CARACTERIZAÇÃO E MODELOS PARA AVALIAR O DESEMPENHO DE REDES DE ACESSO RESIDENCIAL BASEADOS EM APRENDIZADO DE MÁQUINA

Gustavo Henrique Alves dos Santos

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

Orientador: Edmundo Albuquerque de Souza e Silva

Rio de Janeiro
Março de 2015

CARACTERIZAÇÃO E MODELOS PARA AVALIAR O DESEMPENHO DE REDES DE ACESSO RESIDENCIAL BASEADOS EM APRENDIZADO DE MÁQUINA

Gustavo Henrique Alves dos Santos

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

Examinada por:

Prof. Edmundo Albuquerque de Souza e Silva, Ph.D.

Prof. Ana Paula Couto da Silva, D.Sc.

Prof. Rosa Maria Meri Leão, Dr.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2015

Santos, Gustavo Henrique Alves dos

Caracterização e modelos para avaliar o desempenho de redes de acesso residencial baseados em aprendizado de máquina/Gustavo Henrique Alves dos Santos. – Rio de Janeiro: UFRJ/COPPE, 2015.

XIII, 92 p.: il.; 29, 7cm.

Orientador: Edmundo Albuquerque de Souza e Silva

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

Referências Bibliográficas: p. 62 – 65.

1. Redes de acesso residencial. 2. Avaliação de desempenho. 3. Aprendizado de máquina. I. Silva, Edmundo Albuquerque de Souza e. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Agradecimentos

Gostaria de agradecer primeiramente a minha família, minha mãe Rosane, meu padrasto Máiquel e meus irmãos Cássio, Pedro e Marco por todo o suporte durante a minha longa formação profissional. Nada do que obtive seria possível sem todo o apoio dado por eles durante os muitos percalços desta longa jornada. Agradeço também ao Cássio e a minha cunhada Juliana por trazerem ao mundo minha sobrinha e afilhada Ana Luísa, uma pequena faixa de luz que ilumina qualquer ambiente em que se encontra.

Agradeço às professoras Rosa e Ana Paula pelas importantes contribuições dadas para o resultado final deste trabalho. Agradeço ao meu orientador Edmundo por toda a ajuda e orientação durante estes 3 anos de trabalho, que contribuíram imensamente para o meu crescimento profissional.

Agradeço a meus colegas de TGR, Gabriel e Guilherme, por suas contribuições diretas e indiretas para o resultado final deste trabalho. Agradeço também a todos os colegas de LAND pelas discussões técnicas e filosóficas durante todo este período e a Carol por toda a ajuda e gentileza.

Por fim, mas não menos importante, agradeço imensamente a minha namorada Andressa pela enorme compreensão com meus constantes períodos de indisponibilidade e por todo o apoio dado durante este período cheio de desafios.

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

CARACTERIZAÇÃO E MODELOS PARA AVALIAR O DESEMPENHO DE REDES DE ACESSO RESIDENCIAL BASEADOS EM APRENDIZADO DE MÁQUINA

Gustavo Henrique Alves dos Santos

Março/2015

Orientador: Edmundo Albuquerque de Souza e Silva

Programa: Engenharia de Sistemas e Computação

A medida que o acesso a internet utilizando redes residenciais ganha importância para a sociedade se torna essencial garantir que a qualidade do serviço oferecido seja adequada. Diversos trabalhos na literatura realizam medições de desempenho em redes residenciais, mas poucos propõem a criação de modelos a partir dos dados coletados. Neste trabalho são analisados o processo de perda e a dinâmica do tráfego observado em redes de acesso residencial. Propõe-se a utilização de modelos baseados em técnicas de aprendizado de máquina, criados a partir dos dados coletados por um *software* embarcado em roteadores residenciais. Os resultados obtidos mostram que cadeias de Markov ocultas conseguem modelar de maneira precisa as distribuições do tamanho das rajadas de perda e intervalo entre perdas. Observa-se que o tráfego medido possui alta autocorrelação e apresenta rajadas em diferentes escalas de tempo, característica que ainda não foi observada em tráfego residencial. Mostra-se também que cadeias de Markov ocultas podem ser utilizadas para representar de maneira precisa as características do tráfego.

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

CHARACTERIZATION AND MODELS FOR PERFORMANCE EVALUATION OF RESIDENTIAL ACCESS NETWORKS BASED ON MACHINE LEARNING

Gustavo Henrique Alves dos Santos

March/2015

Advisor: Edmundo Albuquerque de Souza e Silva

Department: Systems Engineering and Computer Science

Residential access to the Internet has become increasingly relevant to society. Therefore, it is essential that the service provided to clients has an adequate level of performance. There are a few recent papers in the literature that report results from measurement studies performed at the home gateway router. But the literature lacks studies aimed at building accurate models obtained from data collected at the home gateway. This work is the result of a measurement study conducted at residential clients of a major network provider. We focus on two metrics: the total traffic process generated by clients and the loss process collected from UDP samples generated from clients of a large set to a given reference server. Models based on machine learning techniques were developed from the data collected using an embedded software at the home routers we helped developing. After comparing the models, we conclude that HMMs are accurate for modeling the loss process. We also find that the residential home traffic is bursty at different time scales, and that the autocorrelation slowly decays with increasing time lag. This characteristic has not been observed before for residential traffic. We also conclude that HMMs can be used to accurate represent traffic characteristics.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os serviços disponibilizados pela internet tem se tornado cada vez mais diversificados com o passar dos anos. A medida em que conexões de banda larga residencial se popularizam, uma parte maior da população adquire acesso a serviços como *e-shopping*, voz sobre IP, *streaming* de vídeo sob demanda, *internet banking*, entre outros. A internet já é vista como uma infraestrutura essencial para a sociedade, como água, eletricidade e telefone [1]. Nesse contexto, se torna cada vez mais importante que o serviço de banda larga residencial oferecido seja capaz de garantir um desempenho aceitável.

No entanto, a própria definição de desempenho aceitável para o serviço de internet é desafiadora. Diferentes tipos de aplicação exigem diferentes características da rede representadas por um conjunto de métricas de desempenho, porém a definição das métricas de qualidade de rede e de técnicas de medição também não é simples. Como mostrado em BAUER *et al.* [2], mesmo métricas difundidas entre usuários de banda larga, como velocidade de internet, podem ser definidas e medidas de diversas formas.

Por conta da crescente importância das redes de acesso residencial, entidades reguladoras de telecomunicações de diversas partes do mundo tem voltado sua atenção para a avaliação de desempenho da internet, em um esforço para melhorar a qualidade do serviço prestado aos usuários. Uma discussão ampla sobre métodos de regulação da internet pode ser encontrada em LEHR *et al.* [1].

A *Federal Communications Commission* (FCC), agência reguladora dos Estados Unidos, implementou em 2010 o programa *Measuring Broadband America*, através do qual voluntários instalam em suas residências um roteador que realiza medições em intervalos regulares de diversas métricas de rede, como *throughput*, latência, perda e *jitter*. A metodologia aplicada é aberta e disponibilizada publicamente pela agência reguladora, assim como os resultados das medições são disponibilizados anualmente. LEHR *et al.* [1] define a abordagem implementada pelo FCC como regulação implícita, realizada através das forças de competição do mercado.

No Brasil, a regulação da qualidade da rede de acesso residencial é realizada de maneira explícita. Em 2011 a Agência Nacional de Telecomunicações (ANATEL) criou a resolução número 574, através da qual se define um conjunto de valores mínimos a serem alcançados por métricas de desempenho de rede como latência, *jitter*, perda, *throughput* e disponibilidade. Porém, não se define a metodologia a ser utilizada durante as medições destas métricas. Os provedores de internet que não alcançam as metas propostas são multados pela ANATEL. Assim como no programa americano, um conjunto de voluntários instala em suas residências o roteador responsável por medir as métricas de interesse.

Com o objetivo de garantir o cumprimento das metas da ANATEL, avaliar o desempenho obtido pelo usuário final e melhorar a qualidade do serviço oferecido, a operadora de telecomunicações NET iniciou em 2012 um projeto em parceria com a universidade e com uma empresa incubada no Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia (COPPE) para a realização de medições em redes de acesso residencial. Com este intuito, o módulo de medição de rede do software Tangram-II [3] foi modificado e adaptado para que pudesse ser embarcado em roteadores residenciais com o sistema operacional OpenWrt [4], uma versão do Linux criada especificamente para sistemas embarcados. Um conjunto de colaboradores instalam em suas residências os roteadores com o software embarcado, os quais realizam as medições de maneira periódica. Dentre as métricas coletadas estão *throughput*, *jitter*, latência, perda, disponibilidade e tempo de resolução DNS. O projeto tem como meta distribuir pelo país 4000 roteadores residenciais.

Em 2015 uma extensão deste projeto, financiado pela Financiadora de Estudos e Projetos (FINEP), será iniciada. Nesta nova etapa, além das medições realizadas a partir das redes de acesso residencial serão coletadas métricas de desempenho do núcleo da rede. A combinação dos resultados obtidos será utilizada para a criação de modelos com o objetivo de guiar o processo de planejamento e gerenciamento da rede. Dentre as tarefas propostas estão o planejamento de capacidade, a previsão de crescimento do tráfego agregado e a simulação da dinâmica de roteamento do núcleo da rede.

Um pré-requisito para a implementação destas técnicas é a criação de modelos matemáticos a partir dos dados coletados. Através da utilização de modelos matemáticos é possível adquirir maior entendimento do processo envolvido, inferir métricas de interesse e prever o comportamento futuro da métrica considerada.

Neste trabalho são discutidos modelos para o processo de perda e para o tráfego gerado em redes de acesso residenciais. A abordagem é baseada em técnicas de aprendizado de máquina, utilizadas para estimar os parâmetros associados aos modelos considerados.

Nossas contribuições são resumidas a seguir:

- Participamos da construção de módulos de software que permitem a coleta de dados de tráfego e do processo de perda tratados neste trabalho. Resolvemos problemas práticos de implementação para que a coleta fosse feita de forma confiável e sem onerar os equipamentos (roteador) dos usuários. Um exemplo é a coleta de tráfego do cliente a cada segundo, que gera uma quantidade grande de informação que não pode ficar armazenada no roteador por limitação de memória.

- Implementamos o software para tratamento dos dados de forma a possibilitar a construção dos modelos

- Comparamos e escolhemos os modelos mais precisos para as métricas consideradas

- Identificamos uma característica do tráfego residencial que não havia sido medida anteriormente: o tráfego medido é *bursty* para várias escalas de tempo. Essa característica tem impacto no planejamento de redes de acesso. Ela foi reportada para tráfego agregado de redes de maior porte [5], mas nenhum estudo existe sobre rede residencial.

- Verificamos que rajadas de perda longas (maior que 2 pacotes consecutivos) tem probabilidade de ocorrência não desprezível.

O trabalho é organizado como se segue. No Capítulo 2 são apresentados trabalhos da literatura relacionados a medições residenciais e a modelos de tráfego e perda. O Capítulo 3 descreve os principais conceitos relacionados a este trabalho. O Capítulo 4 descreve o ambiente de medição e o processo de modelagem dos dados. O Capítulo 5 apresenta e avalia os modelos de tráfego em redes de acesso residencial criados neste trabalho. No Capítulo 6 é realizada a avaliação do desempenho de modelos do processo de perda parametrizados a partir de dados coletados dentro do âmbito da parceria em redes de acesso residencial. Por fim, o Capítulo 7 discute as conclusões e os trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

A análise do desempenho de redes residenciais tem sido o foco de trabalhos recentes da literatura [6–8]. Em alguns casos as métricas são coletadas a partir de dispositivos finais, como em KREIBICH *et al.* [7], que analisa os resultados coletados através da ferramenta *Netalyzr*[1], e em CANADI *et al.* [8], que avalia o desempenho de redes de acesso residencial de diversas partes do mundo utilizando o software *speedtest*[2]. Medições coletadas em dispositivos finais podem ser afetadas por diversos fatores externos, como tráfego concorrente, interferência em redes sem fio e sobrecarga no dispositivo final. Em SUNDARESAN *et al.* [6] são analisados os resultados de medições de métricas como latência, perda e vazão coletadas em mais de 4000 *gateways* residenciais dos Estados Unidos utilizando o software desenvolvido pela empresa SamKnows [9]. A abordagem utilizada para a coleta dos dados de [6] é a mesma adotada neste trabalho, onde um roteador residencial conectado diretamente ao modem de acesso a rede coleta as métricas de interesse, evitando a presença de fatores externos de degradação de desempenho. No entanto, [6] tem caráter descritivo e não aprofunda a análise sobre as métricas coletadas. Neste trabalho são propostos modelos criados a partir dos dados coletados e são analisadas características como a rajada de perda e a autocorrelação do tráfego medido.

Em relação a modelos para o processo de perda, YAJNIK *et al.* [10] propõe a modelagem da perda através de um modelo de Markov de ordem k e mostra que é possível encontrar autocorrelações entre perdas para escalas de tempo de até 1 segundo. Em SALAMATIAN e VATON [11] são criados modelos de perda utilizando uma cadeia de Markov oculta que gera 2 símbolos com o objetivo de inferir o estado da rede a partir do modelo. Em SILVEIRA e DE SOUZA E SILVA [12] o processo de perda é modelado utilizando uma cadeia de Markov oculta hierárquica com o objetivo de prever o comportamento futuro e ajustar a redundância de algoritmos de *Forward Error Correction* (FEC). Um modelo de Gilbert simplificado associado

[1] Acessível através da página www.netalyzr.icsi.berkeley.edu
[2] Acessível através da página www.speedtest.net

a cada estado oculto é responsável por gerar os símbolos que representam perda ou sucesso durante a transmissão do pacote. A idéia por trás desta modelagem é que as correlações de curto prazo podem ser capturadas por um modelo simples como a cadeia de Markov de 2 estados, enquanto as correlações de longo prazo são modeladas através da dinâmica associada aos estados ocultos. Em [13] é avaliado o desempenho de modelos de perda presentes na literatura e é proposto um modelo de Markov de 2 níveis para a modelagem da perda observada em redes residenciais. No entanto, o modelo criado em [12] não é avaliado, e o processo de treinamento do modelo de 2 níveis é realizado a partir de parâmetros obtidos de maneira empírica, não sendo proposto nenhum algoritmo para o cálculo destes parâmetros. Neste trabalho analisa-se a acurácia dos modelos propostos em [11, 12], utilizando como métricas a distribuição da rajada de perda e do intervalo entre perdas.

A maior parte dos modelos de tráfego presentes na literatura consideram o tráfego agregado de diversos clientes, coletado a partir de medições realizadas no *backbone* de operadoras ou na infraestrutura de redes empresariais, como nos trabalhos [5, 14]. Neste trabalho são criados modelos de tráfego para redes de acesso residencial, a partir dos quais se torna possível analisar como a dinâmica do tráfego agregado varia de acordo com cada perfil de tráfego medido.

Trabalhos recentes sobre comportamento do tráfego de usuários são em geral descritivos, apenas analisando dados como padrões diários de tráfego ou saturação do canal [15–17]. Por outro lado, trabalhos como [18–20] analisam a dinâmica do tráfego coletado em dispositivos finais ou em redes de acesso residencial. O trabalho de SIMPSON JR. *et al.* [18] propõe a utilização de distribuições empíricas para métricas como o número de bytes enviados e recebidos e o tempo entre requisições do usuário. As distribuições são criadas a partir dos dados coletados em dispositivos finais utilizando a ferramenta NETI@HOME [21]. No entanto, o tráfego é separado de acordo com o protocolo de transporte utilizado e as distribuições empíricas não são mapeadas em distribuições reais. Em KIHL *et al.* [19] são analisadas diversas características do tráfego gerado em uma rede municipal da Suécia, dentre elas o padrão diário de tráfego e a distribuição da taxa de tráfego, sendo proposta a utilização da distribuição de Weibull para modelar a taxa de tráfego. No entanto, é considerado apenas o período ativo do usuário e o tráfego é coletado com granularidade de 5 minutos. Por fim, em AGOSTA *et al.* [20] é proposta a utilização de misturas de distribuições exponencial e pareto para modelar o tráfego medido em dispositivos finais. Além de ter foco em redes empresariais, [20] modela apenas o número de fluxos[3] em um determinado período, enquanto este trabalho foca na quantidade de tráfego gerado.

[3]Neste trabalho um fluxo é definido como um conjunto de pacotes que utilizam o mesmo protocolo de transporte e que possuem os mesmos IP e porta de origem e de destino

Capítulo 3

Conceitos básicos

3.1 Cadeia de Markov oculta

Detalha-se nesta seção o principal modelo baseado em aprendizado de máquina utilizado neste trabalho, a cadeia de Markov oculta. Além da cadeia de Markov oculta discreta em que os símbolos são gerados de acordo com uma probabilidade fixa, é discutido um modelo hierárquico no qual os símbolos são gerados por cadeias de Markov representando o segundo nível de hierarquia. Desta forma, a cada estado oculto é associada uma cadeia de Markov.

Uma cadeia de Markov oculta é um processo duplamente estocástico composto de um processo oculto, representado por uma cadeia de Markov, e um processo observável, determinado de acordo com o estado oculto. Uma boa referência sobre cadeias de Markov ocultas pode ser encontrada em [22, 23].

Apenas cadeias de Markov cujo conjunto de símbolos é discreto serão consideradas neste trabalho. No entanto, é possível a utilização de um conjunto contínuo de símbolos. A notação utilizada para definir o modelo é similar a definida em [22].

Uma cadeia de Markov discreta é composta de N estados ocultos e M símbolos observáveis. Seus estados ocultos são definidos através do conjunto $\mathbf{S} = \{S_0, S_1, ..., S_N\}$ e o conjunto de símbolos observáveis é definido por $\mathbf{V} = \{v_1, v_2, ..., v_M\}$. A sequência de símbolos observados é representada utilizando o vetor $\mathbf{O} = \{O_1, O_2, ..., O_T\}$, enquanto o estado em que a cadeia de Markov se encontra no tempo t é representado por s_t. Utilizando esta notação, podemos definir o vetor π de probabilidades iniciais para os estados ocultos, e as matrizes $\mathbf{A}$ e $\mathbf{B}$ de transição entre estados ocultos e de emissão de símbolos em cada estado oculto:

$$\pi(i) = P[s_1 = S_i] \tag{3.1.1}$$

$$A(i,j) = P[s_t = S_j | s_{t-1} = S_i] \tag{3.1.2}$$

$$B(j,k) = P[O_t = v_k | s_t = S_j] \tag{3.1.3}$$

Uma cadeia de Markov oculta discreta pode ser definida através do número de estados N, do número de símbolos M, do vetor de probabilidades iniciais π, da matriz de transição entre estados $\mathbf{A}$ e da matriz de geração de símbolos $\mathbf{B}$. Desta forma, o modelo pode ser definido através da notação compacta $\lambda = (\pi, \mathbf{A}, \mathbf{B})$.

