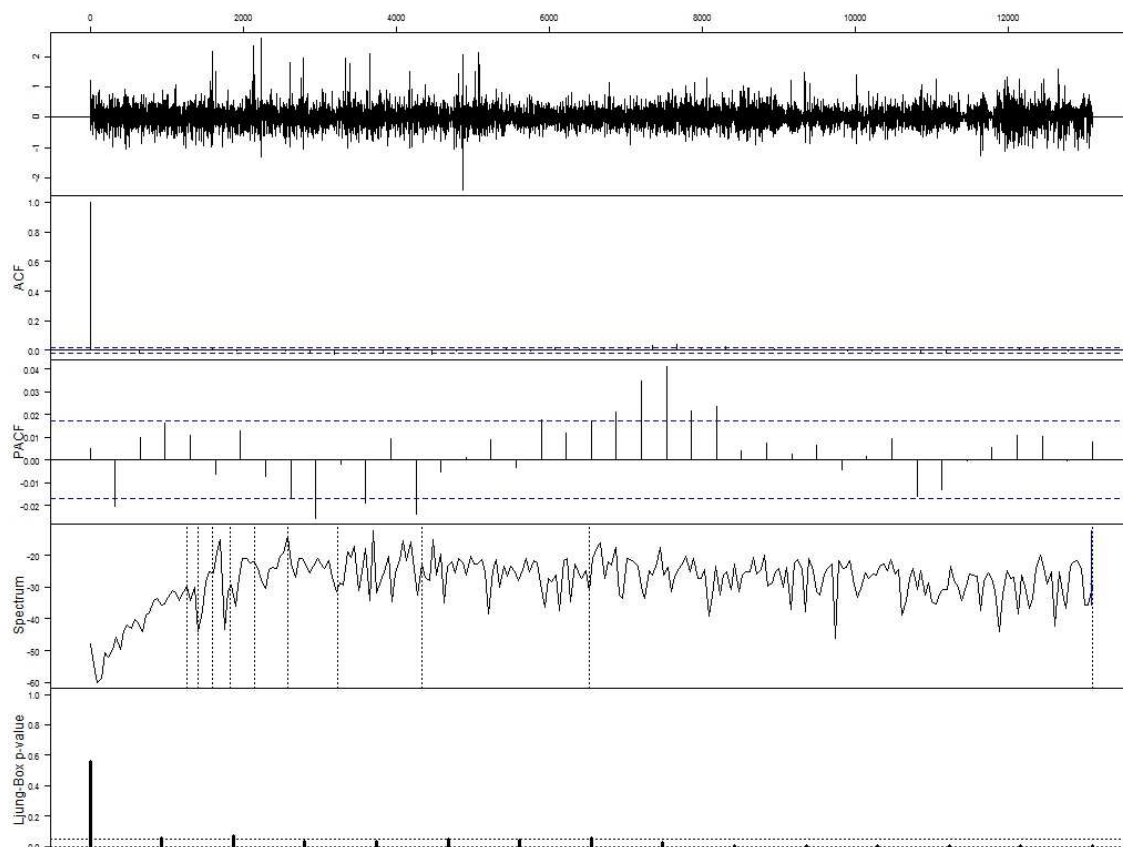


Figura 7.24 - Análise do resíduo resultante da aplicação de auto-regressão sazonal com $p = 1$, $d = 1$, $q = 2$, $P = 1$, $D = 0$ e $Q = 1$ ao fluxo TCP, já aplicada a transformação logarítmica, com janela de amostragem hora

Para determinar qual o melhor conjunto de parâmetros usou-se o critério Akaike.

Para a janela de amostragem dia verifica-se que o critério Akaike é mínimo para os parâmetros $p = 1$, $d = 0$, $q = 2$, $P = 0$, $D = 0$ e $Q = 0$. Logo, para esta janela de amostragem, a sazonalidade não é fator fundamental podendo ser desconsiderado.

O processo de auto-regressão integrada sazonal acabou por determinar o menor valor para o critério Akaike os mesmos parâmetros já determinados para o processo de auto-regressão integrada e já calculados na Seção 7.5 .

Para a janela de amostragem hora, o critério Akaike é mínimo para os parâmetros $p = 1$, $d = 1$, $q = 2$, $P = 2$, $D = 0$ e $Q = 1$. A análise dos resíduos da aplicação destes parâmetros ao fluxo TCP com janela de amostragem hora encontra-se na Figura 7.24.

Da análise do resíduo verifica-se que a série original dos fluxos TCP com janela de amostragem hora gera um resíduo gaussiano quando aplicado modelo de auto-regressão sazonal – SARIMA.

8 PREDIÇÃO DE EVENTOS FUTUROS

O objetivo deste capítulo não é o de desenvolver uma teoria estatística para prever eventos futuros mas sim aplicar conhecimentos já consagrados desta área e, a partir dos resultados, verificar se é possível antecipar a geração de um alerta no caso de um ataque maciço à parcela brasileira da Internet.

Várias combinações de modelos matemáticos foram testados. Adotou-se, porém, para este trabalho, a apresentação dos modelos que, da análise realizada no Capítulo 7, verificou-se serem os com melhores resultados estatísticos: filtragem e auto-regressivo integrado. Estes modelos apresentaram resultados satisfatórios tanto para as janelas de amostragem dia como para hora.

A previsão não é, por si só, um evento com certeza absoluta uma vez que o processo de modelagem adotado é estocástico. Um intervalo de confiança, pela própria teoria da previsão, é o intervalo dentro do qual as observações reais podem ocorrer mantendo a previsão correta. Entretanto, o que se deseja é verificar, nos instantes seguintes aos das observações já levantadas, como será o comportamento do ruído de fundo para permitir a geração de alertas precoces.

Assim, neste trabalho, embora possível, não será tratado o intervalo de confiança. O valor predito será assumido como provável para o comportamento da série e analisado em relação à geração de alertas.

Como um possível forma de validação e/ou comparação entre os modelos de previsão será calculado o erro médio quadrático entre os valores previstos e os futuramente observados.

Parece intuitivo que quanto maior o número de observações disponíveis melhor será a modelagem e, conseqüentemente, a previsão realizada. Entretanto, este número deve ser tal que não comprometa o tempo computacional que inviabilize a obtenção de resultados dentro de um tempo útil de utilização do mesmo.

Para a realização da predição de eventos futuros, com janela de amostragem dia, serão modelados os dados dos primeiros 534 dias (de 01/01/2005 até 18/06/2006) e previstos os próximos 12. Os valores preditos serão comparados com os dados observados entre 19/06/2006 e 30/06/2006. Em termos práticos a janela de 1 dia não é útil para a geração de alertas precoces! Porém, permite a validação do modelo matemático com um menor esforço computacional. A previsão poderia ser de qualquer número de dias. O valor 12 está sendo usado para manter compatibilidade com a previsão que será feita com a janela de amostragem hora.

Para predição de eventos futuros, com janela de amostragem hora, serão modelados os dados das primeiras 13.092 horas (de 01/01/2005 às 00:00 até 30/06/2006 às 11:00) e preditas as observações das próximas 12 horas. Os valores preditos serão comparados com os dados observados em 30/06/2006, das 12:00 às 23:00.

8.1 Filtragem

A predição de eventos futuros para a série modelada com o modelo matemático da filtragem é obtida através da formulação da média móvel ponderada já ilustrada na equação 6.7. Os pesos a serem aplicados são os mesmo usados na Seção 7.3.2 .

Desta forma, os eventos futuros a serem determinados serão obtidos a partir da Equação 8.1. Lembrando que os coeficientes já foram anteriormente explicados e representam a equação com menor erro quadrático em relação às observações.

$$\hat{b}_T = 0,85 * x_{T-1} + 0,10 * x_{T-2} + 0,05 * x_{T-3} \quad (8.1)$$

A comparação entre os valores preditos e os observados, para a janela de amostragem dia, encontra-se representada na Figura 8.1. Ela apresenta dados observados, no intervalo de tempo 500 a 534. A previsão realizada entre os dias 535 e 546 é representada como uma linha contínua. Os valores observados no período predito estão representados como pontos com círculos. O erro entre o valor predito e o observado por uma linha pontilhada na cor cinza.

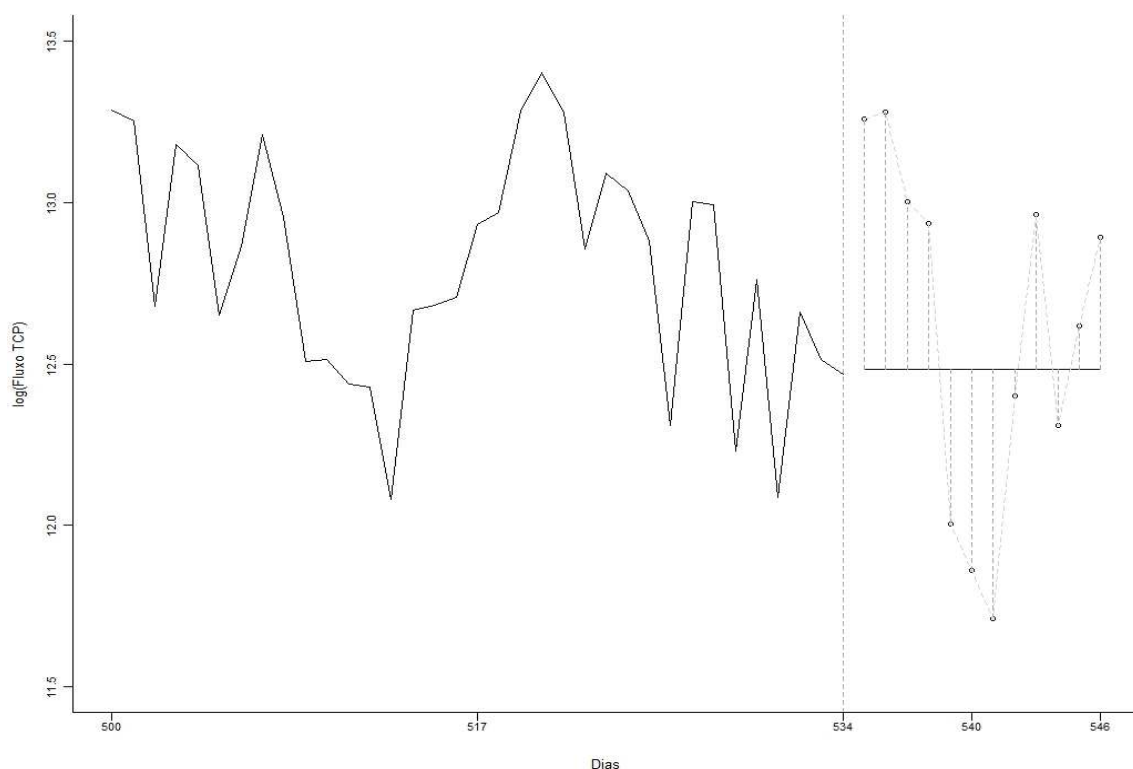


Figura 8.1 - Previsão do comportamento do ruído de fundo, do fluxo TCP, da parcela brasileira da Internet, já aplicada a transformação logarítmica, com janela de amostragem dia

O erro médio quadrático para esta predição é de 3.378496.

A comparação entre os valores preditos e os observados, para a janela de amostragem dia, encontra-se representada na Figura 8.1. Nela verifica-se que o comportamento linear de predição para a janela de amostragem dia não se mostrou adequada.

Entretanto, para a janela de amostragem hora, que possui mais observações, a análise da Figura 8.2 permite inferir que os primeiros dados preditos são muito bons quando comparados com os valores observados (os erros quadráticos são 0.0002062256 e 0.0006632945 para os primeiro e segundo pontos, respectivamente).

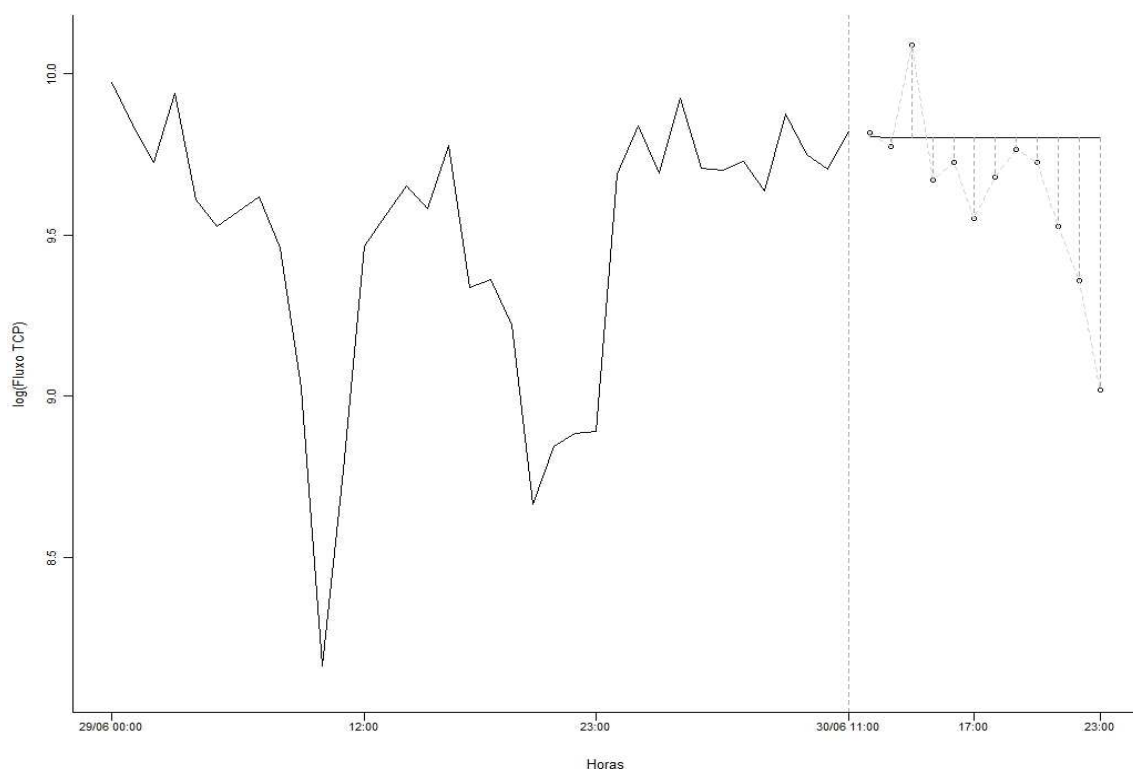


Figura 8.2 - Previsão do comportamento do ruído de fundo, do fluxo TCP, da parcela brasileira da Internet, já aplicada a transformação logarítmica, com janela de amostragem hora

O erro médio quadrático para esta predição é de 1.071593.

E o que esta predição está querendo mostrar? Que na próxima hora há tendência de se continuar o movimento ascendente de observações mas que, a partir daí, haverá uma diminuição.

Na prática, se estivermos olhando o comportamento de um agente malicioso, como um worm, por exemplo, o que deveria estar observando é um aumento exponencial no tráfego do ruído de fundo.

O comportamento esperado de um worm é uma fase inicial de descoberta de vulnerabilidades aonde ele, geralmente, não consegue ser detectado. A partir daí entra numa fase explosiva onde, se traçado graficamente, dará uma representação exponencial uma vez que sua propagação é maciça e muito rápida.

O que se verificou com a predição foi uma tendência de manutenção do status atual ou mesmo uma elevação com consequente diminuição o que foge à característica do comportamento de ataques maciços não sendo necessária a geração de alertas.

As observações dos próximos eventos comprovam a conclusão observada.

8.2 Processo auto-regressivo integrado

O processo matemático para determinação de eventos futuros é inerente ao processo auto-regressivo integrado e não será explorado neste trabalho.

A comparação entre os valores preditos e os observados, para a janela de amostragem dia, encontra-se representada na Figura 8.3.

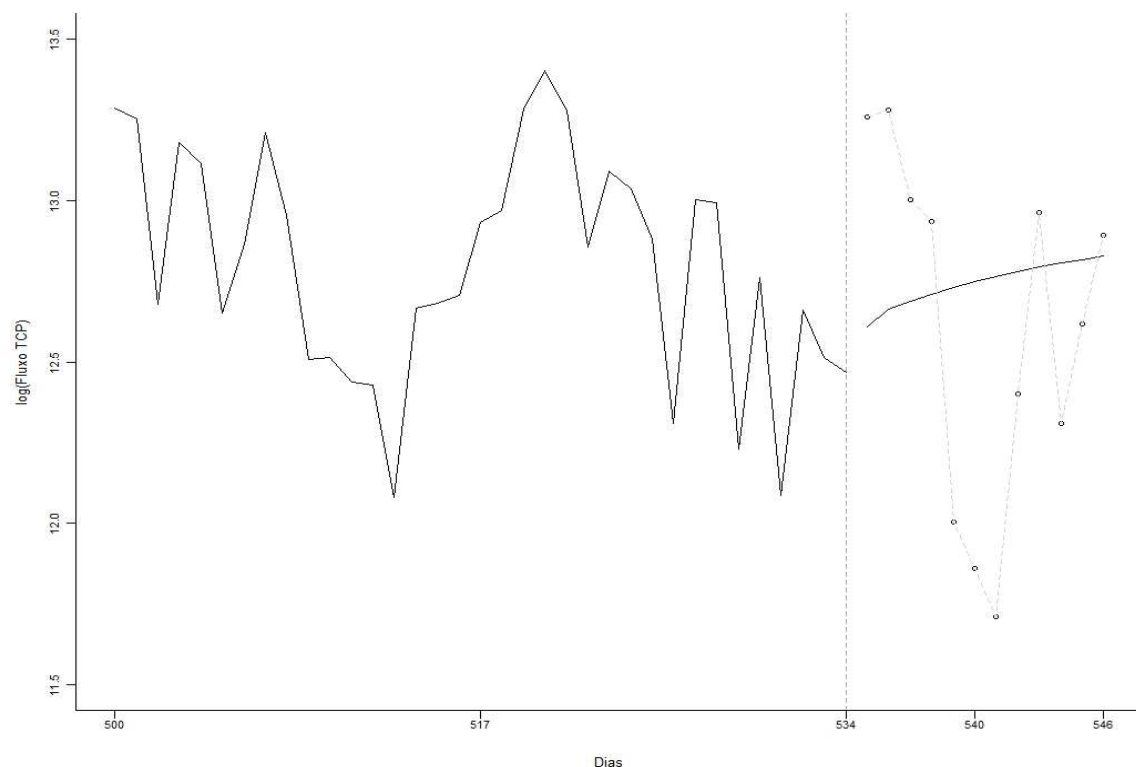


Figura 8.3 - Previsão do comportamento do ruído de fundo da parcela brasileira da Internet, com $p = 1$, $d = 0$ e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem dia

A Figura 8.3 apresenta parte dos dados observados, no intervalo de tempo 500 a 534. A

previsão realizada entre os dias 535 e 546 está representada por uma linha contínua. Os valores observados no período predito estão representados como pontos com círculos. O erro entre o valor predito e o observado por uma linha pontilhada na cor cinza.

O erro médio quadrático para esta predição é de 3.84345.

Embora a janela de observação dia não seja adequada ao tipo de processamento desejado é interessante notar a tendência de crescimento do fluxo TCP no ruído de fundo com o passar dos dias. Entretanto, este crescimento não é explosivo o que não configura um ataque na parcela brasileira da Internet.

A comparação entre os valores preditos e os observados, para a janela de amostragem hora, encontra-se representada na Figura 8.4.

O erro médio quadrático para esta predição é de 0.8516938.

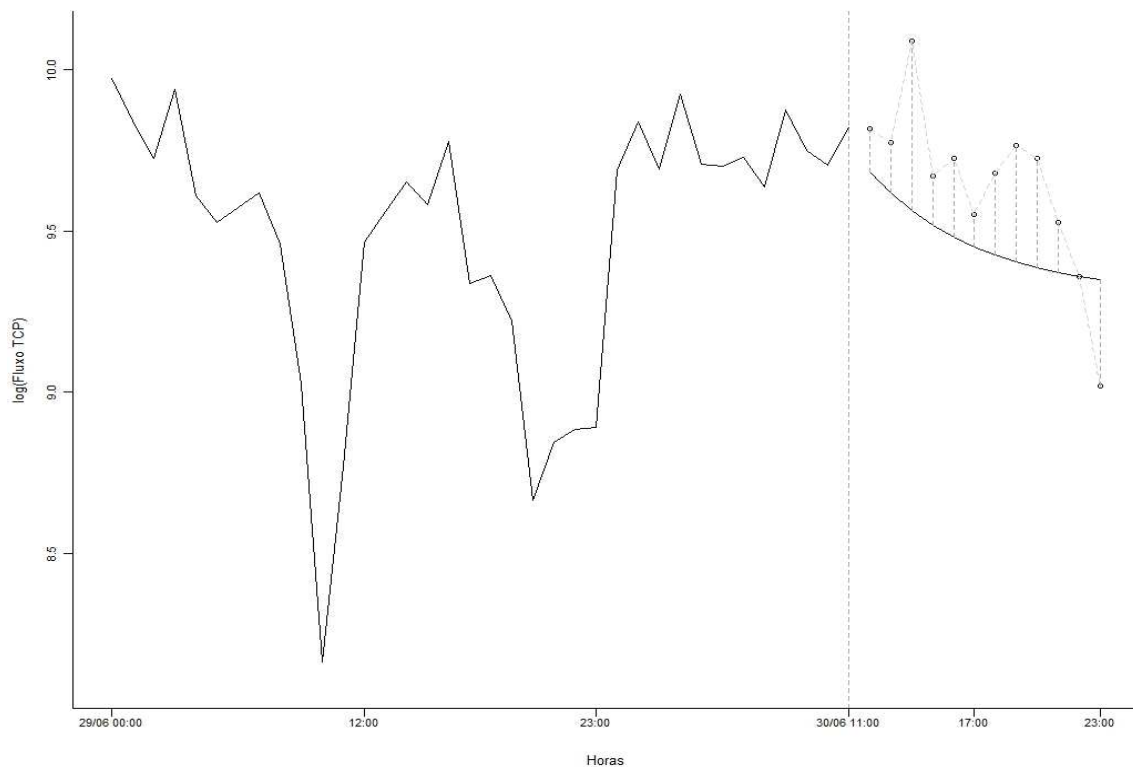


Figura 8.4 - Previsão do comportamento do ruído de fundo da parcela brasileira da Internet, com $p = 1$, $d = 1$ e $q = 2$, no fluxo TCP original, já aplicada a transformação logarítmica, com janela de amostragem hora

Para a janela de amostragem hora, que possui mais observações, os primeiros dados preditos são muito bons quando comparados com os valores observados (os erros quadráticos são 0.01815895 e 0.02457722 para os primeiro e segundo pontos, respectivamente).

E o que se obtém da Figura 8.4? A informação de que, para esta janela de observação, não há um movimento do ruído de fundo que caracterize uma atividade hostil maciça na parcela brasileira da Internet que necessite da geração de um alarme, nas próximas horas.

9 FLUXOS POR PORTAS

De posse da metodologia de análise dos fluxos pode-se estender a análise realizada aos fluxos TCP direcionados às portas do CBH que mais receberam tráfego. Estas portas estão relacionadas na Tabela 4.1.