Após a definição dos parâmetros da cadeia de Markov oculta se torna possível definir um conjunto de variáveis de interesse. Seja o tempo total de observação T, a partir de uma sequência de símbolos observados, representada utilizando o vetor $\mathbf{O} = \{O_1, O_2, ..., O_T\}$, são definidas as seguintes variáveis:

$$\alpha_t(i) = P[O_1, O_2, ..., O_t, s_t = S_i | \lambda] \tag{3.1.4}$$

$$\beta_t(i) = P[O_{t+1}, O_{t+2}, ..., O_T | s_t = S_i, \lambda] \tag{3.1.5}$$

$$\begin{aligned} \gamma_t(i) &= P[s_t = S_i | \mathbf{O}, \lambda] \\ &= \frac{\alpha_t(i).\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i).\beta_t(i)} \end{aligned} \tag{3.1.6}$$

$$\begin{aligned} \varepsilon_t(i,j) &= P[s_t = S_i, s_{t+1} = S_j | \mathbf{O}, \lambda] \\ &= \frac{\alpha_t(i).A(i,j).B(j,O_{t+1}).\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i).A(i,j).B(j,O_{t+1}).\beta_{t+1}(j)} \end{aligned} \tag{3.1.7}$$

A partir destas variáveis se torna possível a obtenção de métricas de interesse, como o *likelihood* e a sequência de estados ocultos mais provável dada uma sequência de observações. Para calcular o *likelihood* de uma sequência de símbolos $\mathbf{O}$, representado por $P(\mathbf{O}|\lambda)$, utiliza-se o algoritmo *Forward-Backward*, implementado utilizando programação dinâmica a partir das variáveis α ou β. A obtenção da sequência de estados mais provável dada uma sequência de símbolos depende diretamente da definição de sequência de estados mais provável. Quando considerada como sequência de estados mais provável aquela em que cada estado é individualmente mais provável define-se que para qualquer t:

$$s_t = \underset{1 \leq i \leq N}{argmax}[\gamma_t(i)], \ 1 \geq t \geq T \tag{3.1.8}$$

A sequência de estados individualmente mais prováveis não necessariamente representa uma sequência possível de ser obtida pelo modelo, pois se torna viável gerar uma sequência em que o estado s_{t-1} não possua transição para o estado s_t. Para considerar apenas sequências de estados válidas a sequência passa a ser definida como o caminho de estados mais provável $\mathbf{s} = \{s_1, s_2, ..., s_T\}$, representado por $P(\mathbf{s}|\mathbf{O}, \lambda)$. Para encontrar esta sequência utiliza-se o algoritmo de Viterbi, mostrado em detalhes em [24].

O processo de estimação dos parâmetros é um passo importante na criação de um modelo. O número de estados N e o número de símbolos M devem ser escolhidos pelo usuário do modelo, sendo em geral definidos de acordo com algum conhecimento prévio sobre o problema estudado ou através de heurísticas. Os parâmetros π, $\mathbf{A}$ e $\mathbf{B}$ são estimados utilizando o algoritmo de Baum-Welch [25], uma variação do algoritmo de *Expectation-Maximization* [26]. Neste algoritmo os parâmetros do modelo são obtidos de maneira iterativa a partir das variáveis α, β, γ e ε. Uma descrição detalhada do algoritmo pode ser encontrada em BAUM *et al.* [25].

O primeiro passo do algoritmo de Baum-Welch é a definição de valores iniciais para os parâmetros do modelo. A seguir os parâmetros são reestimados de maneira iterativa de acordo com as seguintes fórmulas:

$$\pi(i) = \gamma_1(i) \tag{3.1.9}$$

$$A(i,j) = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.1.10}$$

$$B(i,k) = \frac{\sum_{t=1, O_t=v_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{3.1.11}$$

É possível mostrar que cada iteração deste algoritmo aumenta o *likelihood* da sequência de símbolos em relação ao modelo $P(\mathbf{O}|\lambda)$ até que se alcance um máximo local da função de *likelihood* [25]. O máximo local alcançado depende dos valores iniciais escolhidos para os parâmetros do modelo, de forma que diferentes inicializações podem levar a melhores máximos locais da função de *likelihood*.

3.2 Cadeia de Markov oculta hierárquica

Com o objetivo de modelar o processo de perda SILVEIRA e DE SOUZA E SILVA [12] propõe a criação de uma cadeia de Markov oculta hierárquica na qual cada

estado oculto gera símbolos de acordo com uma cadeia de Markov denominada modelo de Gilbert simplificado. O modelo de Gilbert simplificado é uma versão modificada do modelo original de GILBERT [27], sendo definido como uma cadeia de Markov de 2 estados em que cada visita ao estado 0 corresponde a um pacote transmitido com sucesso, enquanto visitas ao estado 1 representam a perda de um pacote.

A idéia por trás do modelo é que as correlações de curto prazo podem ser modeladas utilizando um modelo simples como o modelo simplificado de Gilbert, enquanto as correlações de longo prazo são modeladas pela cadeia de Markov oculta. Neste caso, os estados ocultos representariam diferentes estados de congestionamento da rede.

Este modelo trabalha com o conceito de observações em lote, no qual cada visita a um estado oculto i gera b símbolos antes de realizar uma transição para um estado oculto j. Desta forma, o modelo considera que a sequência de símbolos é segmentada em lotes de tamanho b. O valor de b é um parâmetro de entrada do modelo que deve ser definido pelo usuário.

Por conta das observações em lote cada visita a um estado oculto gera um vetor de símbolos. Desta forma, a sequência de símbolos gerada é redefinida utilizando a notação $\mathbf{O} = \{\mathbf{O_1}, \mathbf{O_2}, ..., \mathbf{O_T}\}$, onde $\mathbf{O_t} = [O_{t,1}, O_{t,2}, ..., O_{t,b}]$.

Também é necessário estabelecer a notação utilizada para definir os parâmetros relacionados a geração de símbolos. A Figura 3.1 exemplifica a notação utilizada para representar o modelo de Gilbert simplificado, cujos parâmetros são definidos para o estado oculto S_i por:

$$r_i = P[O_{t,1} = 1 | s_t = S_i] \tag{3.2.1}$$

$$p_i = P[O_{t,j} = 1 | O_{t,j-1} = 0, s_t = S_i] \tag{3.2.2}$$

$$q_i = P[O_{t,j} = 0 | O_{t,j-1} = 1, s_t = S_i] \tag{3.2.3}$$

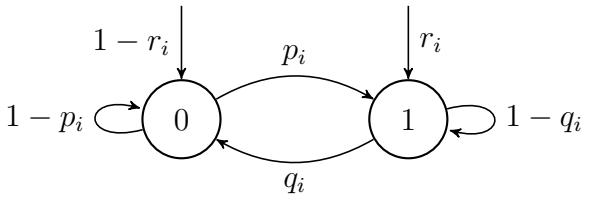


Figura 3.1: Modelo de Gilbert simplificado

Os novos parâmetros podem ser representados utilizando os vetores $\mathbf{r} = \{r_1, r_2, ..., r_N\}$, $\mathbf{p} = \{p_1, p_2, ..., p_N\}$ e $\mathbf{q} = \{q_1, q_2, ..., q_N\}$. O vetor de probabilidades iniciais π e a matriz de transição $\mathbf{A}$ continuam a ser definidos de acordo com as

fórmulas (3.1.1) e (3.1.2). Desta forma, o modelo pode ser definido de acordo com a notação compacta $\lambda = (\pi, \mathbf{A}, \mathbf{p}, \mathbf{q}, \mathbf{r})$.

Para realizar a estimativa dos parâmetros do modelo é realizada uma modificação no método de Baum-Welch [12]. É possível provar que o cálculo da matriz $\mathbf{A}$ e do vetor π continuam de acordo com as fórmulas (3.1.9) e (3.1.10), uma vez que estes não foram modificados em relação as cadeias de Markov ocultas originais. A estimativa dos parâmetros associados ao modelo de Gilbert gerada em cada iteração do algoritmo é dada por:

$$r_i = \frac{\sum_{t=1}^{T} \mathbb{I}(O_{t,1} = 1)\gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{3.2.4}$$

$$p_i = \frac{\sum_{t=1}^{T} G^{01}_{\mathbf{O_t}} \gamma_t(i)}{\sum_{t=1}^{T} (G^{00}_{\mathbf{O_t}} + G^{01}_{\mathbf{O_t}})\gamma_t(i)} \tag{3.2.5}$$

$$q_i = \frac{\sum_{t=1}^{T} G^{10}_{\mathbf{O_t}} \gamma_t(i)}{\sum_{t=1}^{T} (G^{10}_{\mathbf{O_t}} + G^{11}_{\mathbf{O_t}})\gamma_t(i)} \tag{3.2.6}$$

Onde $G^{ij}_{\mathbf{O_t}}$ representa o número de transições entre os estados i e j do modelo de Gilbert simplificado dado o lote de observações $\mathbf{O_t}$.

Uma generalização do modelo hierárquico criado em [12] é proposta no trabalho de DE VIELMOND *et al.* [28]. O novo modelo hierárquico possui dentro de cada estado oculto uma cadeia de Markov geral responsável pela geração de símbolos. A Figura 3.2 exemplifica os parâmetros de uma cadeia de Markov do tipo nascimento e morte associada a um estado oculto i.

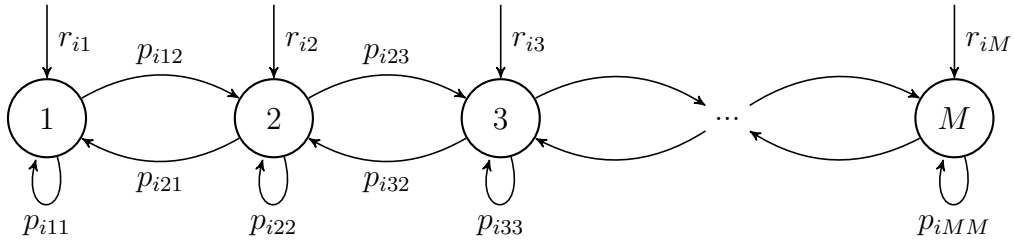


Figura 3.2: Cadeia de Markov dentro de um estado oculto i

Como mostrado na Figura 3.2, é associado a cada estado oculto um vetor de probabilidades iniciais $\mathbf{r_i}$ e uma matriz de transição entre estados $\mathbf{p_i}$ definidos por:

$$r_{ik} = P[O_{t,1} = k | s_t = S_i] \tag{3.2.7}$$

$$p_{ilm} = P[O_{t,j} = m | O_{t,j-1} = l, s_t = S_i] \tag{3.2.8}$$

Nenhuma modificação é realizada no vetor de probabilidades π e na matriz de transição entre estados ocultos $\mathbf{A}$, de forma que a estimativa destes parâmetros continua sendo dada pelas fórmulas 3.1.9 e 3.1.10. A estimativa dos parâmetros associados a geração dos símbolos é dada pela generalização para um conjunto de estados M das fórmulas 3.2.4, 3.2.5 e 3.2.6, resultando em:

$$r_{ik} = \frac{\sum_{t=1}^{T} \mathbb{I}(O_{t,1} = k)\gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{3.2.9}$$

$$p_{ilm} = \frac{\sum_{t=1}^{T} G^{lm}_{\mathbf{O_t}} \gamma_t(i)}{\sum_{j=1}^{M} \sum_{t=1}^{T} (G^{lj}_{\mathbf{O_t}})\gamma_t(i)} \tag{3.2.10}$$

Onde $G^{ij}_{\mathbf{O_t}}$ representa o número de transições entre os estados i e j da cadeia de Markov dado o lote de observações $\mathbf{O_t}$.

3.3 Misturas de distribuições

Uma mistura de distribuições é um modelo matemático criado a partir de uma combinação convexa de distribuições básicas. Sua utilização torna possível a aproximação de distribuições complexas de maneira bastante precisa. Misturas são frequentemente aplicadas em processos de clusterização de dados, sendo também utilizadas no contexto de modelagem de distribuições. Uma explicação detalhada sobre misturas de distribuições pode ser encontrada em [29, 30]. A notação utilizada nesta seção é baseada em [29].

Consideremos um conjunto de K densidades de probabilidades p_k que seguem uma mesma distribuição mas possuem diferentes valores associados a seus parâmetros. A mistura de distribuições é obtida a partir da combinação linear das densidades p_k. Nesta combinação associa-se a cada densidade um parâmetro π_k, representando a probabilidade *a priori* da escolha de p_k. Cada densidade da mistura é denominada componente, enquanto cada probabilidade π_k representa o peso associado a componente k. A Figura 3.3 exemplifica este tipo de modelo através

de uma mistura de gaussianas de 3 componentes. A mistura de K distribuições com densidades p_k é definida através da fórmula:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k . p_k(\mathbf{x}) \tag{3.3.1}$$

Onde $\sum_{k=1}^{K} \pi_k = 1$ e $0 \leq \pi_k \leq 1$.

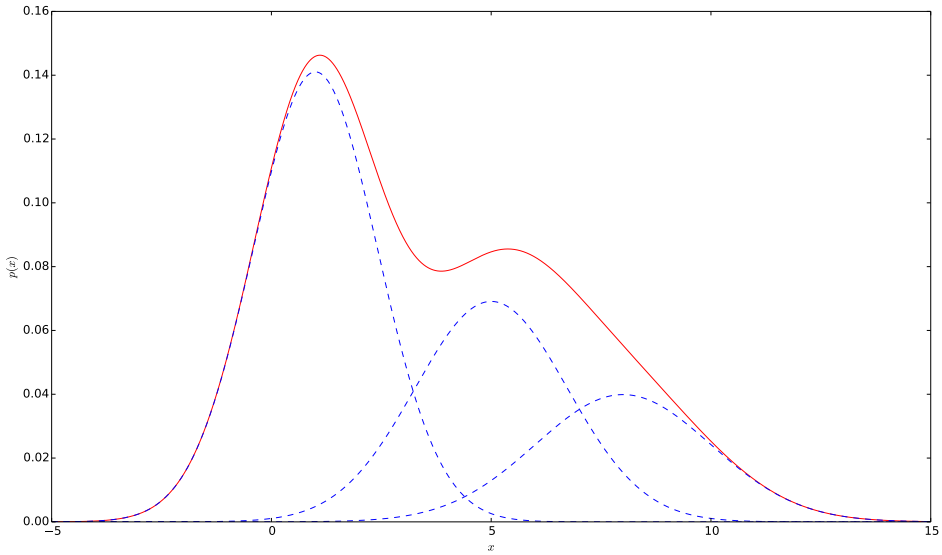


Figura 3.3: Exemplo de mistura de gaussianas de 3 componentes

O treinamento de misturas de distribuições é realizado através do algoritmo de *Expectation-Maximization* [26]. Para a aplicação deste algoritmo é necessária a definição do número de componentes da mistura e de valores iniciais para os parâmetros de cada componente. Prova-se que a cada iteração o algoritmo de EM aumenta o valor da função de *log-likelihood* até encontrar um máximo local desta função. Diferentes valores iniciais para os parâmetros do modelo levam o algoritmo de EM a alcançar diferentes máximos locais, sendo a inicialização uma parte importante da aplicação deste algoritmo.

3.4 Distância de Jensen-Shannon

Muitos conceitos relacionados a teoria da informação podem ser aplicados no contexto de aprendizado de máquina. A partir de métricas como entropia e informação mútua é possível a implementação de técnicas de seleção de modelos e escolha de

variáveis de entrada. Nesta seção é detalhada a distância de Jensen-Shannon, uma métrica utilizada durante o processo de escolha do modelo que melhor representa o conjunto de dados observado. A notação utilizada será similar a encontrada em [30].

O primeiro passo para a obtenção da distância de Jensen-Shannon é a definição de entropia. No contexto de teoria da informação, a entropia representa a quantidade média de informação transmitida durante o envio do valor de uma variável aleatória. Esta métrica também pode ser interpretada como o grau de incerteza associado a uma variável aleatória. Seja uma variável aleatória discreta X que pode assumir K valores distintos, sua entropia $H(X)$ é definida por:

$$H(X) = -\sum_{k=1}^{K} p(X = k)\ log\ p(X = k) \tag{3.4.1}$$

Em geral o logaritmo associado a definição de entropia possui base 2, sendo então a entropia medida em bits. Logaritmos naturais também podem ser utilizados, caso em que a unidade de medida é denominada *nats*. É possível considerar a entropia de distribuições contínuas a partir da generalização da fórmula (3.4.1) [29].

A partir da definição de entropia é possível definir a divergência de Kullback–Leibler (*KL divergence*), também conhecida como entropia relativa. Esta métrica pode ser interpretada como o número extra de bits necessários para se codificar um conjunto de dados que segue uma distribuição p utilizando uma aproximação dada por uma distribuição q. A divergência de Kullback-Leibler também é utilizada para medir a distância entre uma distribuição de interesse p e um modelo cuja distribuição é dada por q. Esta métrica é definida como:

$$\begin{aligned} KL(p||q) &= -\sum_{k=1}^{K} p(X = k)\ log\ \frac{p(X = k)}{q(X = k)} \\ &= -H(p) + H(p, q) \end{aligned} \tag{3.4.2}$$

Onde $H(p, q)$ é denominada entropia cruzada, sendo sua definição dada por:

$$H(p, q) = -\sum_{k=1}^{K} p(X = k)\ log\ q(X = k) \tag{3.4.3}$$

Quanto mais próximo de 0 é o valor da divergência de Kullback-Leibler mais próximas são as distribuições p e q. Esta propriedade torna possível a utilização desta métrica como um critério para a escolha de modelos criados a partir da dis-

tribuição de um conjunto de dados. No entanto, a divergência de Kullback-Leibler não é definida quando $q(X = k_i) = 0$ e $p(X = k_i) \neq 0$.