Como exemplo representativo das portas, tanto para a janela de observação dia como para hora e tanto para os fluxos de varredura como para os de conexão, tomar-se-á, para trabalho, as portas 135, 139, 445, 1080 e 1433.

Além da predição pura e simples, durante a fase de análise das séries temporais destes fluxos alguns outros questionamentos foram investigados, tais como:

- a) o comportamento dos fluxos nas portas está aderente a algum tipo de distribuição estatística; e,
- b) há alguma correlação entre o fluxo de varredura e o de conexão.

Como já utilizado nos capítulos anteriores, para evitar a representação de picos muito desproporcionais ao restante dos dados será usada a série temporal do logaritmo neperiano dos dados.

Como o interesse é a geração de alertas precoces e como já verificado nos capítulos anteriores, a janela de amostragem dia, embora útil para validação do modelo, não se apresenta como uma medida de tempo útil na prática. Assim, neste capítulo, somente serão examinados os fluxos para a janela de amostragem hora.

E, quando não explicitado, estará sendo tratado somente o fluxo de conexão uma vez que a meta é tentar prever um aumento no número de conexões de uma porta indicando a existência de um código malicioso infestando as infraestruturas críticas a partir de um serviço ou congênere que apresenta vulnerabilidades.

Quando da análise das séries em serão utilizados os modelos matemáticos já discutidos

no Capítulo 7 que, para aqueles fluxos, tiveram o gráfico dos p-valores exibido, isto é, os modelos de filtragem e auto-regressivo.

9.1 Aderência dos dados à distribuição normal

Como o modelo estatístico simplificado sendo usado neste trabalho exige que os dados sendo tratados sejam aderentes a uma distribuição normal, parece ser verificar se há aderência entre os dados e a distribuição normal.

Para tal será usado um gráfico de probabilidade – técnica gráfica para avaliar o quanto um conjunto de dados segue uma distribuição como a normal. Os dados são desenhados contra uma distribuição teórica, no caso a normal. Se ele seguirem esta distribuição deve ser formada uma linha reta.

No eixo horizontal dos gráficos apresentam-se valores teóricos da distribuição sendo testada, a normal e, na vertical valores observados. Se confirmada esta característica modelos estatísticos mais simples podem ser aplicados.

Verifica-se, na Figura 9.1 que há diversos intervalos que não apresentam dados. Estas lacunas, embora sejam observações válidas, diminuem o número total de observações da amostra. Entretanto, a despeito destas observações sem valores verifica-se que há aderência da amostra com a distribuição normal.

Esta verificação nos permitirá aplicar a metodologia desenvolvida neste trabalho.

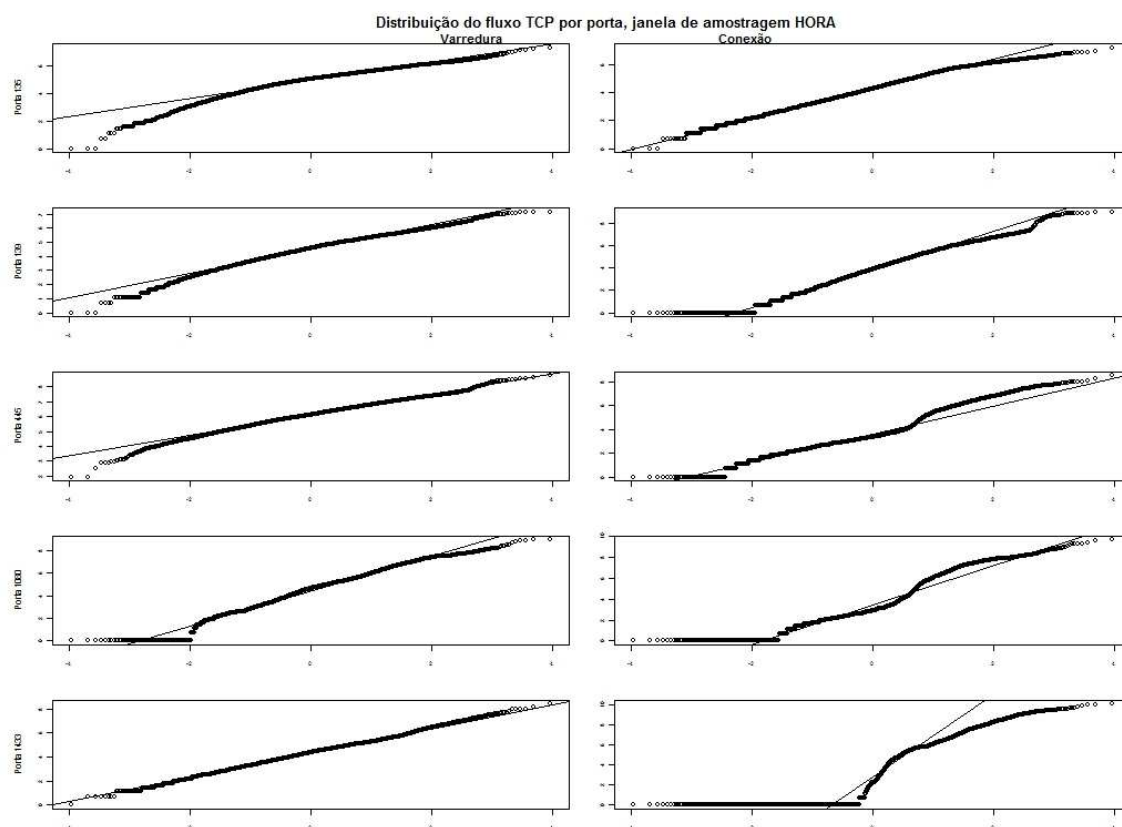


Figura 9.1 - Distribuição do fluxo TCP para portas 135,139, 445, 1080 e 1433, já aplicada a transformação logarítmica, com janela de amostragem hora

9.2 Análise conjunta dos fluxos TCP

Uma vez que as duas séries, tanto de fluxo de varredura como de conexão, se apresentaram aderentes à distribuição normal levantou-se a pergunta se os dois conjuntos de observações pertenceriam a uma mesma distribuição.

Para verificar traça-se um gráfico Q-Q (Quantile-Quantile) – uma técnica gráfica para determinar se dois conjuntos de dados vêm de populações com uma distribuição comum. É uma representação gráfica dos quantis¹ do 1º conjunto de dados contra os quantis do 2º.

¹ Quantil é a fração ou percentual de pontos abaixo de um determinado valor. Por exemplo, o quantil 0,3 é o ponto no qual 30% dos dados localizam-se abaixo do mesmo e 70% acima.

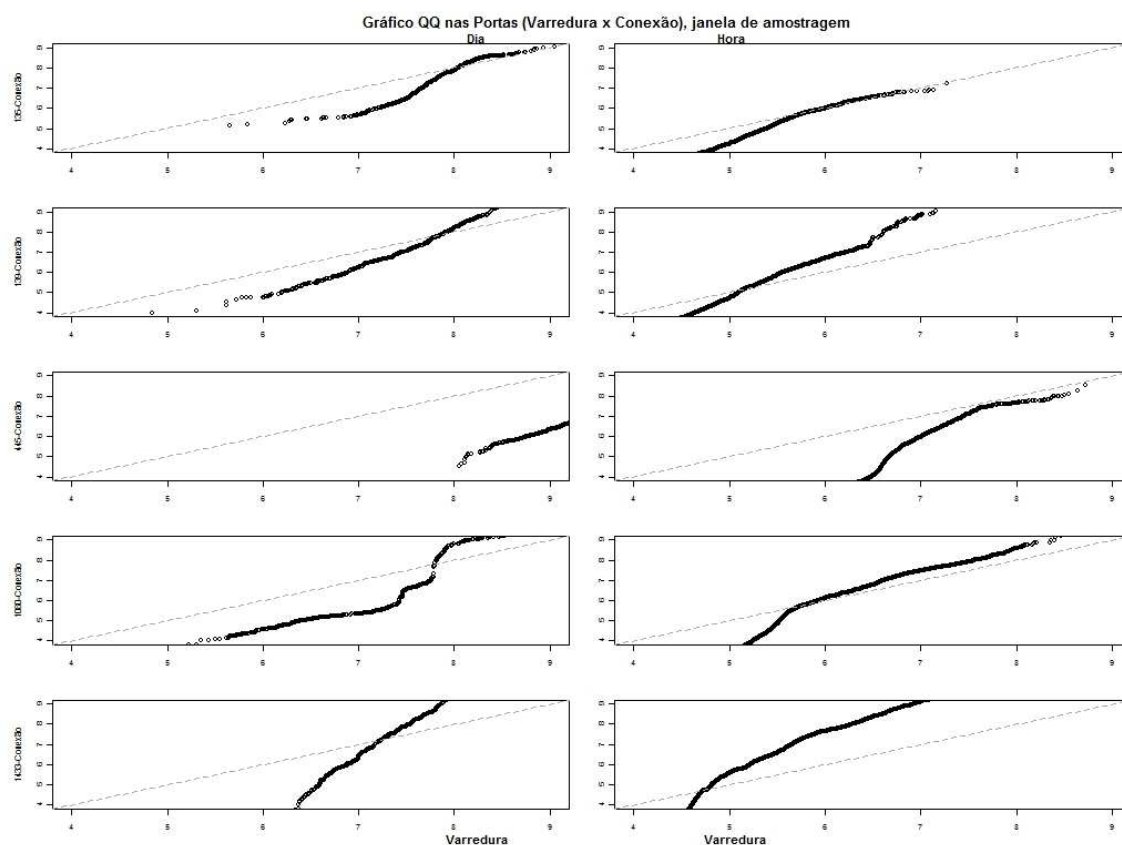


Figura 9.2 - Representação do gráfico QQ para portas 135,139, 445, 1080 e 1433, já aplicada a transformação logarítmica, para as janelas de amostragem dia e hora considerando tráfego de varredura e de conexão

Uma linha de 45° é traçada. Se os dois conjuntos vem da mesma distribuição os pontos devem cair, aproximadamente sobre esta linha de referência. Quanto maior a distância a esta linha mais evidente fica que os conjuntos de dados vem de populações com diferentes distribuições.

Embora não se esteja analisando os fluxos com janela de amostragem dia, na Figura 9.2 faz-se a verificação se os fluxos de conexão e de varredura pertencem a uma mesma distribuição só para efeitos ilustrativos.

Verifica-se, na Figura 9.2 que os fluxos não pertencem à mesma distribuição uma vez que não se encontram sobre a reta inclinada de 45°.

9.3 Correlação entre os fluxos TCP

Examina-se, ainda, se há alguma forma de correlação cruzada entre os fluxos, isto é, se há alguma relação entre o fluxo TCP com menos de 3 pacotes, de varredura, e o com 3 ou mais pacotes, de conexão. Interessa-nos saber se, por exemplo, um certo aumento da varredura implica em um aumento do número de conexões.

O que se deseja é visualizar valores de uma série x_t contra outra série deslocada no tempo de vários intervalos y_{t-h} . A Figura 9.3 apresenta a série do fluxo com menos de 3 pacotes na vertical e a série com três ou mais pacotes, deslocados de um intervalo de tempo, na horizontal.