Outra métrica aplicada no processo de seleção de modelos é a distância de Jensen-Shannon [31], uma versão suavizada e simétrica da divergência de Kullback-Leibler. Para definir esta métrica é necessária a criação de uma distribuição m definida por:

$$m(X = k) = \frac{1}{2}(p(X = k) + q(X = k)) \tag{3.4.4}$$

Utilizando esta distribuição podemos definir a distância de Jensen-Shannon através da fórmula:

$$JSD(p||q) = \frac{1}{2}KL(p||m) + \frac{1}{2}KL(q||m) \tag{3.4.5}$$

Valores mais próximos de 0 para a distância de Jensen-Shannon indicam que as distribuições p e q são mais próximas entre si, mesmo comportamento observado a partir da divergência de Kullback-Leibler. No entanto, a distância de Jensen-Shannon é simétrica e definida para qualquer valor de x, sendo preferível a sua utilização para o processo de comparação entre modelos.

Capítulo 4

Metodologia geral

4.1 Ambiente e metodologia de medição

Nesta seção é detalhado o ambiente de medição utilizado para a coleta das métricas de desempenho de rede analisadas neste trabalho. São discutidas a arquitetura utilizada para a coleta dos resultados e a metodologia aplicada para a obtenção do processo de perda e do tráfego em redes de acesso residencial.

4.1.1 Ambiente de medição

A Figura 4.1 mostra a arquitetura do ambiente de medição utilizado para a obtenção das métricas de desempenho de redes de acesso residenciais. As medições são realizadas a partir de um roteador conectado diretamente ao modem de acesso residencial. Esta abordagem difere da encontrada em diversos trabalhos na literatura, nos quais as medições são realizadas a partir de dispositivos finais [7, 8, 20]. A utilização de dispositivos finais tem como principal desvantagem a degradação de performance gerada por fatores externos, como a utilização de redes sem fio, o tráfego concorrente e a carga no dispositivo, enquanto medições realizadas a partir do gateway residencial diminuem o impacto destes fatores [6]. No entanto, roteadores residenciais possuem baixa memória e capacidade de processamento, o que implica em desafios na implementação do *software* de medição a ser utilizado.

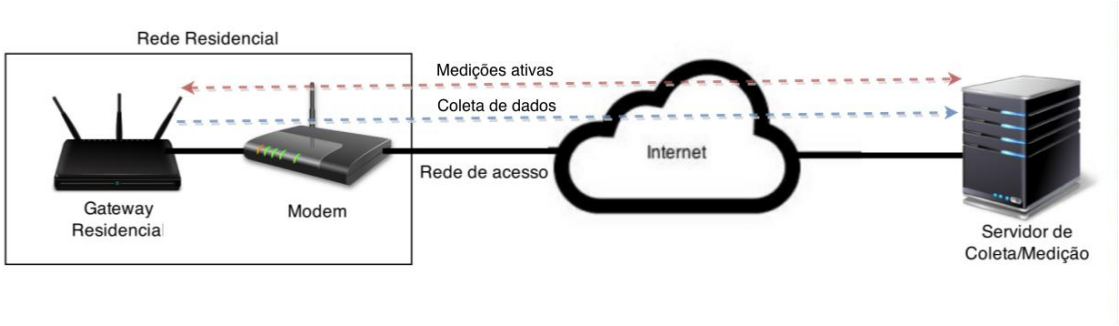


Figura 4.1: Ambiente de medição utilizado para coleta dos dados

Durante este trabalho participamos do processo de implementação de um *software* criado para obter métricas de rede de maneira ativa e passiva. Este programa foi embarcado em roteadores com o sistema operacional OpenWrt [4], uma versão do Linux adaptada para sistemas embarcados. A implementação é baseada no módulo de medições presente na ferramenta Tangram-II [3].

A baixa capacidade da memória de armazenamento do roteador representa um desafio de implementação, inviabilizando a execução do *software* de medição a partir deste tipo de memória. Em nossa implementação a ferramenta de medição é armazenada e executada a partir da memória volátil, representada no sistema operacional pelo diretório */tmp*. A memória de armazenamento contém apenas um conjunto de *scripts* responsáveis pela atualização remota do *software* presente na memória volátil e pela implementação de configurações definidas remotamente.

Por conta da baixa memória encontrada em *gateways* residenciais os resultados das medições precisam ser enviados periodicamente a um servidor de coleta. Durante a execução de medições ativas o servidor não apenas recebe os resultados obtidos como também participa do processo de medição. O envio dos resultados das medições ativas e passivas é realizado de maneira segura através do protocolo SSL.

Diversas métricas são coletadas em nossas medições, como latência, *jitter* e *throughput*. Neste trabalho são analisados apenas o processo de perda e o tráfego observados em redes de acesso residencial.

4.1.2 Tráfego residencial

Um parâmetro importante a ser considerado em medições de tráfego é a granularidade com a qual as amostras são coletadas. Idealmente a coleta do tráfego seria realizada a nível de pacote, uma vez que a partir desta abordagem se torna possível a obtenção da informação de interesse com o máximo de precisão. No entanto, a implementação de medições a nível de pacote não é viável em *gateways* residenciais por conta de sua baixa capacidade de memória e processamento. A maior parte dos trabalhos de medição de tráfego presentes na literatura coletam seus resultados

a partir de roteadores presentes no núcleo da rede, os quais informam o tráfego agregado com granularidade de alguns minutos [14, 19]. Outros trabalhos coletam o tráfego com baixa granularidade, mas suas medições não são obtidas a partir de redes residenciais [5, 20]. As medições deste trabalho coletam o tráfego gerado em redes de acesso residencial com a granularidade de 1 segundo. Não é de nosso conhecimento a existência de outro trabalho de medição de tráfego residencial com granularidade baixa.

As medições foram realizadas a partir de roteadores instalados de forma a garantir que todo o tráfego residencial fosse observado. Para medir o tráfego foram amostrados em intervalos de aproximadamente 1 segundo o número total de bytes trafegados pela rede de acesso, ou seja, o tráfego roteado através da interface *WAN* do roteador. Esta informação é obtida através do arquivo */proc/net/dev* disponibilizado pelo sistema operacional do roteador.

Por conta da baixa capacidade de armazenamento dos roteadores o tráfego medido é automaticamente enviado a um servidor a cada hora, e a cada envio as medições são interrompidas por alguns segundos para evitar a interferência destas transmissões nas estatísticas coletadas. Tanto o tráfego *upstream* quanto o tráfego *downstream* foram coletados, mas neste trabalho considera-se apenas o tráfego *downstream*, uma vez que usuários em geral utilizam pouco sua capacidade de *upload*.

4.1.3 Perda bidirecional

O processo de perda em redes é um assunto amplamente estudado na literatura [6, 10–12]. Dentre as diversas técnicas de medição de perda bidirecional a mais utilizada consiste no envio de pacotes ICMP com intervalos de 1 segundo implementado através da ferramenta *ping*. A utilização de um intervalo entre pacotes alto não permite a análise de correlações de curto prazo, o que impede a coleta precisa de métricas como o tamanho da rajada de perda. Outra técnica utilizada consiste no envio de pacotes cujos intervalos são aleatórios e seguem uma distribuição de Poisson. A idéia é que, por conta da propriedade *Poisson Arrivals See Time Averages* (PASTA) [32] associada a distribuição de Poisson, a utilização destes intervalos torna possível a obtenção de uma média no tempo da métrica de interesse sem viés. No entanto, a aplicação desta técnica em ambientes reais gera estimativas imprecisas em muitos casos [33]. Além disso, a partir de seus resultados não é possível obter informações sobre o comportamento em rajada da perda de pacotes.

Neste trabalho a perda bidirecional é coletada a partir do envio de um trem de 30 pacotes UDP com intervalo entre pacotes de 1 milissegundo e com *payload* de 32 bytes. A partir de um trem de pacotes é possível a obtenção de estatísticas relacionadas ao comportamento do processo de perda em um curto prazo, como o tamanho

da rajada de perda e o intervalo entre perdas. O número de pacotes foi determinado de maneira empírica com o objetivo de gerar medições de curta duração e baixo impacto, uma vez que as medições são ativas e realizadas em residências de voluntários. O intervalo entre pacotes e o tamanho do *payload* foram determinados com o objetivo de não gerar perdas artificiais no buffer do modem residencial. Pacotes que chegam após 3 segundos do início de uma medição são descartados e considerados perdidos. As medições são programadas para ocorrer a cada 30 minutos.

4.2 Processo de modelagem dos dados

Uma mesma metodologia foi adotada para a modelagem do processo de perda e do tŕafego residencial. O processo de modelagem adotado é dividido em 3 partes: pré-processamento dos dados, treinamento dos modelos e avaliação dos modelos. Nesta seção é descrito o funcionamento geral de cada etapa do processo de modelagem aplicado neste trabalho.

4.2.1 Pré-processamento

O primeiro passo necessário para a modelagem de métricas de interesse é o pré-processamento dos dados coletados. É durante esta etapa que os dados são validados e formatados para que possam ser utilizados como entrada de algoritmos responsáveis pela parametrização dos modelos. O processo de validação é especialmente importante quando se utilizam medições reais, uma vez que neste contexto diversos fatores externos podem causar inconsistências e resultados não previstos podem ser observados.

O processo de formatação dos dados inclui não só a transformação dos resultados das medições para um formato compatível com o modelo a ser utilizado como a definição do método de discretização das métricas utilizadas como entrada de modelos discretos. A formatação correta dos dados é essencial para a obtenção de bons modelos, evitando a geração de ruídos que dificultem ou impossibilitem a escolha de parâmetros ótimos. No entanto, cada métrica modelada exige um tipo específico de formatação que deve ser escolhida a partir do funcionamento do modelo utilizado e do conhecimento específico sobre o problema estudado.

Também está inclusa na etapa de pré-processamento a escolha das entradas que serão passadas para o modelo. É importante que apenas entradas relacionadas com as métricas de interesse obtidas a partir do modelo sejam utilizadas, evitando a presença de ruído que diminua a precisão dos algoritmos de escolhas de parâmetros. Um bom conhecimento sobre o problema estudado ajuda na escolha de entradas corretas, sendo também possível a utilização de heurísticas.

4.2.2 Treinamento de modelos

A etapa seguinte ao pré-processamento das entradas consiste no treinamento dos modelos considerados. É nesta etapa que são escolhidos os tipos de modelos que serão utilizados e os parâmetros associados ao processo de treinamento. A escolha de parâmetros é desafiadora e tem grande impacto no resultado final, uma vez que modelos baseados em aprendizado de máquina dependem de boas inicializações para a obtenção de resultados precisos. Estes valores são determinados a partir de conhecimento do problema e da utilização de heurísticas.

A escolha dos tipos de modelos utilizados neste trabalho foi baseada na complexidade de cada modelo e na métrica de interesse extraída a partir do resultado final. Modelos mais simples são mais fáceis de parametrizar, mas podem não ser capazes de capturar todas as informações de interesse. Modelos mais complexos, por outro lado, contém mais informação mas possuem processo de treinamento mais desafiador.

Esta abordagem é exemplificada a partir dos modelos escolhidos para capturar o tráfego residencial. O *fitting* de distribuição é o modelo mais simples considerado para a captura do tráfego. O algoritmo de *Maximum Likelihood Estimation*, utilizado para parametrizar distribuições, não necessita da determinação de valores iniciais para os parâmetros estimados, o que torna o processo de treinamento mais simples. No entanto, as amostras de uma distribuição são independentes, não sendo possível capturar a correlação presente nos dados de entrada. Além disso, o *fitting* de uma única distribuição pode não ser capaz de modelar de maneira precisa a distribuição observada a partir dos dados.

Em seguida são considerados modelos baseados em misturas de distribuições. Neste tipo de modelo é realizada a combinação de um conjunto de distribuições básicas, possibilitando a captura de distribuições mais complexas. O processo de estimativa de parâmetros é realizado pelo algoritmo de *Expectation-Maximization* [26]. Para a aplicação deste algoritmo é necessário determinar o número de componentes da mistura e os valores iniciais para os parâmetros a serem estimados, sendo o resultado final dependente dos valores escolhidos. Apesar da capacidade de modelar de maneira mais precisa distribuições mais complexas, misturas de distribuições também não consideram a correlação entre as amostras coletadas.

Por fim, são considerados os modelos de Markov ocultos. O treinamento deste tipo de modelo é mais desafiador, uma vez que parâmetros como o número de estados, o número de símbolos, a estrutura do modelo e as probabilidades iniciais devem ser determinados antes da aplicação do algoritmo de Baum-Welch [25]. A definição destes parâmetros depende do conhecimento do problema e da experiência do usuário, sendo o resultado final dependente da escolha de valores apropriados. Apesar da

maior complexidade durante o processo de treinamento, cadeias de Markov ocultas contém mais informação pois são capazes de capturar a autocorrelação presente nas variáveis de entrada.

4.2.3 Avaliação dos modelos

A última etapa da modelagem consiste na avaliação dos modelos gerados em relação a um conjunto de métricas de interesse. São comparados nesta etapa diversos tipos de modelos e parâmetros escolhidos durante o processo de treinamento. Para a realização desta avaliação é necessária a extração das métricas de interesse a partir dos modelos, que podem ser obtidas a partir da geração de amostras sintéticas ou a partir da estrutura do modelo. É também nesta etapa que são analisadas possíveis interpretações para as estruturas dos modelos gerados a partir de métodos de aprendizado de máquina.

Capítulo 5

Modelos de tráfego residencial

Entender a dinâmica do tráfego gerado em uma rede de telecomunicações é essencial para que um provedor de internet seja capaz de garantir a qualidade do serviço prestado. Um modelo de tráfego pode ser utilizado em diversos cenários, como na identificação de anomalias, na criação de perfis de usuários e no planejamento de capacidade da rede. Diversos trabalhos na literatura já propuseram modelos para o tráfego agregado em redes empresariais ou no núcleo da rede [5, 14]. No entanto, poucos trabalhos lidaram com modelos para o tráfego gerado por redes de acesso residenciais[18, 19]. Neste capítulo são comparadas diversas técnicas de modelagem para o tráfego residencial baseadas em aprendizado de máquina.

5.1 Dataset

A coleta do tráfego residencial com baixa granularidade pode gerar algum tipo de questionamento pelo usuário monitorado em relação a privacidade. Isso porque a partir destes dados é possível conhecer padrões de utilização da rede e inferir com alguma precisão o tipo de aplicação utilizada, dependendo da granularidade das observações e do tipo de aplicação. Por conta disso, as medições foram realizadas em residências de 9 voluntários que permitiram a coleta do número de bytes trafegados a cada segundo. Apesar do baixo número de usuários, a aplicação de técnicas de aprendizado de máquina torna os modelos flexíveis, sendo facilmente aplicáveis a outros padrões de tráfego.

O período de coleta de dados de cada usuário varia entre 6 semanas e 4 meses, com as primeiras medições iniciadas em outubro de 2014 e as últimas finalizadas em fevereiro de 2015. Dentre os voluntários existem aqueles com redes de acesso via cabo e via VDSL, e a capacidade nominal, que representa o valor contratado pelo usuário, varia entre 10 e 35 Mbps. A Tabela 5.1 mostra a velocidade nominal e o throughput máximo medido para cada um dos voluntários. Nota-se pelas medições que a maior parte dos usuários atinge um throughput aproximadamente 10% maior

que sua velocidade nominal, o que indica que este parâmetro não é um bom indicador para a capacidade real da rede. Por outro lado, um dos voluntários não alcançou em nenhuma medição a capacidade nominal de sua rede.

Tabela 5.1: Capacidade nominal e Throughput máximo medido para cada voluntário

	Capacidade Nominal	Throughput máximo medido
Voluntário 1	20 Mbps	22.0 Mbps
Voluntário 2	15 Mbps	15.7 Mbps
Voluntário 3	10 Mbps	11.1 Mbps
Voluntário 4	10 Mbps	11.1 Mbps
Voluntário 5	30 Mbps	32.8 Mbps
Voluntário 6	10 Mbps	11.1 Mbps
Voluntário 7	35 Mbps	26.3 Mbps
Voluntário 8	35 Mbps	36.9 Mbps
Voluntário 9	10 Mbps	11.1 Mbps

Os desafios encontrados para a criação dos modelos incluem a presença de tráfego em rajadas e a baixa utilização da banda disponível nas redes de acesso. A Figura 5.1 mostra a CDF empírica do tráfego entre 0 e 100 kbps gerado por 3 usuários diferentes. Percebe-se pelo gráfico que ao menos 60% das medições de tráfego de um usuário são menores que 20 kbps e que este valor chega a mais de 90% das medições para um dos clientes.

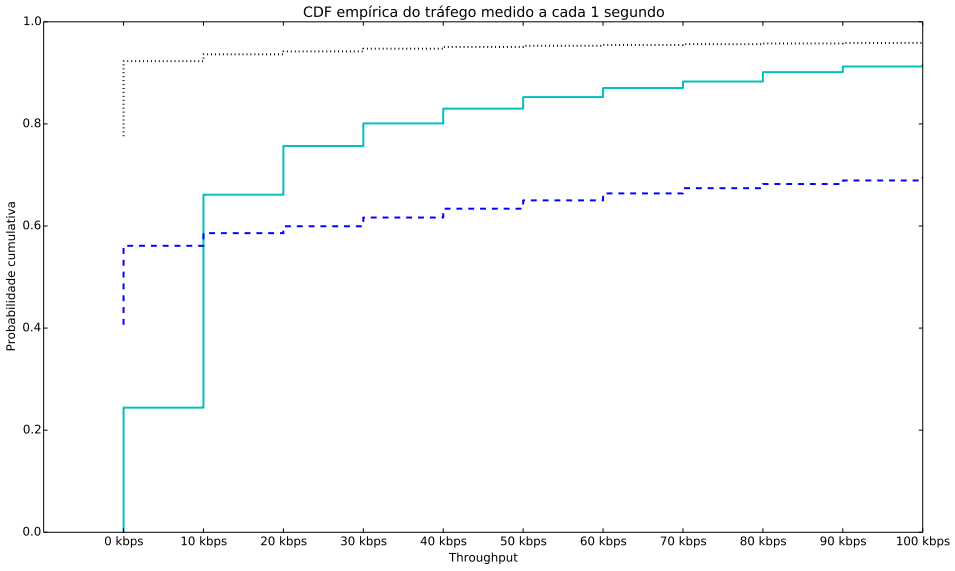


Figura 5.1: CDF empírica do tráfego de 3 voluntários medido a cada segundo

Outra característica das medições coletadas é a presença de rajadas de tráfego em diferentes escalas de tempo. A Figura 5.2 mostra o tráfego coletado na rede de um

voluntário durante 1 semana agregado em escalas de tempo variando entre 1 segundo e 30 minutos. Essa característica é semelhante a reportada em [5] para um tráfego agregado na rede empresarial da Bellcore. Enfatizamos que a rede empresarial de [5] tem características bastante diferentes das observadas em uma rede residencial. Além disso, a medição reportada em [5] foi realizada há 23 anos atrás e, portanto, as aplicações utilizadas atualmente são bastante distintas. É então surpreeendente que uma característica semelhante seja encontrada. Do nosso conhecimento nenhum trabalho reportou tal fato. Um modelo preciso de tráfego residencial deve levar em consideração estas características.

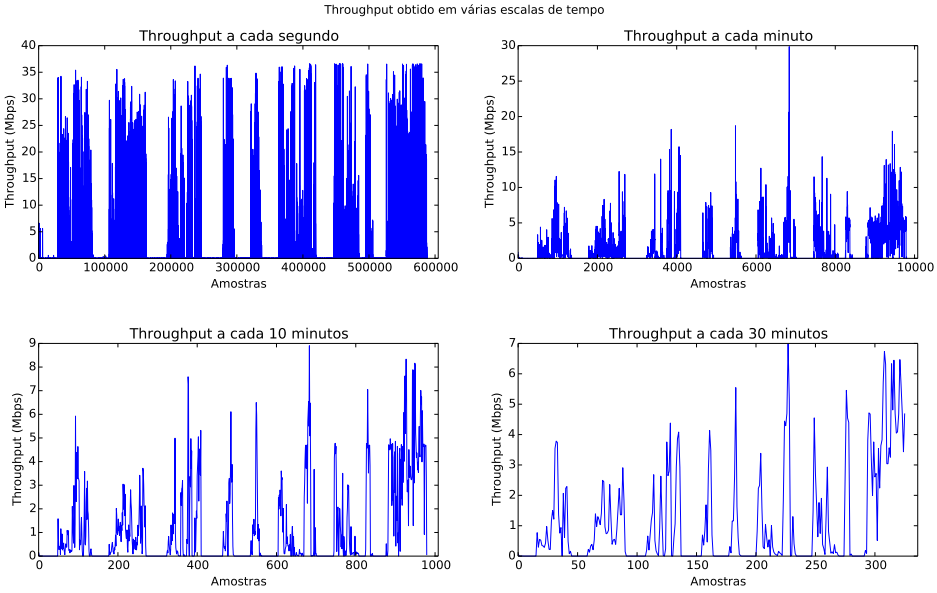


Figura 5.2: Tráfego medido agregado em diferentes escalas de tempo

A presença de rajadas em diferentes escalas de tempo pode ser observada pela autocorrelação do tráfego medido. As Figuras 5.3 e 5.4 mostram a autocorrelação obtida a partir das medições de dois usuários. Observa-se que a autocorrelação tem decaimento lento a medida que o *lag* considerado aumenta. Este fato também está de acordo com [5], segundo o qual uma autocorrelação de decaimento lento sugere uma dependência de longo prazo. Percebe-se pela Figura 5.4 que a autocorrelação pode apresentar um comportamento periódico para certos clientes, o que pode ser explicado pela utilização de algoritmos de compressão de vídeo [34].

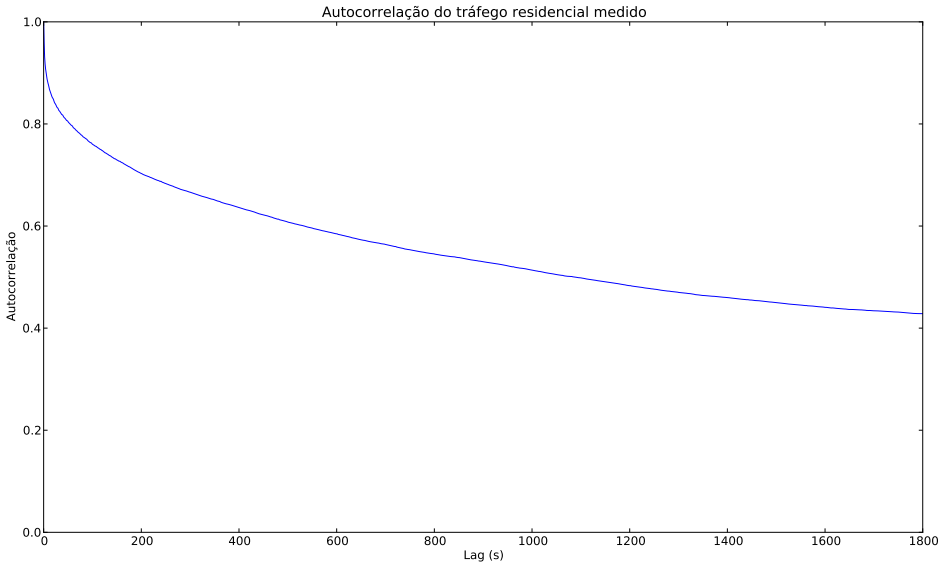


Figura 5.3: Autocorrelação do tráfego residencial gerado pelo voluntário 6

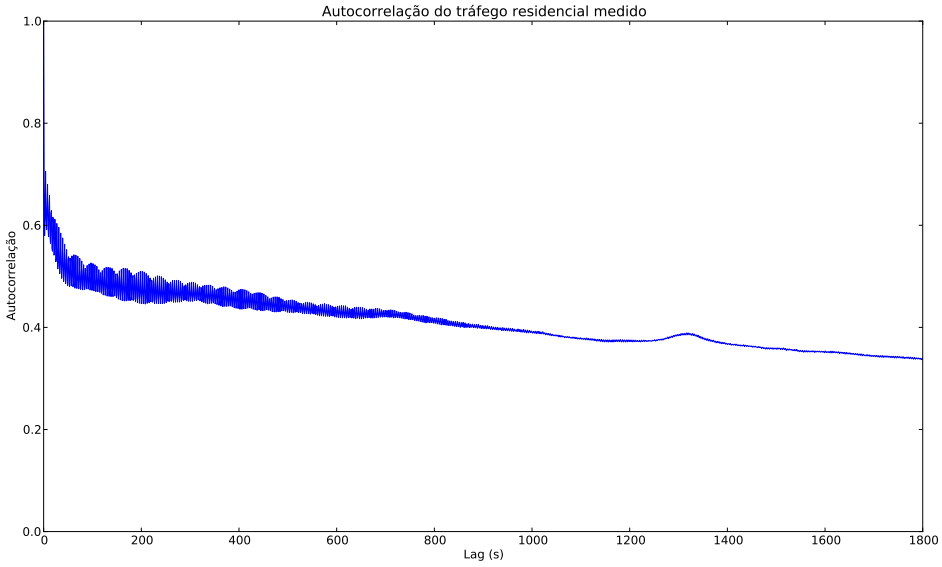


Figura 5.4: Autocorrelação do tráfego residencial gerado pelo voluntário 9

Para entender os diferentes perfis de tráfego de nossos voluntários foi realizada uma clusterização baseada na fração de medições cujo tráfego é baixo (abaixo de 1 Mbps), médio (entre 1 e 6 Mbps) e alto (acima de 6 Mbps) para cada cliente. A clusterização foi implementada utilizando o algoritmo de *k-means* [35] com 3 *clusters*.

Inicialmente foram consideradas para a clusterização apenas amostras com tráfego acima de 70 kbps. O objetivo desta filtragem é eliminar períodos de inatividade e o tráfego irrisório, uma vez que este valor representa menos de 1% da velocidade

nominal dos clientes de 10 Mbps. A Tabela 5.2 mostra os resultados obtidos. Percebe-se a formação de um *cluster* com apenas um cliente (voluntário 5), cujo perfil de tráfego difere dos outros usuários: quase 80% de suas medições são associadas a tráfego baixo. O tráfego abaixo de 1 Mbps inclui o acesso a páginas web e o *streaming* de áudio. O voluntário 5 confirmou que boa parte de seu tráfego corresponde a utilização de aplicações de *streaming* de áudio. Outro agrupamento reúne 3 voluntários que possuem grande parte de suas medições associadas a tráfego alto. O tráfego acima de 6 Mbps pode ser associado a transferência de arquivos grandes, quando o *throughput* máximo disponível tende a ser alcançado, ou ao tráfego gerado por *streaming* de vídeo. Os voluntários associados a este *cluster* confirmaram que boa parte de seu tráfego corresponde a utilização de aplicações de *streaming* de vídeo.

Tabela 5.2: *Clusters* obtidos a partir do tráfego acima de 70 Kbps

	Fração de medições por faixas de tráfego			
	Tráfego baixo	Tráfego médio	Tráfego alto	Voluntários
Centróide 1	0,270	0,270	0,460	2, 8, 9
Centróide 2	0,637	0,279	0,084	1, 3, 4, 6, 7
Centróide 3	0,793	0,153	0,054	5

Em seguida foi realizada a clusterização considerando todas as amostras do tráfego coletado com o intuito de entender a fração de tempo total em que o tráfego se encontra em cada faixa de *throughput*. Os resultados obtidos são mostrados na Tabela 5.3. Mais uma vez os voluntários 2, 8 e 9 apresentaram um perfil diferente dos demais. No entanto, esta clusterização mostra que o voluntário 8 gera tráfego médio e alto por uma fração de tempo maior do que a observada a partir das medições dos outros voluntários.

Tabela 5.3: *Clusters* obtidos a partir do tráfego total medido (Mbps)

	Fração de medições por faixas de tráfego			
	Tráfego baixo	Tráfego médio	Tráfego alto	Voluntários
Centróide 1	0,770	0,111	0,119	8
Centróide 2	0,917	0,031	0,052	2, 9
Centróide 3	0,976	0,018	0,006	1, 3, 4, 5, 6, 7

Para entender o perfil de utilização por horário foi realizada a clusterização da fração do tráfego total gerado pelos usuários em 4 faixas de tempo correspondentes aos períodos de manhã (06:00 às 11:59), tarde (12:00 às 17:59), noite (18:00 às 23:59) e madrugada (00:00 às 5:59). Foram gerados 4 *clusters* a partir da aplicação do algoritmo de *k-means*. Os resultados são mostrados na Tabela 5.4. 4 tipos de perfis de utilização por horário são obtidos. No *cluster* 1 o tráfego é distribuído entre

manhã, tarde e noite. O *cluster* 2 apresenta a maior parte de seu tráfego concentrada no período da noite. O tráfego associado ao *cluster* 3 é gerado majoritariamente nos períodos da tarde e da noite. Por fim, o *cluster* 4 distribui o tráfego entre tarde, noite e madrugada. A Figura 5.5 exemplifica o tráfego diário médio gerado a cada hora para voluntários associados aos clusters 1 e 4.

Tabela 5.4: *Clusters* obtidos a partir da fração do tráfego para cada período do dia (Mbps)

	Fração do tráfego total por horário				
	Madrugada	Manhã	Tarde	Noite	Voluntários
Centróide 1	0,034	0,252	0,396	0,318	5,8
Centróide 2	0,090	0,045	0,192	0,673	7
Centróide 3	0,103	0,144	0,273	0,480	2,9
Centróide 4	0,264	0,094	0,260	0,382	1,3,4,6

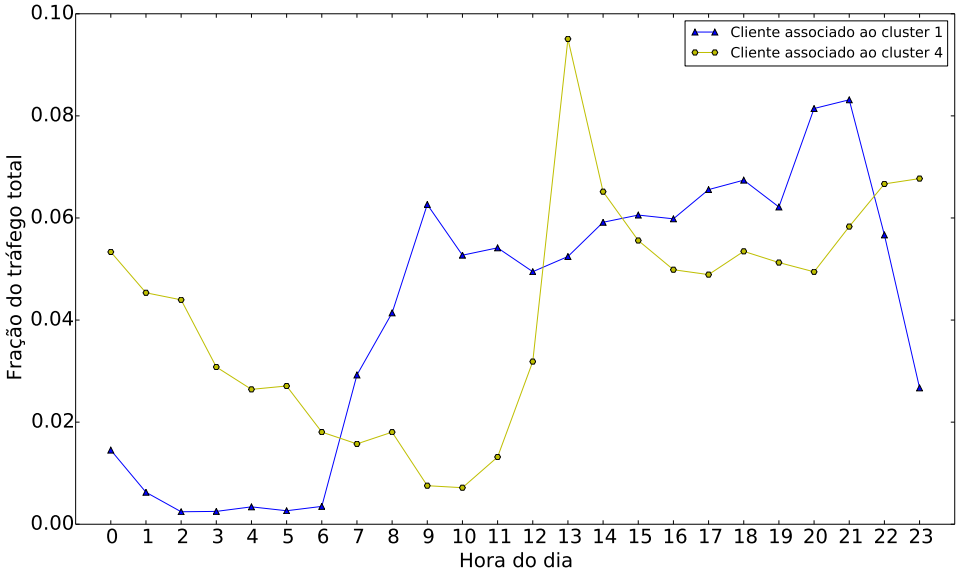


Figura 5.5: Fração do tráfego por hora para voluntários 4 e 8

Por fim, realiza-se para cada uma das 4 faixas de horário um agrupamento da fração de medições em cada um dos 3 níveis de tráfego considerados. Este tipo de agrupamento ajuda a identificar se um determinado conjunto de clientes possui o mesmo perfil de tráfego em diferentes horários do dia. Para esta clusterização são consideradas apenas amostras de tráfego maior que 70 kbps. As clusterizações obtidas são mostradas na Tabela 5.5. Os resultados mostram que os voluntários 2 e 8 são agrupados no mesmo *cluster* em todas as faixas do dia, sendo seu comportamento durante o dia bastante similar. O mesmo fenômeno é observado para as duplas de clientes 6 e 7 e 1 e 3. Percebe-se também que o perfil de tráfego apresenta variações

em diferentes faixas do dia, o que pode ser interpretado como a variação do tipo de aplicação utilizada pelos voluntários em cada hora do dia.

Tabela 5.5: *Clusters* obtidos a partir do tráfego medido em diferentes faixas de horário

		Fração de medições por faixas de tráfego			
		Tráfego baixo	Tráfego médio	Tráfego alto	Voluntários
Madrugada	Centróide 1	0.305	0.335	0.360	2, 8, 9
	Centróide 2	0.602	0.205	0.193	1, 3, 4, 6, 7
	Centróide 3	0.931	0.055	0.014	5
Manhã	Centróide 1	0.196	0.475	0.329	9
	Centróide 2	0.357	0.292	0.351	2, 8
	Centróide 3	0.791	0.125	0.084	1, 3, 4, 5, 6, 7
Tarde	Centróide 1	0.248	0.334	0.418	2, 8, 9
	Centróide 2	0.552	0.240	0.208	4, 6, 7
	Centróide 3	0.665	0.241	0.094	1, 3, 5
Noite	Centróide 1	0.237	0.379	0.384	2, 8, 9
	Centróide 2	0.457	0.249	0.294	6, 7
	Centróide 3	0.607	0.276	0.117	1, 3, 4, 5

5.2 Metodologia

Três tipos de modelos foram utilizados para realizar a modelagem do tráfego coletado: uma cadeia de Markov oculta, uma cadeia de Markov oculta hierárquica que se utiliza dos resultados de [12, 28] e o *fitting* de distribuições mistas. Nas subseções seguintes são considerados separadamente o procedimento utilizado para a modelagem das cadeias ocultas e a estimativa dos parâmetros das distribuições mistas.

5.2.1 Distribuições mistas

Como é mostrado na Figura 5.1 mais de 70% das amostras de tráfego de cada cliente apresentaram um valor muito baixo para o número de bytes transferidos, ou seja, a rede é pouco utilizada pelos usuários. Essa constatação não é surpreendente e por isso a mutiplexação estatística do tráfego gerado por um número grande de usuários é eficiente.

Um modelo de tráfego da rede de acesso residencial tem que levar em consideração a baixa utilização do canal por longos períodos de tempo. Modelos de Markov levam em consideração a correlação temporal das amostras, capturando este tipo de dinâmica. No entanto, quando se considera uma distribuição cujas amostras são independentes se torna interessante realizar a modelagem de períodos com e

sem tráfego separadamente. Com este objetivo é proposta a utilização de uma distribuição mista para modelar o tráfego residencial.

Definimos a variável aleatória X como a quantidade de bits trafegados pela rede de acesso residencial em um segundo. Supomos que X possui uma distribuição mista, onde períodos sem tráfego são modelados com a função degrau u(x), enquanto períodos de utilização do canal são modelados por uma distribuição contínua com *cumulative distribution function* (CDF) G(x) e *probability density function* (pdf) g(x).

Seja $p_0 = P[X = 0]$ e utilizando o teorema da probabilidade total define-se a seguinte função distribuição cumulativa de X:

$$P(X \leq x) = P(X \leq x|X = 0).p_0 + P(X \leq x|X > 0).(1 - p_0) \tag{5.2.1}$$

Onde as probabilidades condicionais são definidas por:

$$P(X \leq x|X = 0) = u(x) \tag{5.2.2}$$

$$P(X \leq x|X > 0) = \begin{cases} 0, & x \leq 0 \\ G(x), & x > 0 \end{cases} \tag{5.2.3}$$

Combinando 5.2.1, 5.2.2 e 5.2.3 chega-se na seguinte fórmula para a CDF de X:

$$P(X \leq x) = \begin{cases} 0, & x < 0 \\ p_0 + G(x).(1 - p_0), & x \geq 0 \end{cases} \tag{5.2.4}$$

A partir de um conjunto de amostras x_i define-se a função *likelihood* de X como:

$$L(\theta, p_0|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} (p_0 + g(x_i|\theta).(1 - p_0)) \quad = \prod_{\forall x_i = 0} p_0 \prod_{\forall x_i \neq 0} g(x_i|\theta).(1 - p_0) \tag{5.2.5}$$

Se os parâmetros p_0 e θ são independentes é possível otimiza-los separadamente [36]. O *Maximum Likelihood Estimation* para a variável Bernoulli é conhecido, sendo o parâmetro p_0 calculado através da fórmula:

$$p_0 = \sum_{i=1}^{n} \frac{\mathbb{I}(x_i = 0)}{n} \tag{5.2.6}$$

Onde $\mathbb{I}$ é a função indicadora.

O conjunto de parâmetros θ é estimado de acordo com a distribuição escolhida para g(x), sendo utilizadas neste trabalho a distribuição Weibull, a mistura de distribuições Weibull e a mistura de distribuições Gama.