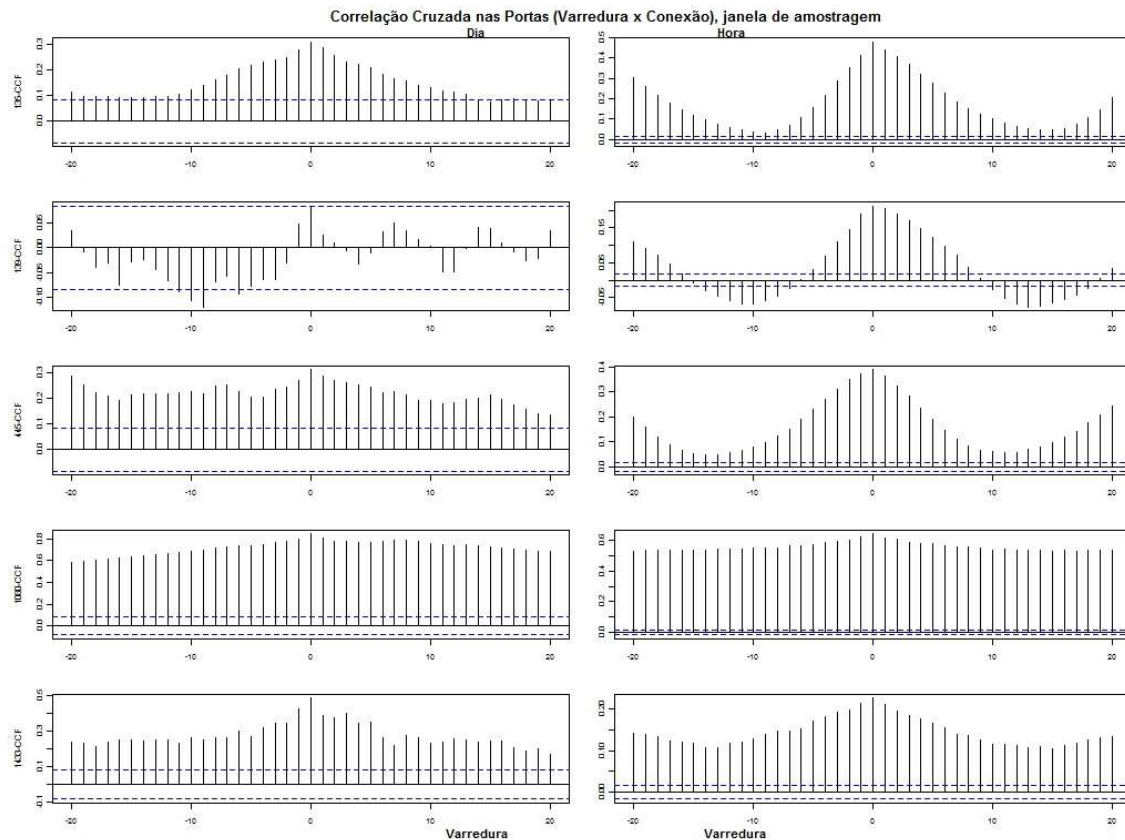


Figura 9.3 - Representação do gráfico de correlação cruzada para portas 135,139, 445, 1080 e 1433, já aplicada a transformação logarítmica, para as janelas de amostragem dia e hora considerando tráfego de varredura e de conexão.

A Figura 9.3 que contém os gráficos de dispersão relacionando os fluxos de varredura e de conexão parece não apresentar nenhuma evidência que correlacione os fluxos. Verifica-se que os picos, os maiores valores, encontram-se quando o valor de afastamento é 0 (zero). O maior valor ser neste ponto é característico de distribuições que não se correlacionam.

9.4 Análise por filtragem

Para utilizar o modelo matemático de filtragem são necessários pesos. Foram testadas janelas de diferentes tamanhos, de 3 a 6, e pesos que variavam entre 0.05 e 0.90, com intervalo de 0.05. O conjunto que apresentou menor erro quadrático foi o com tamanho de janela 3 e pesos 0.9, para o valor mais próximo, e 0.05 para os mais afastados.

A análise dos resíduos foi feita na série resultante da subtração da série filtrada da série original, isto é, na série sem o componente de tendência.

A análise realizada nas séries resultantes, isto é, nas 5 (cinco) séries resultantes da aplicação de filtragem nas séries com fluxos de conexão nas portas selecionadas, não apresentou o gráfico dos p-valores.

Esta ausência indica que a aplicação do modelo de filtragem para remoção do componente de tendência não conduziu a um ruído gaussiano e não deve ser considerado para análise de predição.

9.5 Análise por auto-regressão integrada

No modelo auto-regressivo foram testadas várias combinações de valores para os parâmetros p , d e q . O conjunto que apresentou o menor AIC foi (2, 1, 2) para o fluxo das portas 135, 139 e 1080 e (2, 0, 2) para as portas 445 e 1433.

A Figura 9.4 representa a série resultante da aplicação do modelo auto-regressivo integrado à série representativa dos fluxos TCP, com três ou mais pacotes, para a porta 135 e janela de amostragem hora.

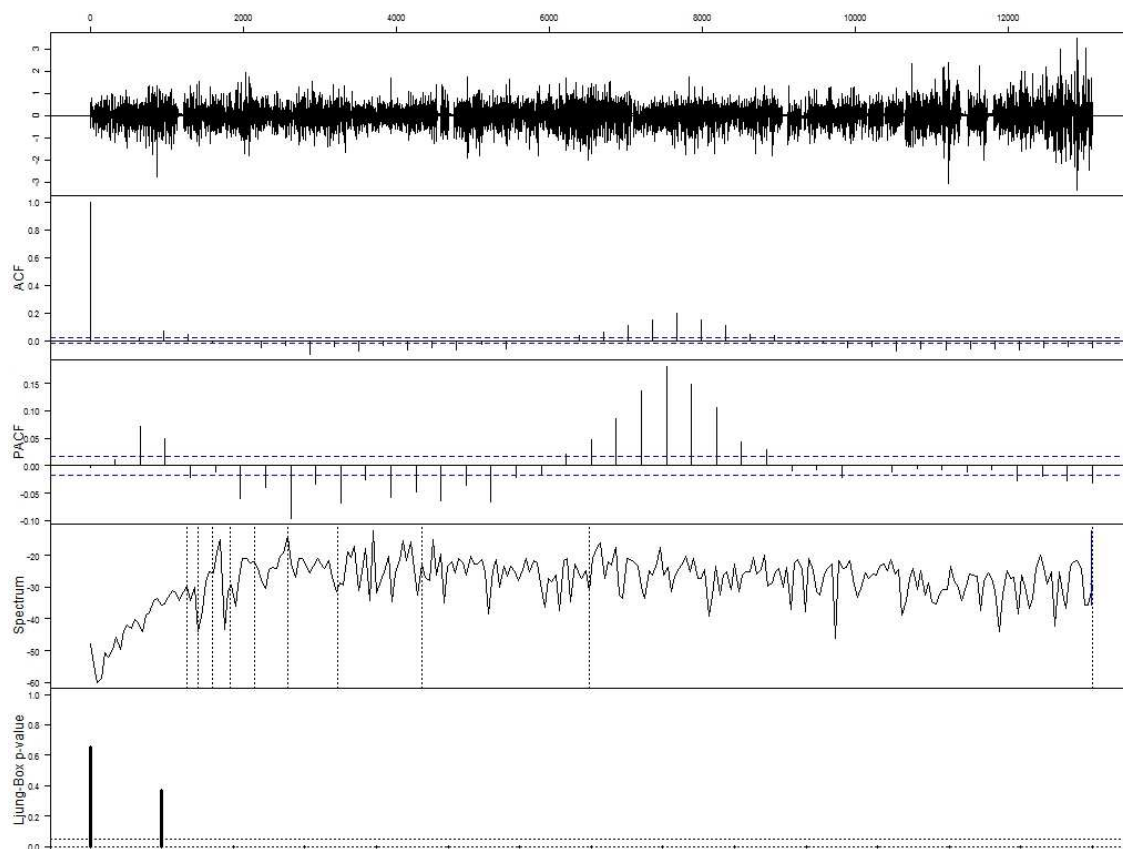


Figura 9.4 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 1$ e $q = 2$, no fluxo de conexão para porta 135, já aplicada a transformação logarítmica, com janela de amostragem hora

A Figura 9.5 representa a série resultante da aplicação do modelo auto-regressivo integrado à série representativa dos fluxos TCP, com três ou mais pacotes, para a porta 139 e janela de amostragem hora.

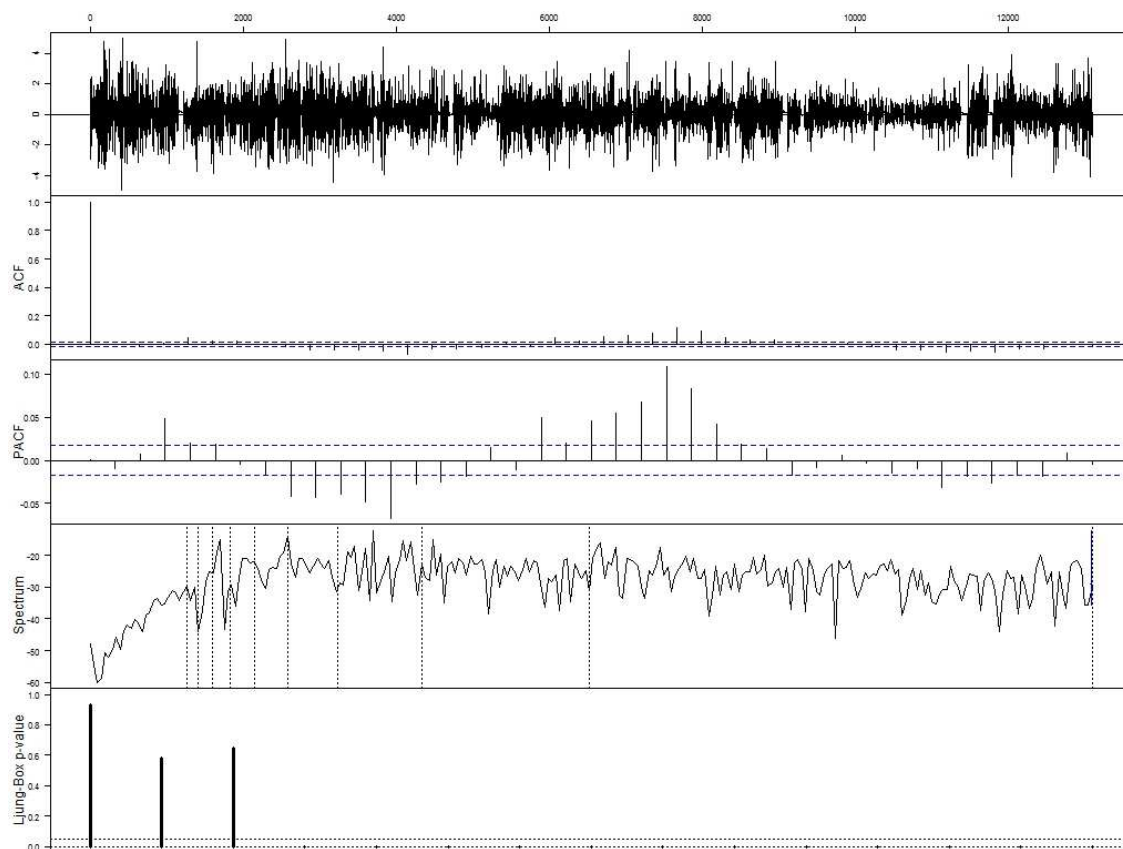


Figura 9.5 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 1$ e $q = 2$, no fluxo de conexão para porta 139, já aplicada a transformação logarítmica, com janela de amostragem hora

A Figura 9.6 representa a série resultante da aplicação do modelo auto-regressivo integrado à série representativa dos fluxos TCP, com três ou mais pacotes, para a porta 445 e janela de amostragem hora.