Para realizar o *fitting* da distribuição Weibull foi utilizado o pacote SciPy [37] implementado em Python, que estima os parâmetros da distribuição através do método de *Maximum Likelihood Estimation*. A mistura de distribuições Gama foi criada através do pacote mixtools [38], que implementa em R o método *Expectation-Maximization* (EM). O pacote mixdist [39], também implementado em R, foi utilizado para estimar os parâmetros da mistura de distribuições Weibull através de uma combinação entre *Expectation-Maximization* e do método de Newton.

O algoritmo de EM requer a definição de valores iniciais para os parâmetros que serão estimados. Os parâmetros iniciais de cada componente da mistura de distribuições Gama foram definidos pela própria biblioteca utilizada, onde componente é definida como uma distribuição utilizada pela mistura. Segundo seus autores, é aplicado o método dos momentos em k partições obtidas a partir dos dados de entrada, onde k corresponde ao número de componentes da mistura.

A biblioteca responsável pela implementação da mistura de distribuições Weibull requer a definição das médias e variâncias iniciais de cada componente da mistura, além da discretização dos dados de entrada. Para a inicialização dos parâmetros foi implementado um algoritmo que ordena os dados de entrada, cria k partições de mesmo tamanho e obtém a média e a variância iniciais a partir do resultado de cada partição. A idéia por trás deste método de inicialização é que cada componente da mistura será responsável por modelar um pedaço da distribuição do tráfego. Para a discretização dos dados de entrada foram utilizados bins de tamanho igual a 10 kbps.

A entrada passada para o *fitting* da distribuição Weibull e da mistura de distribuições Gama é medida em bits por segundo, enquanto a estimativa dos parâmetros da mistura de distribuições Weibull foi feita com o tráfego em Megabits por segundo por conta de limitações da biblioteca utilizada. Antes de serem passadas como entrada para a estimação dos parâmetros as medições de tráfego passaram por um processo de pré-processamento. Neste processo foi atribuído o valor 0 para todas as medições abaixo de um limiar T. O objetivo do pré-processamento é reduzir a estimativa de g(x) ao tráfego de interesse, desconsiderando valores muito baixos de utilização que apenas diminuiriam a precisão do modelo. Para a obtenção dos resultados T foi definido de maneira empírica como 70 kbps. Este valor representa menos de 1% da capacidade nominal de clientes de 10 Mbps, representando uma porção de tráfego irrisório associada a tráfego de sinalização e requisições a páginas web de tamanho pequeno.

Foram geradas misturas de distribuições com o número de componentes variando entre 2 e 7. Após a estimação dos parâmetros cada modelo gerou 7776000 amostras. Uma etapa de pós-processamento foi definida com o objetivo de garantir que os resultados obtidos a partir da distribuição mista corresponderiam a valores válidos de tráfego. Nesta etapa, todas as amostras acima de um limiar M foram truncadas para o valor de M. O parâmetro M é definido de acordo com o cliente onde o tráfego foi coletado e corresponde ao maior valor de *throughput* coletado durante as medições residenciais.

5.2.2 Cadeias de Markov ocultas

Para modelar o tráfego das redes de acesso residencial dois tipos de cadeia de Markov oculta foram considerados: a cadeia oculta em que cada estado oculto gera símbolos i de acordo com uma probabilidade p_i e a Cadeia Oculta Hierárquica definida em [28], onde cada estado oculto gera símbolos de acordo com uma cadeia de Markov. Sempre que não houver ambiguidade a expressão *Cadeia Oculta Simples (COS)* será utilizada para se referir ao primeiro modelo, enquanto para o segundo modelo é utilizada a expressão *Cadeia Oculta Hierárquica (COH)*. A Figura 5.6 mostra um exemplo de COH, enquanto a COS é exemplificada na Figura 5.7.

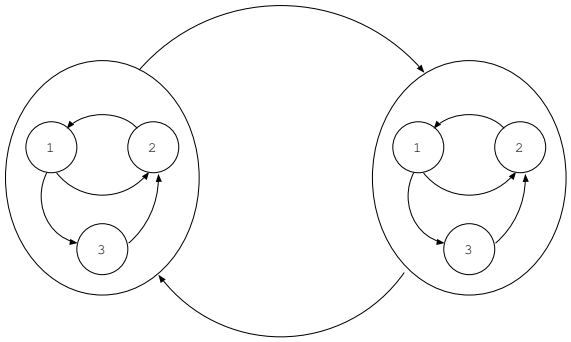


Figura 5.6: Exemplo de Cadeia Oculta Hierárquica

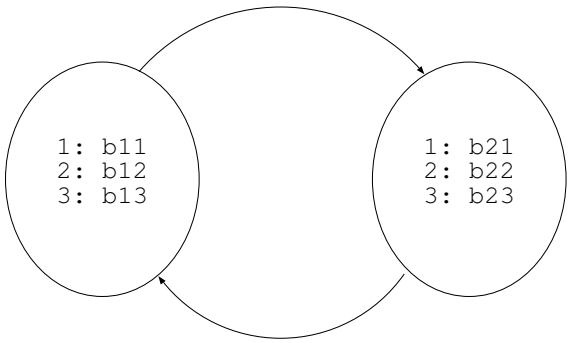


Figura 5.7: Exemplo de Cadeia Oculta Simples

Em ambas as cadeias a estrutura foi definida de maneira a permitir transições

entre todos os estados. A cadeia de Markov associada a cada estado oculto da Cadeia Oculta Hierárquica também não possui nenhum tipo de restrição, sendo possíveis transições entre todos os estados. Todos os parâmetros das cadeias de Markov ocultas consideradas foram inicializados de maneira completamente aleatória.

Para gerar os símbolos utilizados durante o treinamento das cadeias ocultas as medições de tráfego foram discretizadas de maneira uniforme entre 0 e M, onde M corresponde ao maior valor de throughput medido para um determinado voluntário. Desta forma, o tamanho de cada *bin* é igual a M/R, onde R corresponde ao número de símbolos da cadeia oculta.

Ambas as cadeias foram implementadas utilizando o MTK(*Modeling Tool Kit*), o módulo de criação de modelos matemáticos do Tangram-II [3]. O treinamento da Cadeia Oculta Simples é implementado através do algoritmo Baum-Welch, enquanto o treinamento da Cadeia Oculta Hierárquica foi implementado utilizando o algoritmo Baum-Welch com as modificações propostas em [28], ambos já implementados pelo MTK. Cada visita a um estado oculto da COH gera 60 símbolos antes de ser realizada uma transição para outro estado oculto. Outros valores para o número de símbolos gerados a cada visita a um estado oculto foram testados e apresentaram piores resultados. Após o treinamento de cada modelo foram geradas 7776000 amostras. Para calcular o tráfego sintético a partir das amostras geradas cada símbolo foi multiplicado pelo tamanho do *bin* utilizado durante as discretizações.

Para a Cadeia Oculta Simples foram utilizadas cadeias com 4 estados ocultos e 100 símbolos, enquanto a Cadeia Oculta Hierárquica foi implementada utilizando 4 estados ocultos e 10 símbolos. Foram atribuídos menos símbolos para a Cadeia Oculta Hierárquica por conta do maior número de parâmetros associados a este tipo de modelo: como existem transições entre todos os estados da cadeia de Markov responsável pela geração dos símbolos é necessário estimar $R.(R-1)$ parâmetros para cada estado oculto, onde R corresponde ao número de símbolos escolhido.

5.3 Resultados

Os resultados são apresentados em duas partes. Primeiramente é realizada uma comparação entre as três distribuições mistas consideradas e a partir dos resultados é escolhida a melhor distribuição para cada cliente. A seguir, a distribuição escolhida na primeira parte é comparada com os dois tipos de cadeia de Markov considerados. Para a segunda etapa serão utilizados quatro descritores de tráfego: a média, a variância, a distribuição e a autocorrelação.

5.3.1 Comparação entre distribuições mistas

Nesta seção são comparadas a distribuição do tráfego real e a distribuição do tráfego sintético gerado por cada distribuição mista considerada. A mistura de distribuições é sucetível a presença de singularidades, onde uma componente possui variância próxima de 0 e o *likelihood* tende a infinito [29]. Por este motivo não utilizamos para a avaliação dos modelos o *log-likelihood*. As comparações entre as distribuições são realizadas através do gráfico Q-Q plot.

Primeiramente é analisado o efeito da variação do número de componentes das misturas de Gama e Weibull. As Figuras 5.8 e 5.9 mostram os resultados obtidos a partir do tráfego gerado por um dos voluntários. Nota-se que para este cliente o aumento do número de componentes não se traduz em um modelo que aproxime melhor a distribuição de interesse e que os melhores resultados são obtidos com um número pequeno de componentes. Os métodos de inicialização utilizados particionam os dados de acordo com o número de componentes, ou seja, diferentes partições serão obtidas para diferentes números de componentes. A partir de diferentes partições serão calculados diferentes valores iniciais para os parâmetros das distribuições, ou seja, a inicialização pode explicar a diferença observada. O aumento do número de componentes também não apresenta melhores resultados para outros 4 clientes.

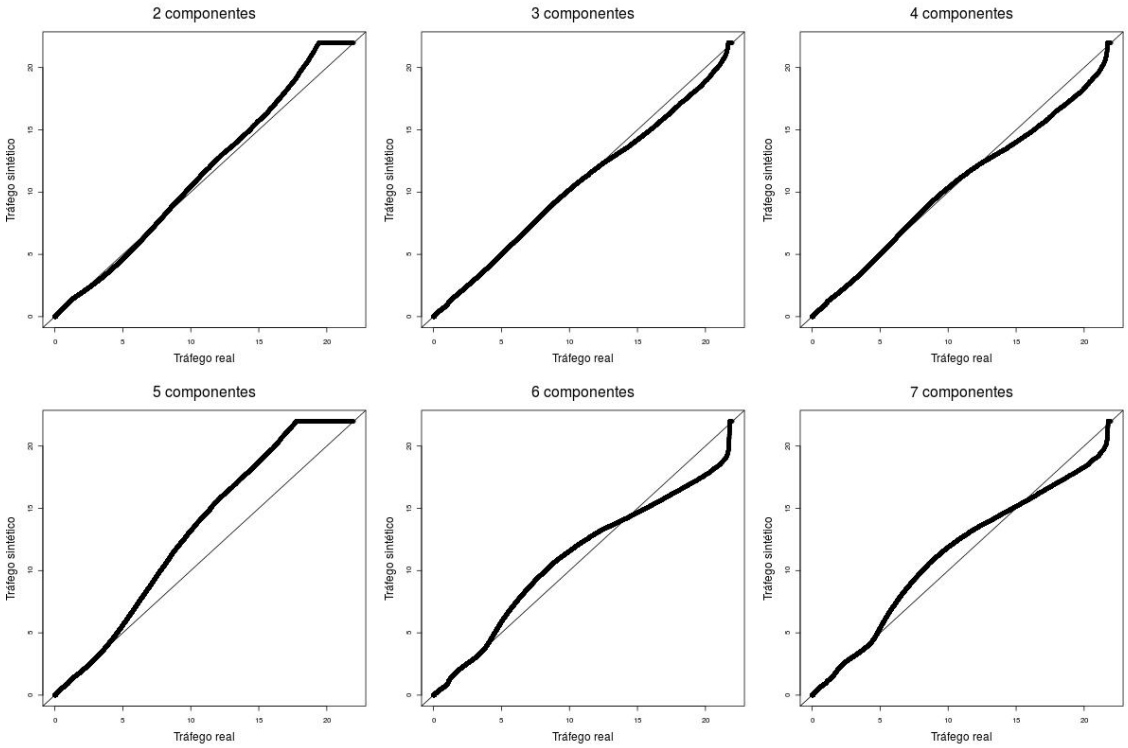


Figura 5.8: Comparação entre misturas de Weibull e tráfego real do voluntário 1

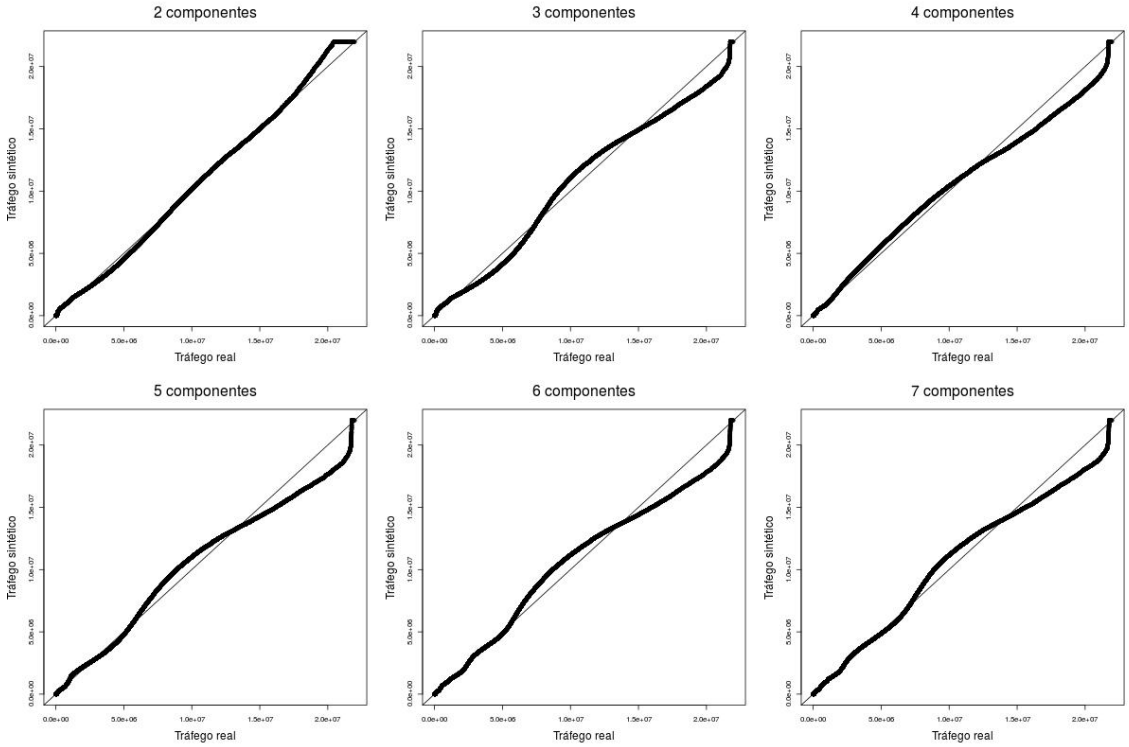


Figura 5.9: Comparação entre misturas de Gama e tráfego real do voluntário 1

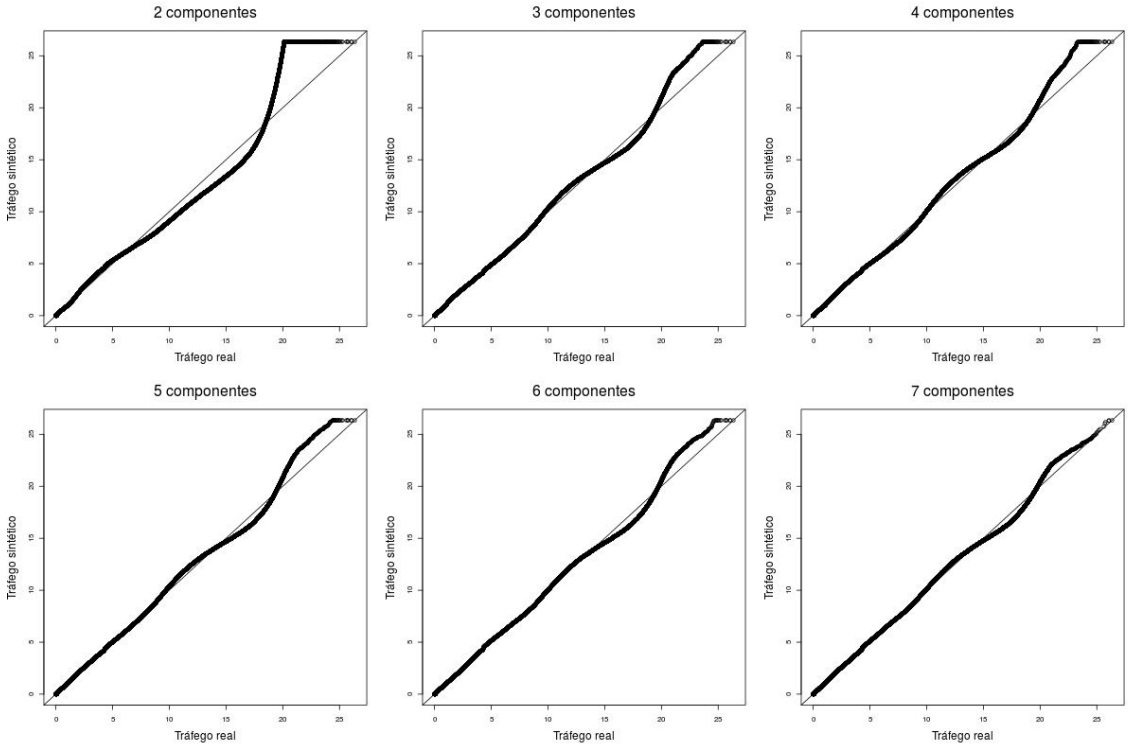


Figura 5.10: Comparação entre misturas de Weibull e tráfego real do voluntário 7

Os resultados dos voluntários 2, 4, 6 e 7 apresentam um comportamento mais próximo do esperado, com o aumento do número de componentes se traduzindo em

melhores estimativas para os parâmetros da distribuição real, como mostrado na Figura 5.10. Neste caso, o melhor resultado é obtido utilizando uma mistura com 7 componentes.

Para cada mistura foi escolhido através da análise do Q-Q plot o número de componentes que melhor aproximava a distribuição real. Em seguida foi realizada uma comparação entre as melhores misturas obtidas na primeira etapa e a distribuição de Weibull. A Figura 5.11 mostra o resultado obtido considerando o tráfego de um dos voluntários. Percebe-se que a distribuição de Weibull não é suficientemente precisa na modelagem do tráfego residencial deste usuário, resultado que se repete para todos os outros clientes. As misturas, por outro lado, modelam de maneira bastante precisa a distribuição original. Nota-se que a diferença entre a mistura de Gama e a mistura de Weibull é mínima, com resultados melhores para a mistura de Weibull. A similaridade entre os resultados de ambas as misturas se mantém para outros voluntários, com resultados melhores observados para a mistura de Weibull na maior parte dos casos. Desta forma, a mistura de Weibull é a distribuição mista escolhida para a comparação dos diferentes modelos utilizando descritores de tráfego.