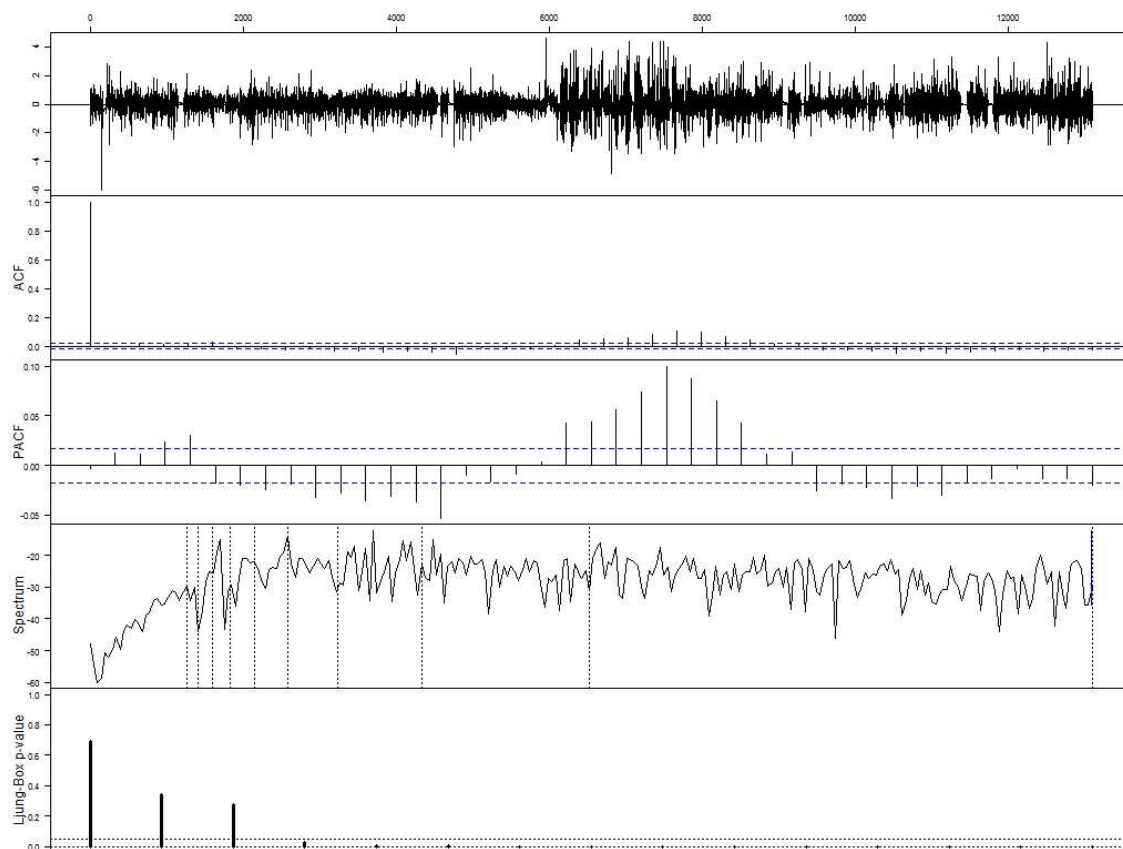


Figura 9.6 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 0$ e $q = 2$, no fluxo de conexão para porta 445, já aplicada a transformação logarítmica, com janela de amostragem hora

A Figura 9.7 representa a série resultante da aplicação do modelo auto-regressivo integrado à série representativa dos fluxos TCP, com três ou mais pacotes, para a porta 1080 e janela de amostragem hora.

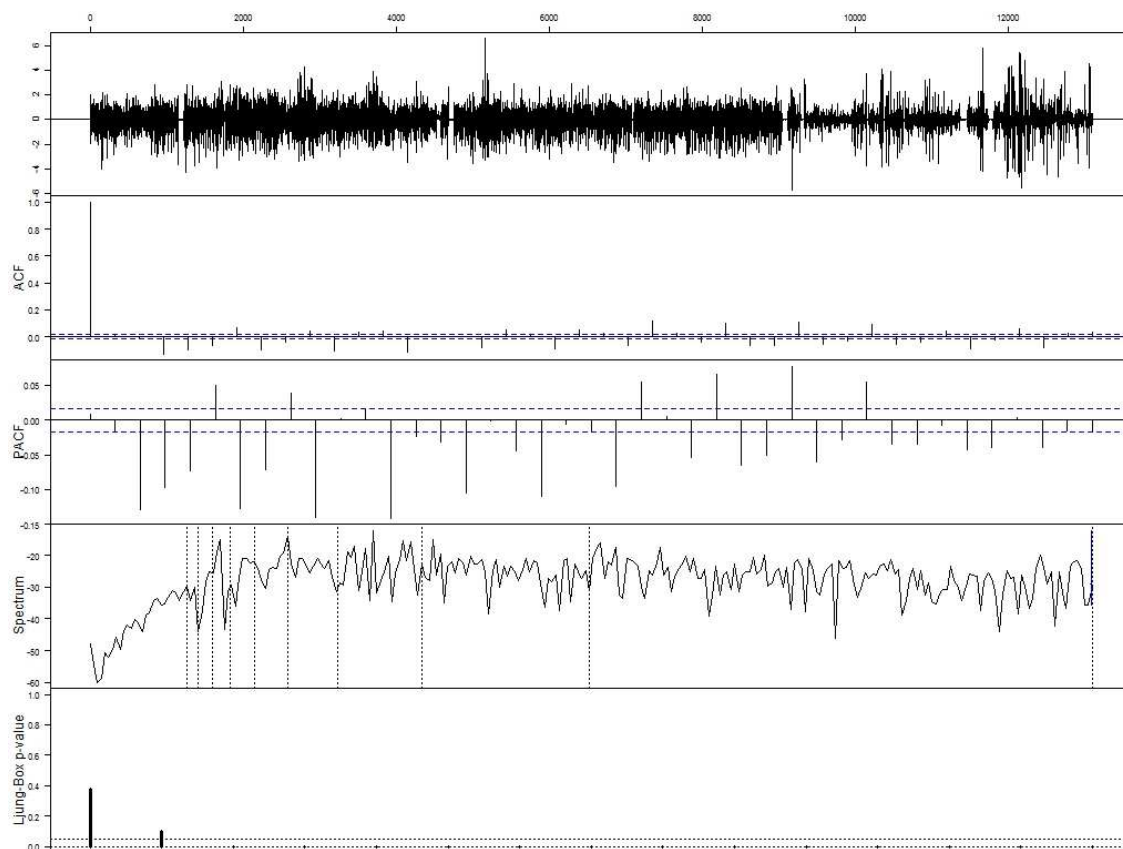


Figura 9.7 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 1$ e $q = 2$, no fluxo de conexão para porta 1080, já aplicada a transformação logarítmica, com janela de amostragem hora

A Figura 9.8 representa a série resultante da aplicação do modelo auto-regressivo integrado à série representativa dos fluxos TCP, com três ou mais pacotes, para a porta 1433 e janela de amostragem hora.

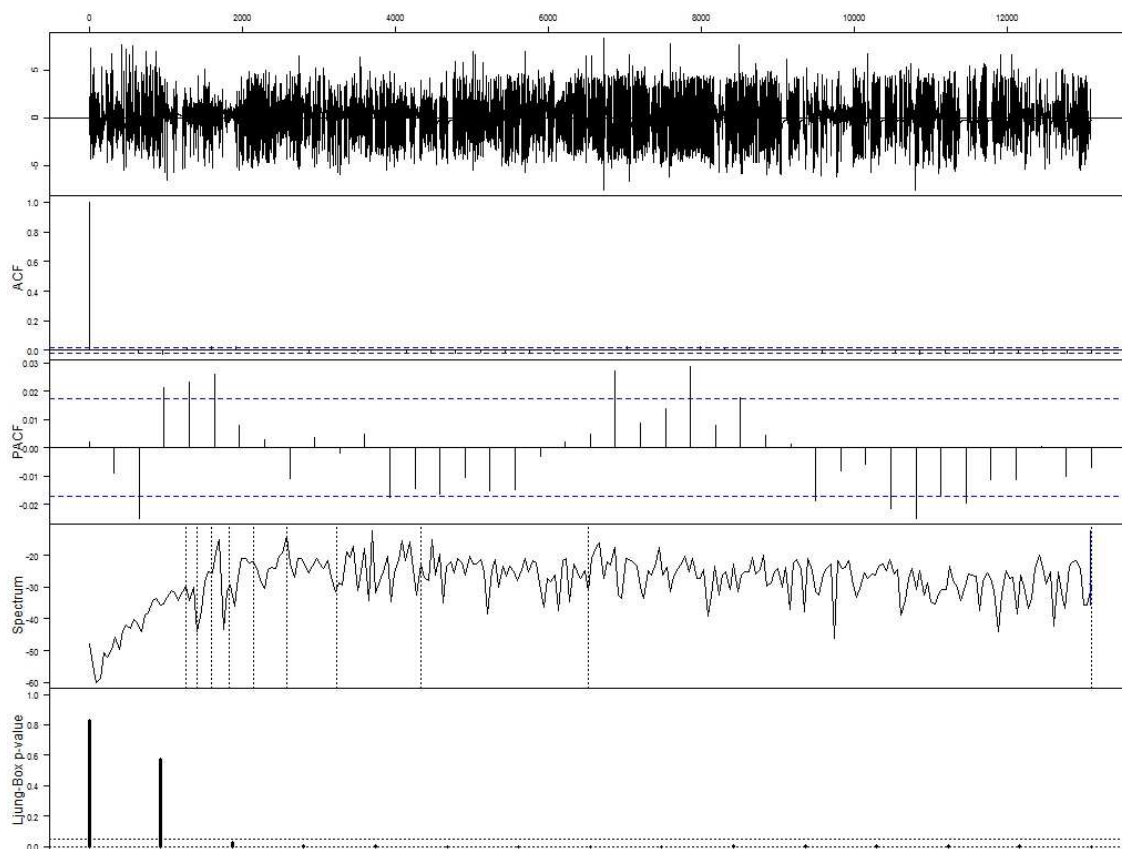


Figura 9.8 - Análise do resíduo resultante da aplicação de auto-regressão com $p = 2$, $d = 0$ e $q = 2$, no fluxo de conexão para porta 1433, já aplicada a transformação logarítmica, com janela de amostragem hora

A análise das séries resultantes, para todas as portas consideradas, apresentou os p-valores indicando que o resíduo gerado é gaussiano e que o modelo pode ser usado para geração de previsões.

9.6 Predição

O processo de predição do comportamento do ruído de fundo será feito considerando-se tão somente os fluxos com 3 ou mais pacotes TCP uma vez que se está desejoso de prever, ou antecipar, a ocorrência de atividades maliciosas na infraestrutura crítica. As fases de varredura dos elementos maliciosos, como já visto anteriormente, não estão correlacionados com os eventos de conexão que caracterizam, ou materializam, o ataque

Assim, pesquisar variações de fluxos com menos de 3 pacotes não necessariamente con-

duziriam a respostas de aumento de atividades de elementos maliciosos. Desta forma são analisados somente os fluxos com 3 ou mais pacotes e com janela de amostragem hora.

Este é outro tópico interessante a ser analisado. Quando do interesse em se tentar prever a infestação maciça de computadores, ou ativos de rede, interconectados à parcela brasileira da Internet, o fator tempo é crucial. Os exemplos dos anos recentes dos worms de infestação mundial mostraram que a curva de propagação é muito rápida. Para que a predição seja eficiente é necessário que o tempo de processamento e de resultados de predição seja compatível com a taxa de propagação.

Não é possível, hoje, com a infraestrutura do CBH realizar processamentos com janelas de amostragem inferior a dia. E, esta janela, não é útil em termos de predição. Entretanto, o amadurecimento da solução prevê que, em breve, ter-se-á a possibilidade de troca de arquivos mais rápidas e o uso de janelas de amostragem hora e, quem sabe, até menores. Assim, o tratamento da predição somente irá considerar o uso da janela de amostragem hora.

E, ainda, devido aos resultados encontrados nas seções anteriores, a predição somente será realizada a partir do modelo matemático auto-regressivo.

Para predição de eventos futuros, com janela de amostragem hora, serão modelados os dados das primeiras 13.092 horas (de 01/01/2005 às 00:00 até 30/06/2006 às 11:00) e preditas as observações das próximas 12 horas. Os valores preditos serão comparados com os dados observados em 30/06/2006, das 12:00 às 23:00.

A Figura 9.9 representa a predição do fluxo TCP para a porta 135 e janela de amostragem hora.

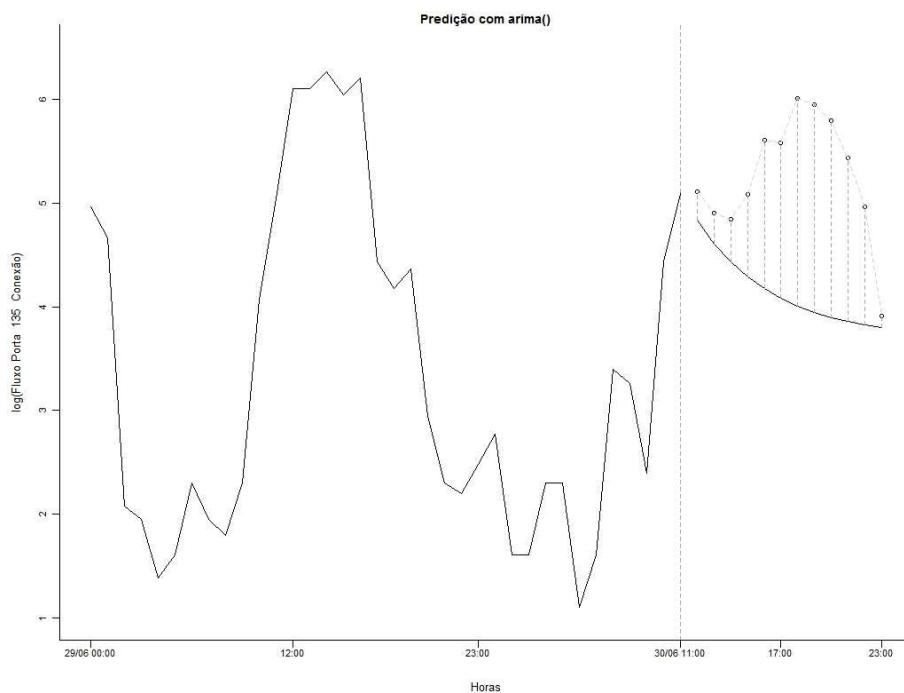


Figura 9.9 - Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 135, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo

Como é possível ver na figura o comportamento esperado para as 3 horas seguintes acompanha o fluxo observado. Desta forma a previsão realizada, para este intervalo que ainda pode ser considerado útil, acompanha os valores observados.

A Figura 9.10 representa a previsão do fluxo TCP para a porta 139 e janela de amostragem hora.

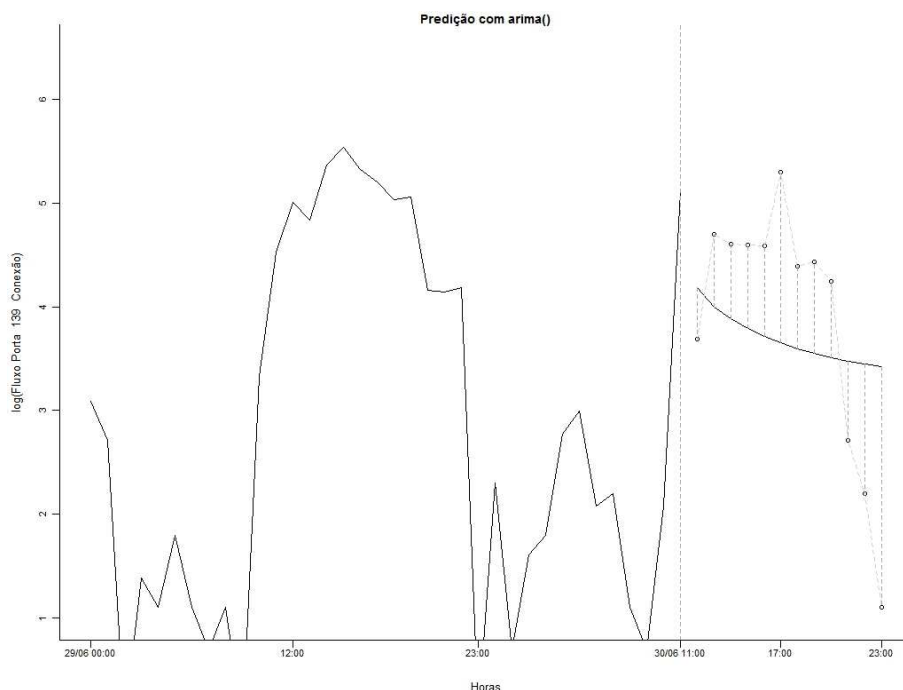


Figura 9.10 - Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 139, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo

Tal como na previsão realizada para a porta 135 a previsão na porta 139 apresenta um comportamento descendente o qual é corroborado pelas observações. As observações das 3 horas seguintes e, neste caso, até 5 horas seguintes são consistentes com as observações. Logo, esta previsão apresenta valores úteis para um sistema de previsão de comportamento do ruído de fundo.

A Figura 9.11 representa a previsão do fluxo TCP para a porta 445 e janela de amostragem hora.

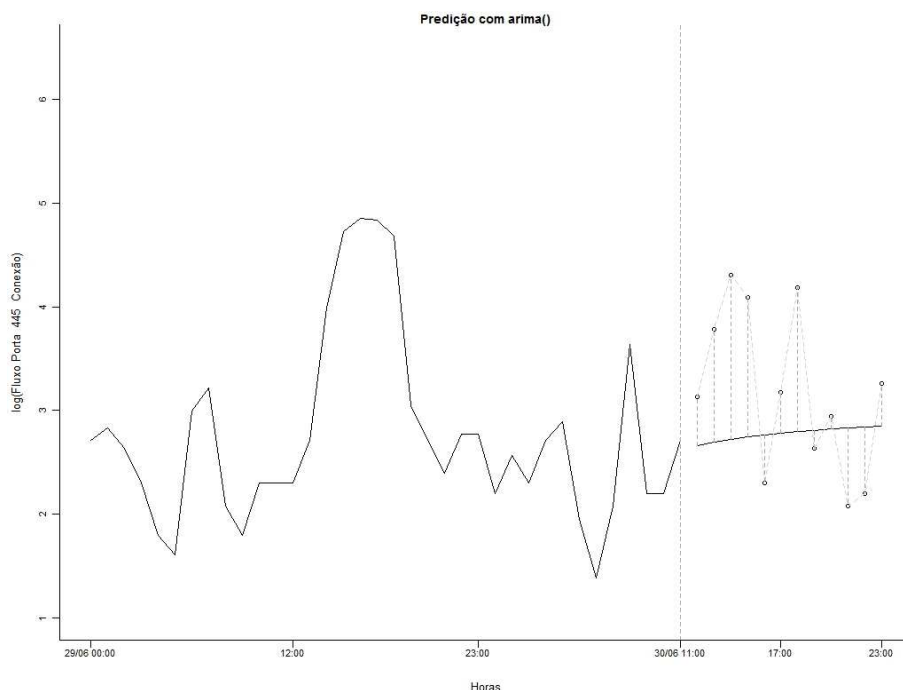


Figura 9.11 - Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 445, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo

A análise da predição na porta 445 apresenta, neste trabalho, a possibilidade de se discutir a geração de um alerta precoce.

A grande pergunta é: uma vez que tenho uma predição quando gerar um alerta precoce? A Figura 9.11 nos permite analisar esta situação.

Inicialmente deve-se reparar que os dados observados já estavam apresentando um movimento ascendente quando se interrompeu as observações e passou-se a gerar as predições. Como um movimento oscilatório dos fluxos é um comportamento aceitável e não um motivo de alerta deve-se analisar não a variação mas a continuidade do movimento ascendente do gráfico indicando uma persistência dos fluxos de conexão num dado intervalo de tempo.

Além da persistência outro fator deve ser analisado, a explosão, isto é, o súbito e rápido crescimento dos valores num pequeno espaço de tempo. Esta observação pode ser encontrada na análise dos worms maliciosos que realizaram ataques mundiais tais como

Code Red, Nimda, etc.

Portanto, acredito que a geração de um alerta somente deve ocorrer numa das seguintes situações:

- a) existência de uma persistência de crescimento por intervalo superior a 3 janelas de observações consecutivas; ou,
- b) crescimento explosivo, isto é, a manutenção de um crescimento superior a 45° em pelo menos duas janelas de observações consecutivas.

O gráfico de predição apresentou somente um dos comportamentos: a característica da persistência. Os dados preditos apresentam, para todos os períodos de predição, uma tendência de crescimento das observações.

Os valores observados também indicam a característica de crescimento entre 11:00 e 14:00 porém sem que, em nenhuma destas janelas de observação haja um crescimento superior a 45° . Neste intervalo o menor dos crescimentos é de 23.144° e o maior de 32.971° .

Portanto, o que fazer neste caso. Acredito ser uma situação em que a predição gerou uma informação que, ao se confirmar na terceira janela de observação a tendência de crescimento continuado um alerta precoce poderia ser gerado e corroborado, ou não, com o passar do tempo. Neste caso um alerta poderia ser gerado às 12:00, mantido até às 15:00 quando poderia ser revisto ou mesmo retirado, se fosse o caso.

As Figuras 9.12 e 9.13 representam a predição do fluxo TCP para as portas 1080 e 1433 e janela de amostragem hora, respectivamente.

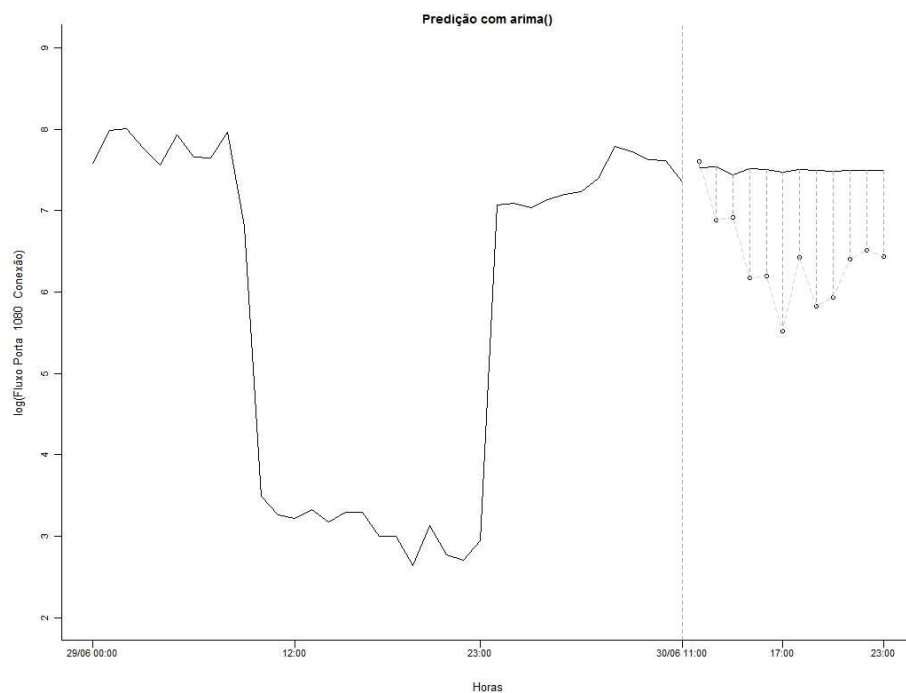


Figura 9.12 - Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 1080, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo