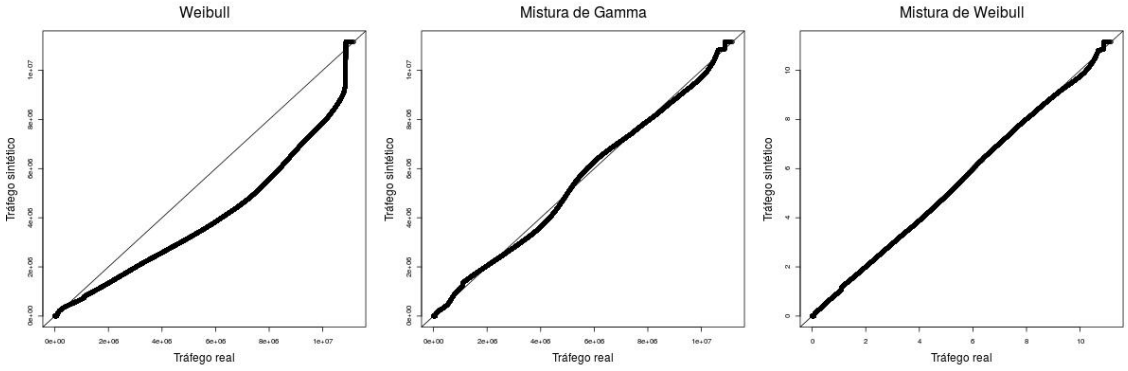


Figura 5.11: Comparação entre distribuições mistas e tráfego real gerado por voluntário 9

Apenas um voluntário não obteve uma distribuição mista que conseguisse aproximar de maneira satisfatória a distribuição real do tráfego, como mostra a Figura 5.12. A Figura 5.13 mostra o histograma do tráfego real com *bins* de tamanho igual a 10 kbps e a pdf da mistura de distribuições Weibull. Nota-se que a mistura não consegue modelar a parte inicial nem o pico no final da distribuição. Como mostramos na Tabela 5.2 este voluntário possui um perfil diferente dos demais clientes, com boa parte de seu tráfego abaixo de 1 Mbps, o que pode explicar a diferença encontrada durante o processo de modelagem.

Para verificar se a inicialização é a responsável por este resultado foi experimentado um segundo método para a inicialização dos parâmetros. Neste método, o intervalo entre 0 e M é dividido uniformemente em k partições e os parâmetros

iniciais de cada componente i são calculados a partir das amostras pertencentes ao i-ésimo intervalo, onde M corresponde ao maior valor de throughput obtido durante as medições. A Figura 5.14 mostra os resultados obtidos com a nova inicialização. Percebe-se que a mistura de Weibull com 6 componentes tem um resultado melhor do que o obtido utilizando a primeira inicialização. Este resultado exemplifica o impacto do método de inicialização para o algoritmo de EM, mas não altera as conclusões obtidas. Um estudo aprofundado sobre diferentes métodos de inicialização de parâmetros para misturas de distribuições está fora do escopo deste trabalho.

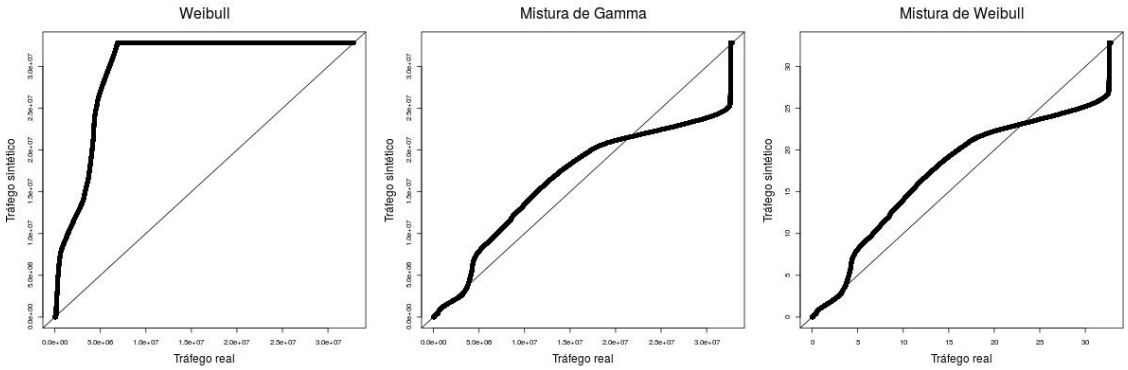


Figura 5.12: Comparação entre distribuições mistas e tráfego real gerado por voluntário 5

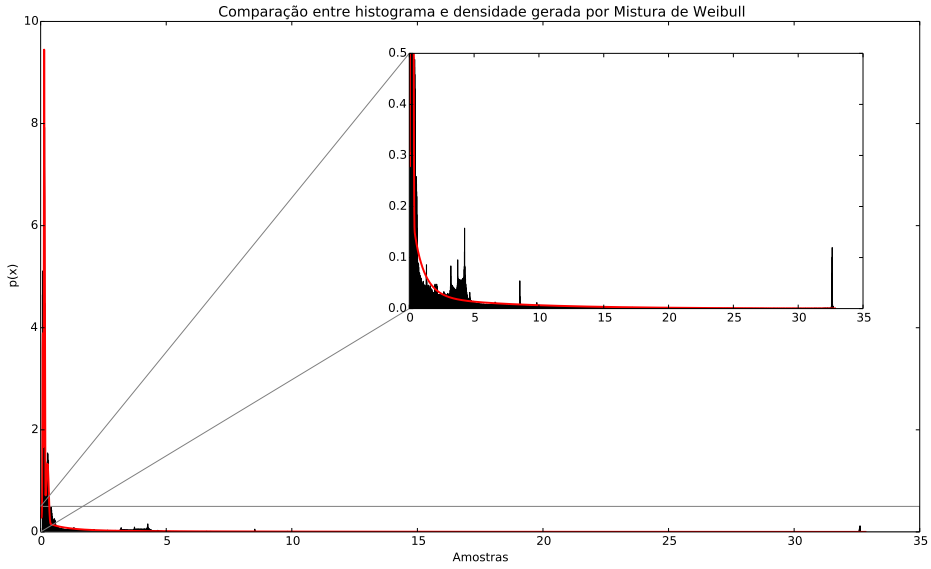


Figura 5.13: Comparação entre histograma dos dados reais e densidade gerada por mistura de Weibull

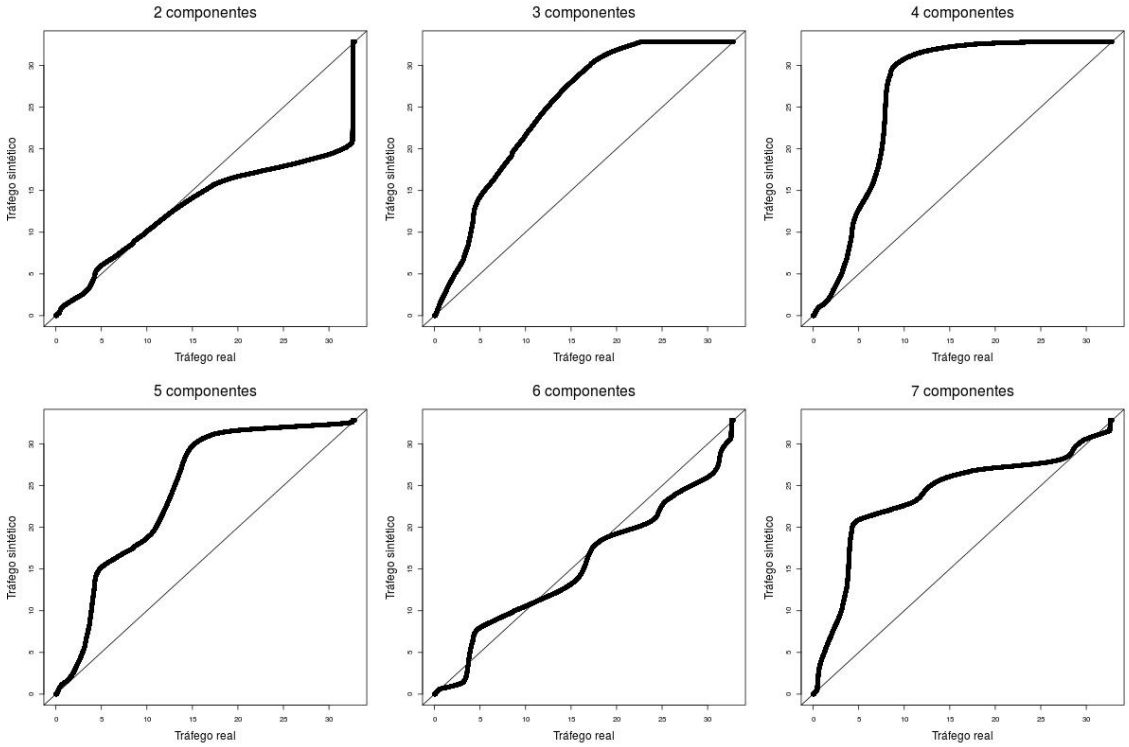


Figura 5.14: Comparação entre misturas de Weibull utilizando segunda inicialização e tráfego real do voluntário 5

5.3.2 Comparação utilizando descritores de séries temporais

Existem na literatura vários descritores de tráfego propostos no passado que se mostraram eficazes para comparar modelos de tráfego e o tráfego real [5, 40, 41]. Estes descritores são métricas de comparação, sem ser necessário comparar a distribuição completa. Alguns deles capturam correlações temporais que por ventura existam no tráfego real. Neste trabalho escolhemos descritores simples como a média e a variância. Usamos também a distribuição e a autocorrelação, este último para comparar correlações temporais. Para a comparação com modelos de cadeias ocultas foi gerado tráfego sintético para cada modelo e, a partir deste *trace*, calculamos os descritores. Para os modelos de distribuição mista as métricas foram calculadas diretamente.

As Figuras 5.15 e 5.16 mostram a média e a variância do tráfego obtidas pelas medições e geradas pelos modelos, enquanto as tabelas 5.6 e 5.7 mostram o erro absoluto obtido por cada modelo em relação a média e a variância reais. Nota-se que ambas as métricas são modeladas de maneira precisa para 8 entre os 9 voluntários. As únicas exceções são os valores de média e variância gerados pela Cadeia Oculta Hierárquica para o voluntário 8. Como mostramos na Tabela 5.3 este cliente apresenta um perfil distinto de tráfego, com fração maior de medições associada a tráfego médio e alto, o que pode explicar a pouca precisão da COH neste caso. Além

disso, o tamanho do *bin* utilizado para a discretização do tráfego deste cliente em símbolos de entrada para a COH é de 3.6 Mbps. Testes realizados com 20 símbolos melhoram os resultados obtidos por este cliente, o que confirma o problema relacionado ao tamanho do *bin*. Como a média não é bem modelada neste caso, a COH não é um bom modelo para o tráfego deste cliente. Quando considerados outros clientes a Cadeia Oculta Hierárquica é razoavelmente precisa mesmo utilizando apenas 10 símbolos.

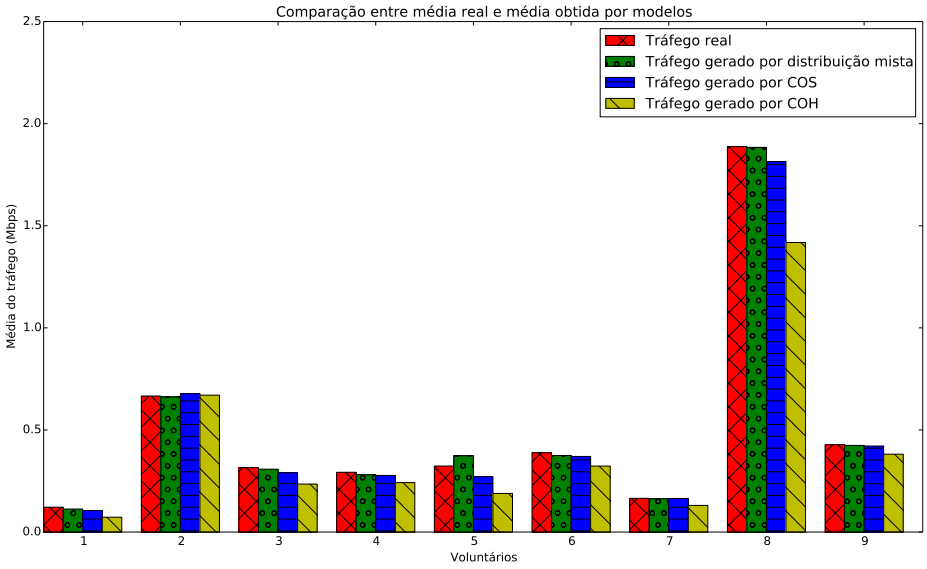


Figura 5.15: Comparação entre média real e média obtida a partir dos modelos

Tabela 5.6: Erro absoluto da média (Mbps)

	Distribuição mista	COS	COH
Voluntário 1	0,01	0,02	0,05
Voluntário 2	0,00	0,01	0,00
Voluntário 3	0,01	0,03	0,08
Voluntário 4	0,01	0,02	0,05
Voluntário 5	0,05	0,05	0,13
Voluntário 6	0,01	0,02	0,07
Voluntário 7	0,00	0,00	0,03
Voluntário 8	0,00	0,07	0,47
Voluntário 9	0,00	0,01	0,05

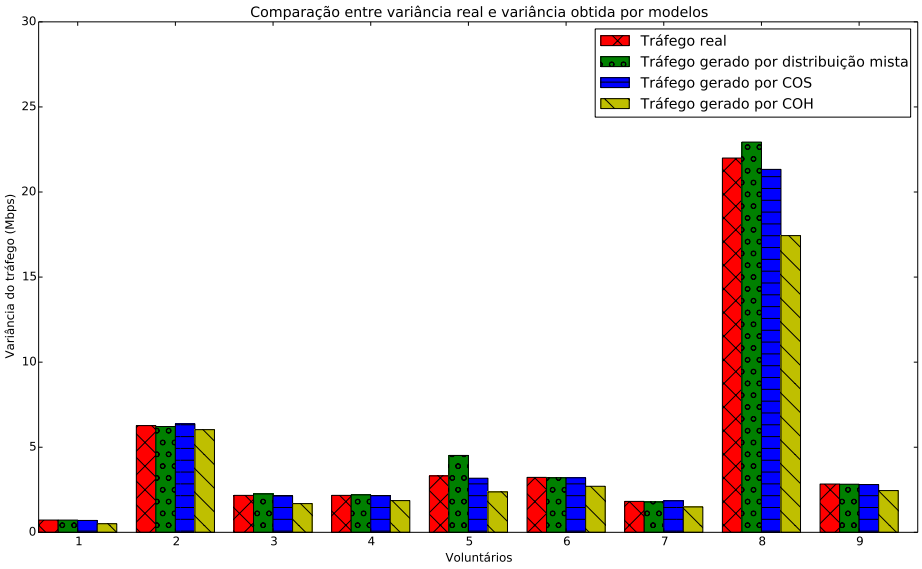


Figura 5.16: Comparação entre variância real e variância obtida a partir dos modelos

Tabela 5.7: Erro absoluto da variância (Mbps)

	Distribuição mista	COS	COH
Voluntário 1	0,00	0,02	0,22
Voluntário 2	0,05	0,10	0,24
Voluntário 3	0,09	0,03	0,48
Voluntário 4	0,03	0,01	0,31
Voluntário 5	1,20	0,13	0,94
Voluntário 6	0,03	0,02	0,52
Voluntário 7	0,04	0,03	0,33
Voluntário 8	0,94	0,67	4,56
Voluntário 9	0,01	0,03	0,39

As distribuições do tráfego são comparadas através da utilização de histogramas com 10 *bins* de tamanho igual a M/10, onde M é o maior valor de throughput medido por um determinado voluntário. O número de *bins* foi determinado através do menor valor de granularidade dentre os modelos considerados, neste caso definida pelas Cadeias Ocultas Hierárquicas de 10 símbolos. Também é calculada para cada modelo a distância de Jensen-Shannon, uma versão simétrica e suavizada da distância de Kullback-Leibler [31]. A partir desta métrica é possível quantificar a distância entre a distribuição do tráfego real e a distribuição gerada pelos modelos. Assim como nos histogramas, a distância de Jensen-Shannon foi calculada utilizando 10 *bins*.

As Tabelas 5.8 e 5.10 mostram as probabilidades associadas ao primeiro bin da distribuição do tráfego real e dos modelos criados para dois clientes. Nota-se que o primeiro *bin* é modelado de maneira precisa por todos os modelos. Para comparar

a distribuição dos outros *bins* é considerada a distribuição condicional dado que o tráfego não pertence ao primeiro *bin*, mostrado nas Figuras 5.17 e 5.18

Percebe-se que a Cadeia Oculta Simples é o modelo que melhor representa a distribuição dos dados reais em ambos os casos. A Cadeia Oculta Hierárquica é pouco precisa na modelagem da distribuição apresentada na Figura 5.17, mas apresenta resultados próximos da distribuição real para o tráfego do voluntário representado pela Figura 5.18. Nota-se ainda que nos histogramas gerados pela distribuição mista os últimos *bins* possuem probabilidade maior do que a observada nos dados reais. Isto pode ser explicado pelo acúmulo de probabilidade na cauda das distribuições mistas, as quais são definidas para valores entre 0 e infinito. Resultados similares são observados para os outros voluntários. As Tabelas 5.9 e 5.11 mostram para cada modelo o cálculo da distância de Jensen-Shannon. Os resultados obtidos confirmam que a COS é o modelo que melhor aproxima a distribuição do tráfego real para ambos os clientes.