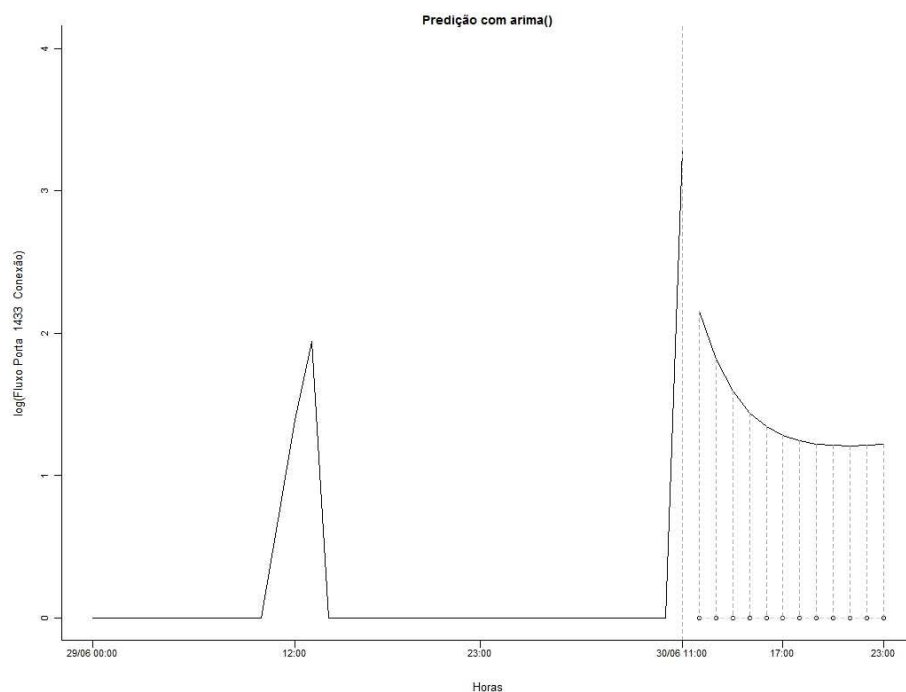


Figura 9.13 - Previsão do comportamento do ruído de fundo, do fluxo de conexões TCP na porta 1433, já aplicada a transformação logarítmica, com janela de amostragem hora e modelo auto-regressivo

A Figura 9.12 mostra um comportamento predito coerente com os valores observados. Uma tendência de diminuição do fluxo na porta 1080.

Já a Figura 9.13 nos permite uma nova análise. Uma porta que tem pouco acesso mantendo os valores observados no 0 ou próximo a ele com alguns picos. Este tipo de comportamento não permite uma predição confiável e é mais fácil considerar o fato de que, em havendo tráfego continuado é mais seguro gerar um alerta.

10 CONCLUSÃO

A pesquisa realizada focou na discussão da necessidade de evolução do atual modelo de segurança, o reativo, para o preditivo. Para tal é necessário a geração de alertas precoces. Esta geração pode ser feita utilizando-se o ruído de fundo da parcela brasileira da Internet obtidos a partir dos dados do Consórcio Brasileiro de Honeypots (CBH).

De posse dos dados foi necessária a criação de uma metodologia para sanitizar os dados brutos recebidos e permitir seu tratamento. Várias possibilidades foram levantadas e algumas testadas sendo que a análise a partir de séries temporais foi a utilizada por esta pesquisa.

O que se deve destacar é que, mesmo com o tráfego da Internet sofrendo mudanças ao longo do tempo, esta pesquisa foca nos dados constituintes do ruído de fundo e, sendo assim, a metodologia proposta permanece válida, independentemente do perfil de tráfego.

Como são vários os passos realizados para a consecução do trabalho, por motivos didáticos a conclusão será dividida em tópicos que abordarão as várias conclusões parciais obtidas ao longo de todo o trabalho.

10.1 Modelo preditivo de segurança

Segurança da informação tem sido, tradicionalmente, tratada somente por especialistas, de forma defensiva, isto é, os responsáveis pela segurança da informação são técnicos que tentam defender a organização através da detecção de falhas na estrutura de defesa que, quando visualizadas, geram processos de reação e de melhorias.

Para que a segurança da informação possa evoluir e fazer frente às novas ameaças a que se está sujeito nos dias atuais, dois paradigmas precisam ser quebrados ou, pelo menos, evoluídos. O primeiro é que segurança é um assunto eminentemente técnico. Não o é! Segurança da informação também é técnico mas é, preliminarmente um processo administrativo que impacta o negócio da empresa. É um processo de gestão que deve estar

alinhado com os objetivos estratégicos da organização. Entretanto, por não ser o foco deste trabalho este assunto não será mais aprofundado aqui.

O segundo paradigma é o do modelo defensivo de segurança. Neste modelo o inimigo tem sempre a iniciativa! Tem-se, sempre, que esperar a concretização de uma atividade maliciosa para que esta possa ser estudada e melhorias no sistema de defesa possam ser propostas e implementadas.

O que se imagina é um novo modelo que rompe o paradigma atual: o preditivo! É importante que os dois modelos complementem-se e não sejam mutuamente excludentes. Porém, a partir deste novo modelo, mecanismos que possam prever atividades maliciosas e as deter, de alguma forma, ainda no seu estágio inicial, estarão em pleno funcionamento.

Na área da segurança da informação ainda há poucos resultados na literatura especializada sobre predição de eventos. Nos trabalhos realizados até hoje os pesquisadores vem aplicando técnicas de controle estatísticos e de validação, através de análise de regressão sobre dados coletados em pequenas redes para geração de alertas precoces com antecedência máxima de um dia.

Este trabalho extrapolou os limites das pesquisas já realizadas aplicando as mesmas técnicas, mas visando a segurança da parcela brasileira da Internet como um todo, gerando eventos preditos com horas de antecedência o que é fundamental para adoção do novo modelo de segurança.

10.2 Características do ruído de fundo

Este trabalho analisou dados coletados pelo CBH durante 546 dias e os organizou de forma a comporem séries temporais agrupados por janelas de amostragem dia e hora. Os dados capturados pelos diversos sensores do CBH constituem o ruído de fundo!

Dentro do ruído de fundo está o tráfego malicioso que interessa ser estudado, entendido e predito além do tráfego não desejado causado por erros de configuração, dentre outros.

Portanto, dentro do tráfego capturado encontra-se a atividade desenvolvida pelos softwares maliciosos que circulam na parcela brasileira da Internet.

Para entender o comportamento dos dados utilizados e se os mesmos apresentam características que possam auxiliar na detecção de anomalias determinou-se alguns parâmetros observáveis das séries temporais. Ao conjunto de valores levantados chamou-se de “característica de normalidade”. Estes valores permitem uma forma de caracterização do ruído de fundo.

O conhecimento dos padrões comportamentais do ruído de fundo é um auxiliar na detecção de atividades não esperadas. Variações neste comportamento podem indicar comportamentos mapeados de agentes infecciosos.

Este trabalho usa, ainda, o conceito de fluxo em substituição ao de tráfego ou de pacotes. Isto foi necessário em razão do tipo de dado disponibilizado pelo CBH.

A “característica de normalidade”, isto é, o que se pode esperar receber numa máquina ao conectá-la na Internet, é constituída por:

- a) 90% de fluxos TCP, 7% UDP e 3% ICMP;
- b) 80% de fluxos < 3 e 20% de fluxos ≥ 3 ;
- c) cada sensor apresenta média de 14.383 ± 33.479 fluxos < 3 .

Estes números representem a “característica de normalidade” da amostra trabalhada. Mesmo se considerada a evolução esperada dos serviços disponíveis para o usuário final estas características tem a tendência de se manterem constantes, ou próximas dos valores obtidos, com o passar do tempo.

Mesmo sendo os dados usados de 2005 e, nos dias atuais, serviços tem sido disponibilizados utilizando-se diferentes protocolos, as características levantadas dizem respeito ao ruído de fundo e não ao tráfego legítimo. É importante ressaltar que as estatísticas dis-

poníveis no sítio oficial do CBH¹ comprovam que as características aqui levantadas com dados de 2005 e 2006 permanecem válidas, particularmente a relação entre os protocolos TCP, UDP e ICMP.

A relação entre varreduras e conexões pode não ser uma relação fixa de 4 x 1 mas, com certeza, indica que circula na rede um número muito maior de varreduras do que de conexões. A ordem de grandeza pode variar mas o fato de ser muito superior, não.

O número de fluxos recebidos por um sensor é um alerta ao administrador. Cada qual terá seu valor próprio. Porém, todo gerente deve incluir na sua capacidade de tráfego, pelo menos, o valor médio levantado. Este número pode e deve ser atualizado ao longo do tempo para que seja mais representativo da realidade daquele momento. Entretanto, serve como base para análise e permite que, quando comparado aos números levantados por outros métodos, verifique-se a existência, ou não, de particularidades da parcela brasileira da Internet.

Características auxiliares levantadas mostram que a grande maioria das varreduras concentra-se em um pequeno número de portas. O mesmo ocorre com a quantidade de portas para fluxos com 3 ou mais pacotes, porém, com uma distribuição diferente.

A grande maioria do ruído de fundo tem origem em endereços IP iniciados por 200, faixa de endereços distribuídos ao Brasil. No tráfego de fluxos com menos de 3 pacotes nem todos os endereços são usados porém, nos fluxos com 3 ou mais pacotes, todos os endereços IP, válidos ou não, são usados. Caracteriza-se, assim, tentativas de ataques usando IP forjados.

10.3 Análise estatística dos dados como séries temporais

Várias tentativas de análise foram realizadas ao longo da pesquisa: programação dinâmica, neuro-computação, etc. Outras foram cogitadas mas não chegaram a ser abordadas como filtro de Kalman, por exemplo. De tudo o que foi pesquisado decidiu-se que a análise inicial a ser feita, particularmente em razão do tipo de dado disponível, seria a

¹ <http://www.honeypots-alliance.org.br/stats/flows/current/#proto>

representação dos mesmos como uma série temporal e a análise desta série a partir dos modelos estatísticos clássicos.

Uma vez tendo sido sanitizados, os dados foram representados como séries temporais. A primeira análise realizada procurava verificar se, estatisticamente, as séries temporais dos dados do CBH eram aderentes à alguma distribuição conhecida.

Verificou-se que as séries são aderentes à distribuição normal! Desta forma técnicas estatísticas consagradas podem ser utilizadas para analisar os dados.

Dos vários modelos matemáticos testados os que apresentaram melhor resultado, para este trabalho isto significa que a série resultante possui ruído gaussiano – na análise realizada tem-se a representação do gráfico dos p-valores – foram a filtragem e o modelo auto-regressivo integrado.

Para a realização de predição as séries com janela de amostragem dia foram usadas somente para validação do modelo matemático. Os valores de interesse foram determinados a partir da janela de amostragem hora. O uso de uma janela de amostragem menor aumenta a quantidade de dados disponíveis para processamento melhorando a predição.

Para testar a aderência da predição realizada aos dados observados foram modelados os dados das primeiras 13.092 horas (de 01/01/2005 às 00:00 até 30/06/2006 às 11:00) e preditas as observações das próximas 12 horas. Os valores preditos foram comparados com os dados observados.

Os valores preditos para ambos os modelos matemáticos aplicados apresentam um erro quadrático aceitável para as previsões das próximas 3 horas.