Tabela 5.8: Probabilidades associadas ao primeiro bin para voluntário 1

	Probabilidade de tráfego estar no primeiro *bin*
Tráfego real	0,983
Tráfego gerado por distribuição mista	0,977
Tráfego gerado por COS	0,984
Tráfego gerado por COH	0,984

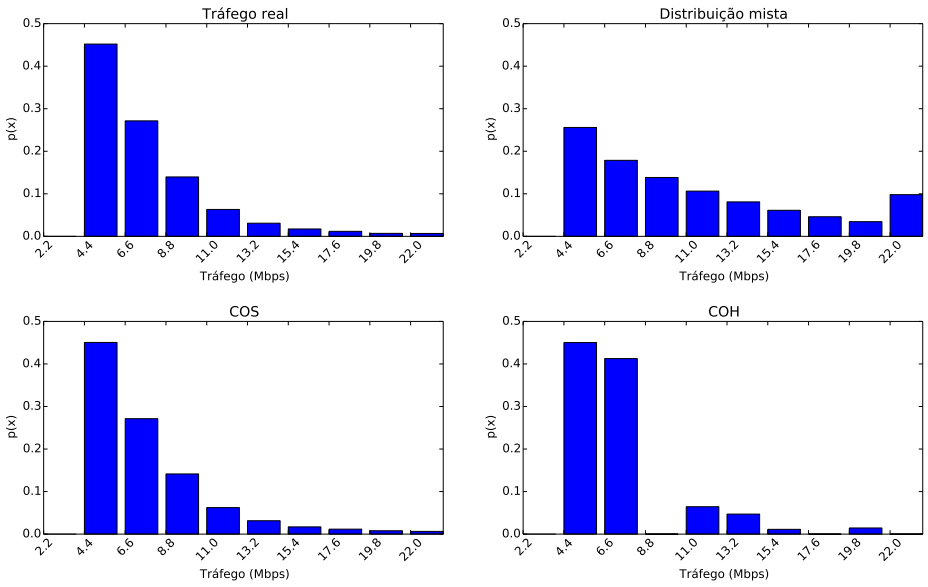


Figura 5.17: Comparação entre distribuições condicionais de tráfego do voluntário 1

Tabela 5.9: Distância de Jensen-Shannon para modelos de tráfego do voluntário 1

	Distância de Jensen-Shannon
Distribuição mista	$1,6 \; . \; 10^{-3}$
COS	$5,1 \; . \; 10^{-7}$
COH	$1,0 \; . \; 10^{-3}$

Tabela 5.10: Probabilidades associadas ao primeiro bin para voluntário 4

	Probabilidade de tráfego estar no primeiro *bin*
Tráfego real	0,957
Tráfego gerado por distribuição mista	0,957
Tráfego gerado por COS	0,957
Tráfego gerado por COH	0,957

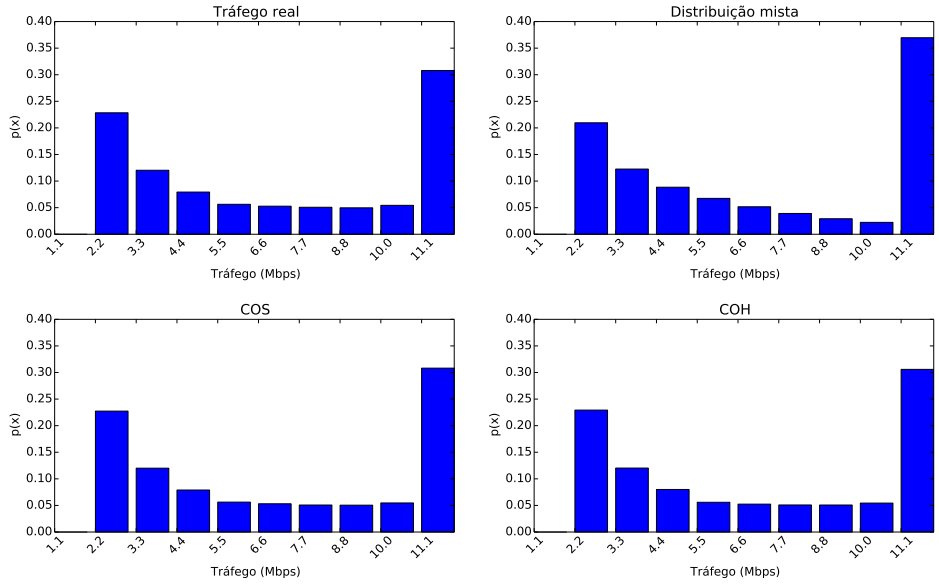


Figura 5.18: Comparação entre distribuições condicionais de tráfego do voluntário 4

Tabela 5.11: Distância de Jensen-Shannon para modelos de tráfego do voluntário 4

	Distância de Jensen-Shannon
Distribuição mista	$3,0 \; . \; 10^{-4}$
COS	$1,3 \; . \; 10^{-7}$
COH	$1,3 \; . \; 10^{-6}$

Por fim, considera-se a autocorrelação capturada por cada modelo. A distribuição mista apresenta autocorrelação 0 por conta da independência entre suas amostras e não terá seus resultados exibidos. As Figuras 5.19 e 5.20 mostram a

autocorrelação obtida pelos dados reais e sintéticos gerados pelos modelos de COS e COH. Percebe-se que em ambos os casos a Cadeia Oculta Hierárquica é o modelo que mais se aproxima da autocorrelação real para valores maiores de lag, tendo um decaimento mais lento do que o observado na Cadeia Oculta Simples, que modela melhor a autocorrelação para *lags* menores mas decai rapidamente. O tráfego do cliente observado na Figura 5.19 possui autocorrelação real muito maior do que a capturada pelos modelos, tendo um decaimento mais lento do que o obtido a partir dos tráfegos sintéticos. Já a autocorrelação obtida pelo tráfego do segundo cliente, mostrada na Figura 5.20, é menor e é modelada de maneira melhor pela Cadeia Oculta Hierárquica. Os outros clientes apresentaram comportamento similar ao encontrado na Figura 5.19.

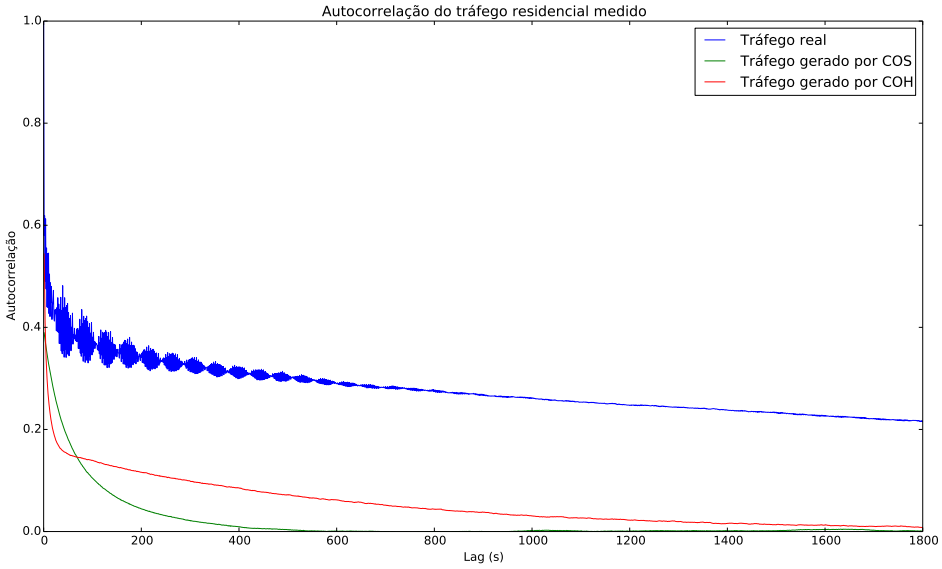


Figura 5.19: Comparação da autocorrelação de dados reais e sintéticos para voluntário 8

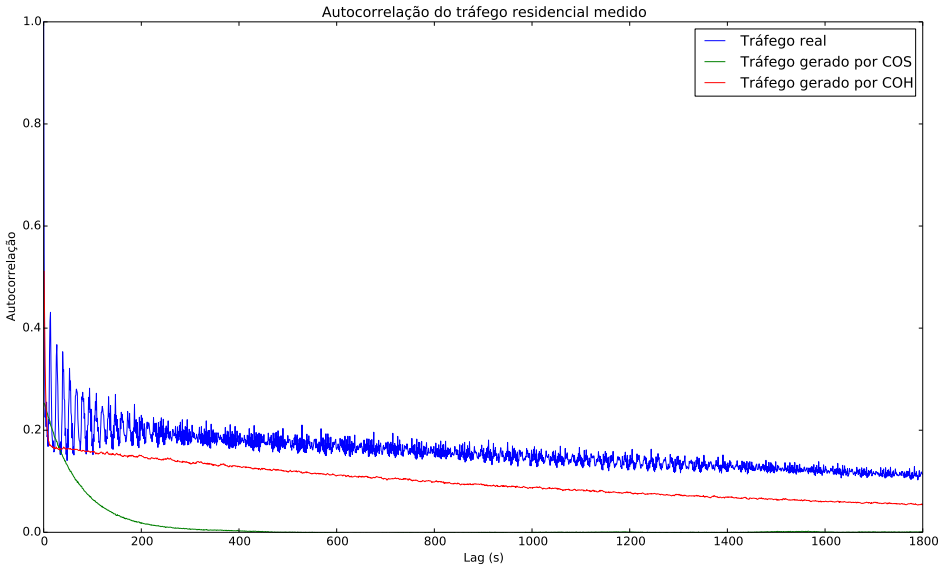


Figura 5.20: Comparação da autocorrelação de dados reais e sintéticos para voluntário 7

Como podemos observar a autocorrelação nao é capturada de maneira precisa pelos modelos de COS e COH. Isso pode ser explicado pelo fato do tráfego real ter rajadas em diferentes escalas de tempo, como observado na Figura 5.2, pois os modelos usados neste trabalho em geral não capturam de maneira suficientemente precisa este comportamento.

As Figuras 5.21, 5.22 e 5.23 mostram a dinâmica do tráfego sintético obtida a partir de diferentes níveis de agregação de 604800 amostras de cada modelo considerado. Percebemos que as cadeias ocultas apresentam tráfego em rajada para diferentes escalas de tempo, enquanto o tráfego gerado pela distribuição mista se torna mais suave a medida em que se aumenta o período de agregação. Estes resultados estão de acordo com o observado em [5], que mostra que uma autocorrelação de decaimento mais lento se traduz em tráfego apresentando rajadas em diferentes escalas de tempo.

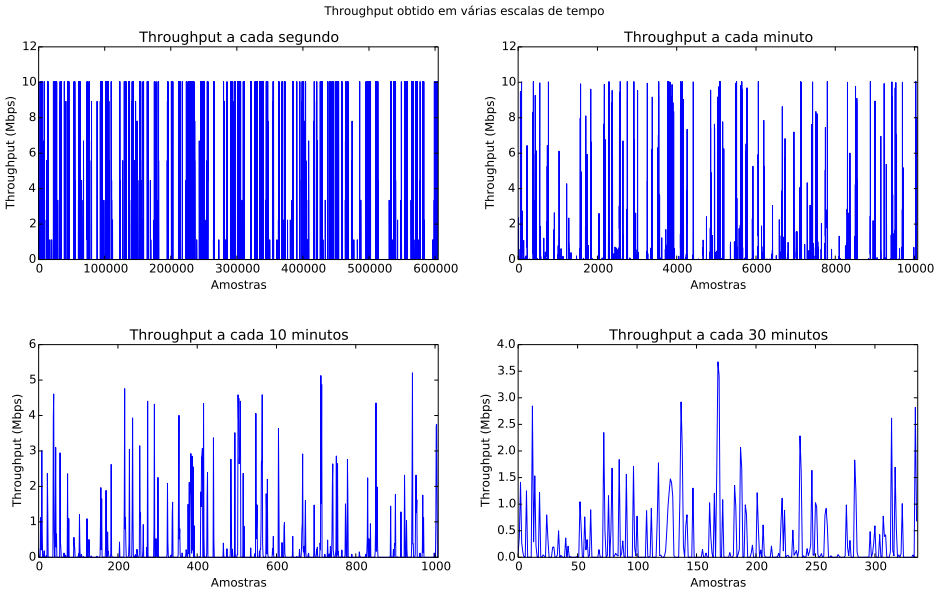


Figura 5.21: Tráfego gerado por Cadeia Oculta Hierárquica agregado em diferentes escalas de tempo

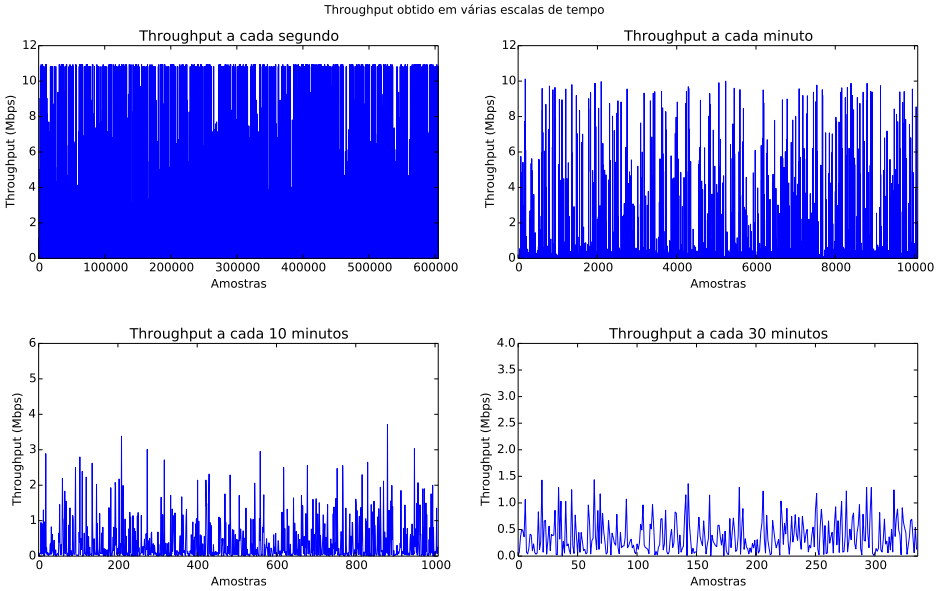


Figura 5.22: Tráfego gerado por Cadeia Oculta Simples agregado em diferentes escalas de tempo

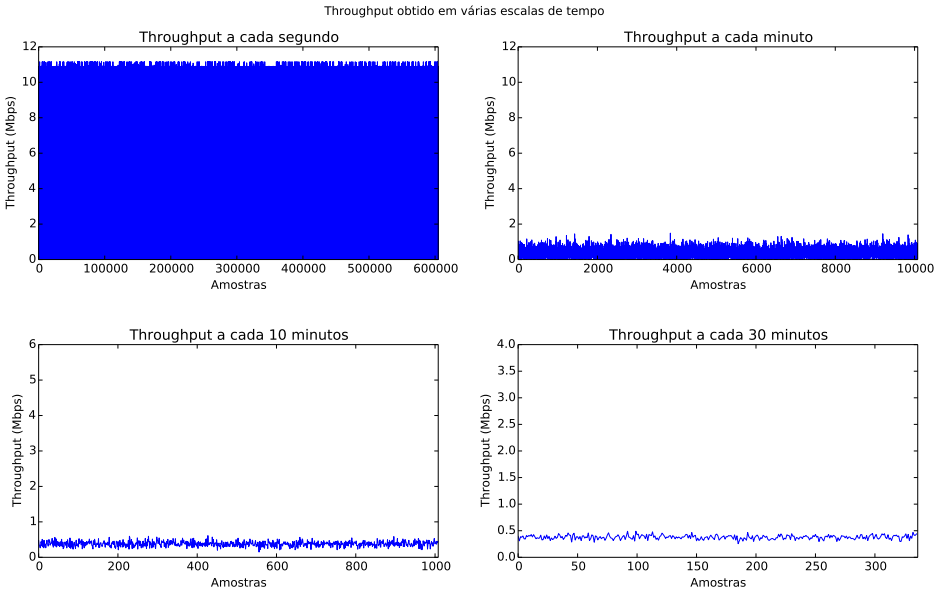


Figura 5.23: Tráfego gerado por distribuição mista agregado em diferentes escalas de tempo

5.4 Conclusões sobre os resultados obtidos

Neste capítulo é realizada a análise do tráfego observado em redes de acesso residencial e 3 tipos de modelos são propostos, a distribuição mista, a Cadeia Oculta Simples e a Cadeia Oculta Hierárquica.

O tráfego coletado apresenta rajadas em diferentes escalas de tempo e possui autocorrelação de decaimento lento, reproduzindo o comportamento observado em [5]. Além disso, na maior parte do tempo o número de bytes trafegados é baixo, o que é esperado pois é este tipo de comportamento que permite a multiplexação estatística.

Quando considerados os modelos, percebe-se que o modelo de distribuição mista apresenta resultados satisfatórios para a modelagem do tráfego, mas não captura a autocorrelação e por isso o tráfego sintético agregado em diferentes escalas de tempo não reproduz a dinâmica observada no tráfego real. A Cadeia Oculta Simples é o modelo que melhor modela a distribuição do tráfego real, mas sua autocorrelação decai rapidamente. A Cadeia Oculta Hierárquica, por outro lado, não apresenta resultados precisos na modelagem da distribuição de alguns clientes, mas sua autocorrelação decai mais lentamente do que a observada na COS. No entanto, nenhum dos modelos considerados modela de maneira precisa a alta autocorrelação observada no tráfego residencial.

Capítulo 6

Processo de perda em redes de acesso residencial

Uma das métricas fundamentais para a definição de qualidade de serviço em redes é a perda de pacotes. A perda interfere diretamente no *throughput* alcançado por conexões TCP, uma vez que o mecanismo de controle de congestionamento altera o tamanho de sua janela de acordo com as perdas inferidas. Além disso, a perda de pacotes é um dos principais fatores de degradação da qualidade de aplicações multimídia como voz sobre IP [42]. A criação de modelos permite um melhor entendimento do processo de perda e torna possível a predição de seu comportamento. Desta forma, seus resultados podem ser aplicados para otimizar mecanismos adaptativos de recuperação de perda como o *Forward Error Correction* (FEC) [12] e para a criação de modelos de simulação para estudar o impacto das perdas em aplicações específicas.

Neste trabalho são comparados diversos modelos para o processo de perda utilizando técnicas de aprendizado de máquina a partir de dados de medições coletadas em redes de acesso residencial.

6.1 Dataset

Para a criação dos modelos foram coletados dados de perda bidirecional de 647 clientes entre julho e dezembro de 2014, totalizando 135779 amostras. As coletas foram realizadas a partir de roteadores residenciais utilizando uma versão modificada e embarcada do módulo de medições de rede do software Tangram-II [3]. Estas medições foram coletadas em residências de colaboradores da NET como parte do projeto Inova Telecom, realizado em parceria com a universidade e com a TGR, uma empresa incubada na COPPE. O servidor utilizado para as medições é controlado pela NET e se encontra dentro da rede do ISP, sendo ligado diretamente a rede

HFC.