10.4 Geração de alertas

A geração de alertas precoces é dependente não somente do cálculo e da análise estatística mas do cenário nos quais os dados tratados estão inseridos. Dentro deste contexto optou-se por estipular parâmetros a partir do estudo do comportamento de agentes maliciosos como os worms.

O comportamento esperado de um worm é uma fase inicial de descoberta de vulnerabilidades aonde ele, geralmente, não consegue ser detectado. A partir daí entra numa fase explosiva que, se traçada graficamente, indica uma representação exponencial uma vez que a propagação é muito rápida. É nesta fase, entre o término da varredura e o início da contaminação maciça que se deve concentrar esforços para a detecção deste agente.

Como as máquinas do CBH representam um conjunto de endereços válidos da Internet brasileira e são máquinas sujeitas a serem exploradas com facilidade, a probabilidade de que estes agentes maliciosos sejam detectados nestes sensores, nesta fase de contaminação inicial, é alta. Daí se esperar que o comportamento gráfico do ruído de fundo também apresente um crescimento explosivo ao longo do tempo assim como uma permanência desta tendência ascendente até que se encontrem vacinas e as mesmas possam ser aplicadas.

Conclui-se, portanto, que o seguinte critério pode ser adotado para a geração de alerta precoce:

- a) existência de uma persistência de crescimento por intervalo superior a 3 janelas de observações consecutivas; ou,
- b) crescimento explosivo, isto é, a ocorrência de um crescimento com inclinação superior a 45° entre 3 janelas de observações consecutivas.

Os dados do CBH, quando analisados como um todo, mostraram, na predição realizada, que não há tendência de variação explosiva nem de crescimento constante ao longo do tempo.

Ao contrário, o que verificou com a predição foi a tendência de manutenção do status atual ou mesmo uma pequena elevação com consequente diminuição o que foge à característica do comportamento de ataques maciços não sendo necessária a geração de alertas.

Na análise realizada no agrupamento das observações, por porta, foram seleccionadas 5

das portas com maior representatividade de fluxo: 135, 139, 445, 1080 e 1433.

Foi analisado somente o fluxo TCP com 3 ou mais pacotes uma vez que se determinou que não há uma correlação explícita entre aumento e/ou diminuição de varreduras com aumentos e/ou diminuições de conexões. Como o foco é a determinação precoce, ainda na fase inicial de contágio, optou-se por examinar o comportamento do fluxo das conexões que, no caso da presença de agente malicioso, deve apresentar comportamento explosivo e/ou ascendente ao longo do tempo.

As predições apresentaram resultados semelhantes ao da análise do fluxo total. Entretanto, para a série temporal representativa dos fluxos na porta 445 há uma característica diferente: apresenta uma tendência ascendente.

É uma tendência persistente porém não explosiva, isto é, sua variação não é superior a 45°. Neste caso a informação gerada pela predição, e confirmada pelas observações, seria a geração de um alerta que seria retirado a partir do momento que se verificasse o encerramento da tendência ascendente.

10.5 Sugestões para pesquisas futuras

Sugere-se, a seguir, alguns temas para pesquisas futuras:

- a) a busca por correlações ou tendências entre os endereços IP de origem e as portas;
- b) uso de metodologias de discretização de grandes conjuntos de dados, como os citados em Lin (2003) para geração, ou busca, de padrões de ataques (assinaturas);
- c) combinação de métodos estatísticos. Uso de série filtrada como entrada de um processo auto-regressivo, por exemplo. Montagem de modelos matemáticos que permitam o uso desta combinação na predição de eventos.

REFERÊNCIAS BIBLIOGRÁFICAS

ARCE, I.; LEVY, E. An analysis of the slapper worm. **IEEE Security & Privacy**, v 1, n. 1. p 82-87, 2003.

BOX et al. **Time series analysis: forecasting & control**. 4. Hoboken, New Jersey: John Wiley & Sons, 2008. p. 746. ISBN: 978-0-470-27284-8.

BAILEY et al. The blaster worm: then and now. **IEEE Security & Privacy**, v 3, n 4, p 26-31, 2005.

BARROS, E. G. et al. Características iniciais do ruído de fundo da internet brasileira a partir de dados do CBH. In: WORKSHOP DOS CURSOS DE COMPUTAÇÃO APLICADA DO INPE, 7.,2007, São José dos Campos-SP. **Anais...** Disponível em <<http://mtc-m18.sid.inpe.br/rep/sid.inpe.br/mtc-m18@80/2010/04.20.18.37>>. Acesso em: 10 de novembro de 2010.

BARROS, E. G. et al. Análise das séries temporais formadas por dados do CBH. In: WORKSHOP DOS CURSOS DE COMPUTAÇÃO APLICADA DO INPE, 8.,2008, São José dos Campos-SP. **Anais...** Disponível em <<http://mtc-m18.sid.inpe.br/rep/sid.inpe.br/mtc-m18@80/2010/07.22.17.32>>. Acesso em: 10 de novembro de 2010.

BARROS, E. G. et al. Características do ruído de fundo da internet brasileira a partir de dados do CBH. In: WORKSHOP DOS CURSOS DE COMPUTAÇÃO APLICADA DO INPE, 10.,2010a, São José dos Campos-SP. **Anais...** Disponível em <<http://mtc-m18.sid.inpe.br/rep/sid.inpe.br/mtc-m18/2010/09.27.19.46>>. Acesso em: 10 de novembro de 2010.

BARROS, E. G. et al. **Predição de eventos na segurança da informação: um novo paradigma?** Aceito pela revista Datagramazero, julho de 2010b.

CAIDA. **Analysis of code-red**. Disponível em: <<http://www.caida.org/research/security/code-red/>>. Acesso em: 13 de outubro de 2010.

CONSÓRCIO BRASILEIRO DE HONEYPOTS – CBH. **Projeto honeypots distribuídos**. Disponível em: <<http://www.honeypots-alliance.org.br/index-po.html>>. Acesso em: 13 de outubro de 2010.

DAGON, D. et al. HoneyStat: local worm detection using honeypots. In: RECENT ADVANCES IN INTRUSION DETECTION INTERNATIONAL SYMPOSIUM, RAID 2004. 7., Sophia Antipolis, France. **Proceedings...** Sophia Antipolis, France: Springer Berlin/Heidelberg, 2004. p. 39 - 58. v. 3224/2004. Lecture Notes in Computer Science.

GRIZZARD, J. et al. Flow based observations from NETI@home and honeynet data. In: IEEE WORKSHOP ON INFORMATION ASSURANCE AND SECURITY. United States 5., 2005, West Point, NY. **Proceedings...** West Point, NY, USA: Military Academy, 2005.

GUIMARÃES, A.M.C. **Empresas de gestão conservadora: potencial de previsão de demanda e simulação computacional.** 100 p. (PUC-RIO 061499/CA). Dissertação (Mestrado em Engenharia Industrial). Pontifícia Universidade Católica do Rio de Janeiro, 2008.

HOEPERS, C. et al. **Honeypots e honeynets: definições e aplicações.** Disponível em: <<http://www.cert.br/docs/whitepapers/honeypots-honeynets/>>. Acesso em: 13 de outubro de 2010.

JENSEN, P. **Operations research models and methods.** Disponível em: <<http://www.me.utexas.edu/~jensen/ORMM/omie/operation/unit/forecast/index.html>>. Acesso em: 13 de outubro de 2010.

KIRK A. **Responding to the modern threat landscape: flexibility in the face of change.** Brasília, DF, 2010. Palestra realizada no Blackbull Security Day no Mercure Hotel Brasília em 14 de abril de 2010.

LEVINE, J. et al. The use of honeynets to detect exploited systems across large enterprise networks. In: IEEE WORKSHOP ON INFORMATION ASSURANCE AND SECURITY. United States 3., 2003, West Point, NY. **Proceedings...** West Point, NY, USA: Military Academy, 2003.

Lin, J. et al. A symbolic representation of time series, with implications for streaming algorithms. IN: 8th ACM SIGMOD WORKSHOP ON RESEARCH ISSUES IN DATA MINING AND KNOWLEDGE DISCOVERY. United States, 2003, New York, NY. **Proceedings...** New York, NY, USA, 2003.

MCCARTY, B. Botnets: big and bigger. **IEEE Security & Privacy**, v 1, n 4, p 87-90. 2003.

MOORE et al. Inside the Slammer Worm. **IEEE Security & Privacy**, v 1, n 4. p 33-39. 2003.

PANG, R. et al. Characteristics of Internet background radiation. In: INTERNET MEASUREMENT CONFERENCE, Itália, 2., 2004, Taomina, Itália. **Proceedings...** Taomina, Itália, 2004.

POLLOCK, D.S.G. **A Handbook of time series analysis, signal processing and dynamics**. 1. London, UK: Academic Press (The University Press, Cambridge), 1999. p. 782. ISBN: 978-0-12-560990-6.

RICHARDSON, D. et al. The limits of global scanning worm detector in the presence of background noise. In: ACM WORKSHOP ON RAPID MALCODE, United States, 3., 2005, Fairfax, VA. **Proceedings...** Fairfax, VA, USA, 2005.

SAVAGE, S. et al. **Center for Internet epidemiology and defenses**. Disponível em: <<http://www.cs.ucsd.edu/~savage/papers/CIEDProposal.pdf>>. Acesso em: 13 de outubro de 2010.

SHUMWAY, R. H.; STOFFER, D. S. **Time series analysis and its applications: with R examples**. 2. New York, NY, USA: Springer Science+Business Media, 2006. p. 575. ISBN: 978-0-387-29317-2.

STANIFORD et al. **How to own the Internet in your spare time**. Disponível em: <<http://www.icir.org/vern/papers/cdc-usenix-sec02/>>. Acesso em: 13 de outubro de 2010.

STATSOFT, INC. **STATISTICS methods and applications**. 1. Tulsa, OK, USA: Statsoft Inc, 2005. p. 828. ISBN: 1-884233-59-7. Disponível em: <<http://www.statsoft.com/textbook/>>. Acesso em: 13 de outubro de 2010.

THE HONEYNET PROJECT & RESEARCH ALLIANCE. **Know your enemy: tracking botnets**. Disponível em: <<http://www.honeynet.org/papers/bots/>>. Acesso em: 13 de outubro de 2010a.

THE HONEYNET PROJECT & RESEARCH ALLIANCE. **Know your enemy: statistics**. Disponível em: <<http://old.honeynet.org/papers/stats/>>. Acesso em: 13 de outubro de 2010b.

TSAY, R. S. **Analysis of financial time series**. 1. New York, NY, USA: Jon Wiley & Sons, 2002. p. 448. ISBN: 0-471-41544-8.

VANDERAVERO, N. et al. The HoneyTank: a scalable approach to collect malicious Internet traffic. In: INTERNATIONAL INFRASTRUCTURE SURVIVABILITY WORKSHOP (IISW'04), Portugal, 1., 2004, Lisbon, Portugal. **Proceedings...** Lisbon, Portugal, 2004.

VIINIKKA J. et al. **Monitoring IDS background noise using EWMA control charts and alert information**. Springer Berlin/Heidelberg, 2004. p. 166 - 187. v. 3224/2004. Lecture Notes in Computer Science.

VIINIKKA, J. et al. Time series modeling for IDS alert management. In: ACM SYMPOSIUM ON INFORMATION, COMPUTER AND COMMUNICATIONS SECURITY (ASIACCS'06), Twain, 1., 2006, Taipei, Twain. **Proceedings...** Taipei, Twain, 2006.

YIN, C. et al. Honeypot and scan detection in intrusion detection system. In: CANADIAN CONFERENCE ON ELECTRICAL AND COMPUTER ENGINEERING, Canada, 17., 2004, Ontario, Canada. **Proceedings...** Ontario, Canada, 2004.