Foi realizada uma comparação entre a taxa de perda obtida a partir de nossas medições e a taxa de perda coletada em setembro de 2013 pelo programa *Measuring Broadband America*, disponível em [43]. Ambas as medições foram coletadas utilizando pacotes UDP, mas a metodologia utilizada pelo FCC é diferente da aplicada durante as medições deste trabalho. Nas medições realizadas nos Estados Unidos até 600 pacotes UDP com 16 bytes de *payload* são distribuídos de maneira aleatória durante o período de 1 hora. Se um determinado pacote não é recebido durante 3 segundos ele é considerado perdido [44]. Nota-se que na metodologia utilizada pelo FCC não é possível medir a rajada de perda, já que os pacotes são enviados individualmente, enquanto a partir de nossas medições é possível a obtenção de métricas relacionadas a rajada de perda.

Ambos os histogramas de taxa de perda foram gerados utilizando bins de tamanho 0,03, pois esta é a menor granularidade possível de taxa de perda considerando o número de pacotes enviado durante as medições deste trabalho. A Figura 6.1 mostra o histograma da taxa de perda gerado a partir das medições do FCC, enquanto a Figura 6.2 mostra o histograma obtido a partir das medições realizadas. Em ambos os casos mais de 95% das medições tem taxa de perda menor que 3%. Além disso, ambos os histogramas contém pequenos picos em taxas altas de perda. Testes preliminares realizados com o protocolo ICMP também obtiveram picos em altas taxas. A presença de indisponibilidades temporárias na conexão entre o servidor e o cliente é uma possível explicação para este tipo de comportamento. Por fim, apesar da diferença das metodologias e do maior número de medições com altas taxas de perda coletadas neste trabalho o comportamento apresentado é similar.

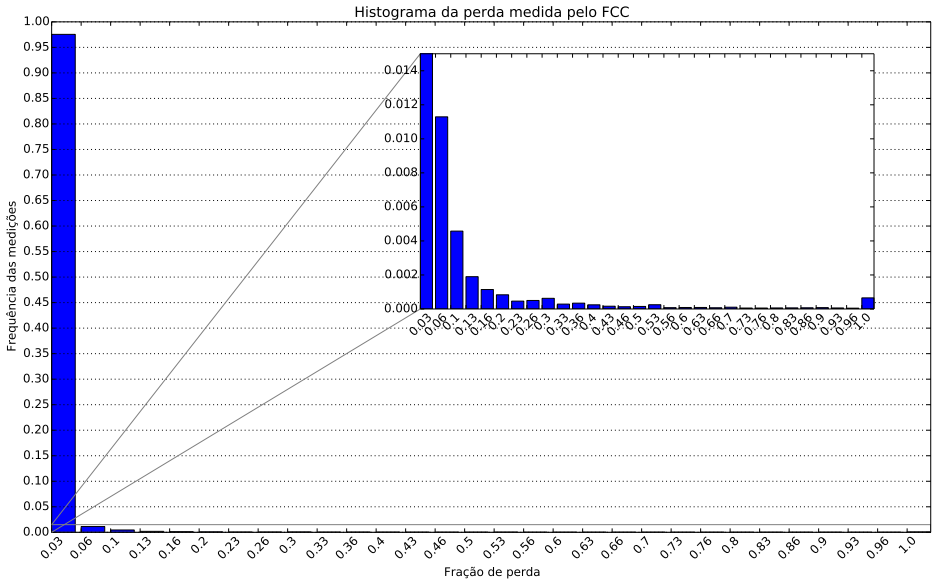


Figura 6.1: Histograma da taxa de perda medida pelo FCC

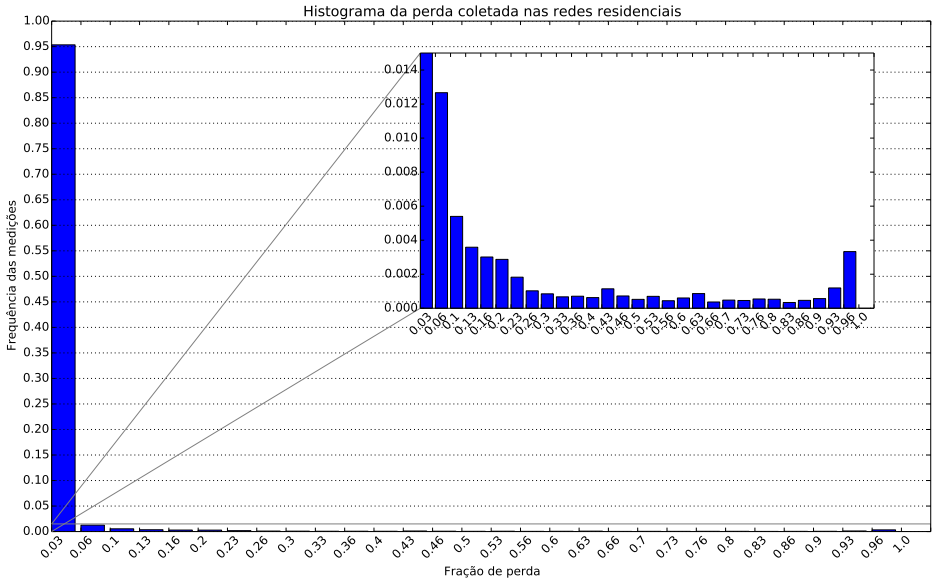


Figura 6.2: Histograma da taxa de perda coletada nas redes residenciais

A aplicação de um algoritmo de clusterização foi utilizada para entender o comportamento da variação da taxa de perda associada aos usuários. Foram calculadas as taxas médias de perda por hora de clientes com ao menos 5 medições em cada hora, totalizando 80 clientes. Em seguida foram gerados 3 *clusters* de 24 dimensões utilizando o algoritmo de *k-means* [35]. Os resultados são mostrados nas Figuras 6.3 e 6.4. Percebe-se a geração de três tipos distintos de *clusters* de usuários. O primeiro *cluster* representa os clientes com taxa de perda próxima de 0 durante a

maior parte do dia, ou seja, aqueles que observam pouco congestionamento na rede. Foram associados a este *cluster* 80% dos usuários considerados. O segundo agrupamento apresenta um centróide cuja taxa média de perda se mantém próxima de 0,05, representando clientes com taxa de perda intermediária. 15 clientes são associados ao segundo *cluster*. Por fim, o terceiro *cluster* é formado por um único usuário e apresenta uma taxa média de perda de até 0,6. Este cliente apresenta uma rede problemática em que se observa congestionamento na rede e altas taxas de perda.

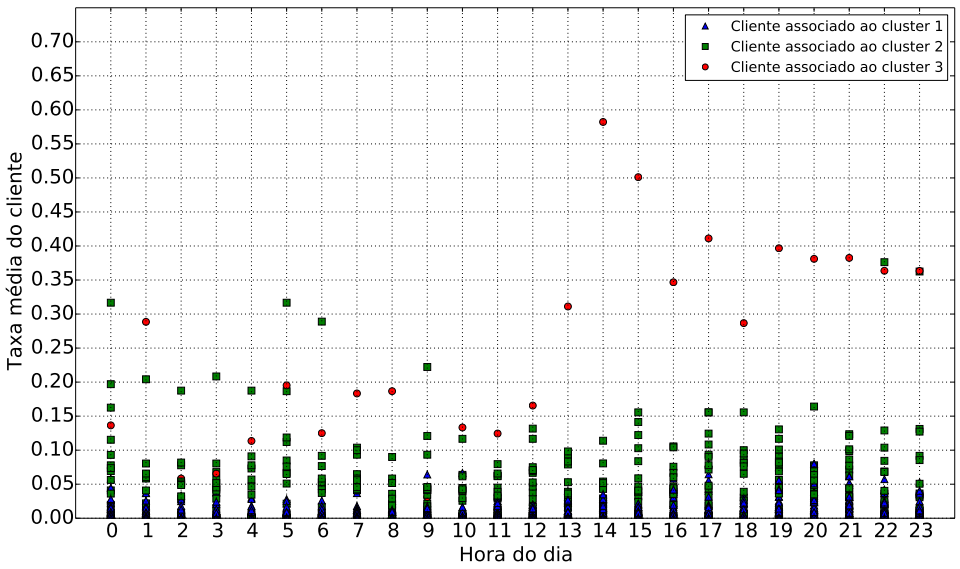


Figura 6.3: *Clusters* obtidos a partir da taxa de perda dos usuários

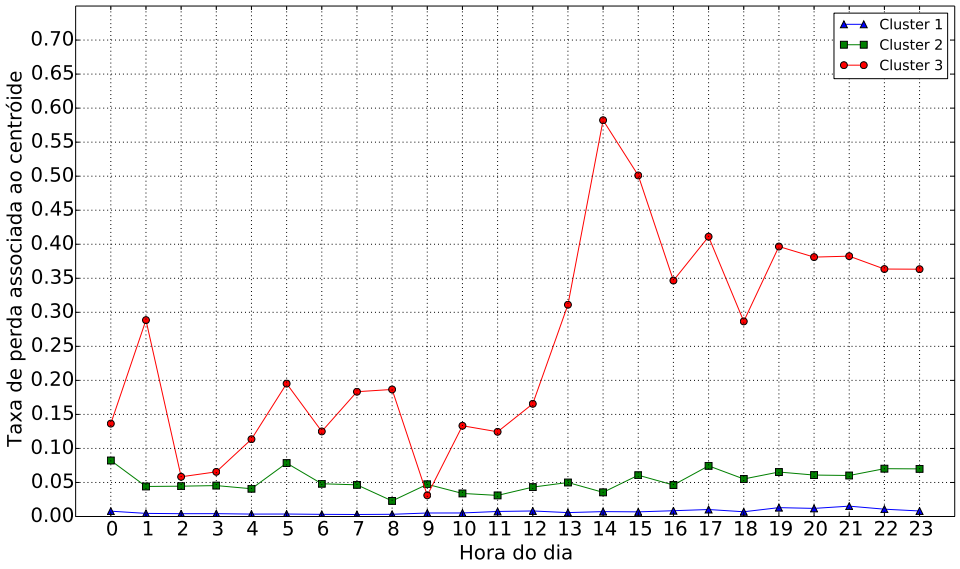


Figura 6.4: Taxa de perda por hora associada a cada centróide

6.2 Metodologia

Três tipos de modelos para o processo de perda foram considerados nas avaliações: o modelo de Gilbert simplificado, a cadeia de Markov oculta em que cada estado oculto i gera símbolos 0 com probabilidade p_i e símbolos 1 com probabilidade $1 - p_i$, descrita em SALAMATIAN e VATON [11], e a cadeia de Markov oculta hierárquica em que cada estado oculto gera símbolos 0 e 1 através de um modelo de Gilbert simplificado, criada em SILVEIRA e DE SOUZA E SILVA [12]. Utiliza-se neste trabalho a expressão *Cadeia Oculta Simples (COS)* quando o modelo considerado é o definido em [11], enquanto para o modelo presente em [12] é utilizada a expressão *Cadeia Oculta Hierárquica (COH)*.

Os modelos utilizados neste trabalho recebem como entrada uma sequência de símbolos extraída a partir dos resultados das medições realizadas em redes de acesso residenciais. Nesta sequência, o símbolo 1 representa a perda de um pacote e o símbolo 0 representa o seu recebimento. A concatenação das sequências de símbolos geradas por todas as medições realizadas foi utilizada durante as fases de treinamento e teste de cada modelo.

O modelo de Gilbert simplificado possui 2 estados, enquanto modelos variando entre 2 e 7 estados foram testados para ambas as cadeias de Markov ocultas. A sequência de entrada para as Cadeias Ocultas Hierárquicas é segmentada em lotes de tamanho b, sendo b um parâmetro a ser passado para o modelo. Para a modelagem deste trabalho b foi definido como 30, baseando-se na idéia de que cada medição captura um estado de curto prazo da rede.

O modelo de Gilbert simplificado e a Cadeia Oculta Simples foram treinados utilizando o algoritmo de Baum-Welch implementado na ferramenta MTK (*Modeling Tool Kit*), o módulo de criação de modelos matemáticos presente no Tangram-II [3]. Para a Cadeia Oculta Hierárquica foi utilizado o Baum-Welch com as modificações propostas em [12], também implementado no MTK.

Neste trabalho são testadas duas estruturas para cadeias de Markov ocultas. Na primeira estrutura nenhum tipo de limitação é imposta ao modelo, sendo permitidas transições entre todos os estados ocultos e não sendo definidas limitações na geração de símbolos de nenhum dos estados. A segunda estrutura também permite transições entre todos os estados ocultos, mas é forçada a existência de um estado que gere apenas o símbolo 0, ou seja, um estado que não gere perdas. A segunda estrutura é inspirada na alta probabilidade de perdas de tamanho 0, pois em mais de 95% dos *traces* coletados nenhuma perda ocorre. Utilizamos a expressão *Estrutura Simples* para fazer referência a primeira estrutura, enquanto a segunda estrutura é denominada *Estrutura Modificada*.

Os valores iniciais para a matriz **A** são definidos de maneira aleatória em todos

os casos. O vetor π, que indica a probabilidade inicial de cada estado, também possui valores iniciais definidos de maneira aleatória para todos os modelos, assim como a matriz $\mathbf{B}$, responsável pela geração de símbolos.

Por conta da utilização de valores aleatórios para a inicialização dos parâmetros cada configuração foi gerada 10 vezes, sendo considerado para os resultados finais apenas o modelo gerado que obteve o maior *likelihood* a partir dos *traces* originais. Foi possível notar durante este processo que os resultados obtidos pela Cadeia Oculta Simples com poucos estados variam bastante quando considerada a Estrutura Simples, diferentemente da Cadeia Oculta Hierárquica, que apresentou maior estabilidade em relação aos parâmetros iniciais. Uma avaliação mais profunda sobre o impacto de diferentes valores para os parâmetros iniciais está fora do escopo deste trabalho.

Três métricas foram definidas para comparar os modelos criados: o *log-likelihood* da sequência de treinamento aplicada ao modelo, a distribuição do tamanho da rajada de perda e a distribuição do intervalo entre perdas. O tamanho da rajada de perda corresponde ao número de perdas consecutivas em um *trace*, representada por uma sequência de símbolos 1. O intervalo entre perdas é calculado como o número de símbolos 0 entre dois símbolos 1. A Figura 6.5 mostra um exemplo de como estas duas métricas são obtidas.

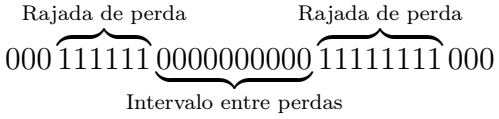


Figura 6.5: Exemplo de obtenção de métricas de interesse

Para avaliar os modelos em relação as distribuições da rajada de perda e do intervalo entre perdas foi realizada uma comparação entre os valores extraídos a partir das medições realizadas e os valores obtidos a partir de *traces* sintéticos gerados por cada modelo. Cada *trace* sintético tinha tamanho de 6000000 e era dividido em pedaços de tamanho 30 para as comparações. A avaliação dos resultados é realizada a partir de histogramas, do gráfico Q-Q plot e da distância de Jensen-Shannon, uma versão simétrica da distância de Kullback–Leibler [31].

6.3 Resultados

A comparação entre os valores de *log-likelihood* obtidos por cada modelo é mostrada na Figura 6.6. A Cadeia Oculta Hierárquica é aquela que apresenta os melhores resultados quando considerada esta métrica, independentemente do número de estados e da estrutura considerada. O modelo de Gilbert simplificado apresenta o pior resultado, o que era esperado por conta de sua grande simplicidade. É possível

perceber que o aumento do número de estados das cadeias de Markov ocultas não gera um aumento significativo no *log-likelihood* para um número maior do que 4 estados. Este resultado está de acordo com [11], que observa que 4 estados ocultos são suficientes para modelar o processo de perda. A Estrutura Modificada gera resultados ligeiramente melhores para Cadeias Ocultas Simples com menos de 3 estados, enquanto a Estrutura Simples é um pouco mais eficiente quando consideradas as Cadeias Ocultas Hierárquicas. No entanto, em ambos os casos os resultados obtidos por cada estrutura convergem a medida que o número de estados aumenta.

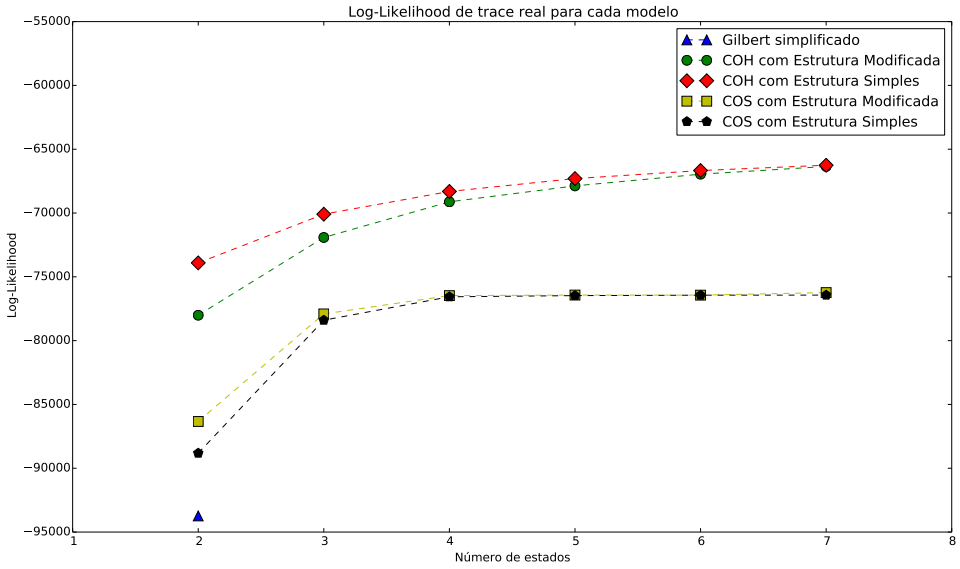


Figura 6.6: *Log-likelihood* de *trace* real para cada modelo

A próxima métrica considerada é a distribuição do tamanho da rajada de perda. Nenhum ganho de desempenho considerável foi observado com a utilização da Estrutura Modificada, então seus resultados serão omitidos. Para analisar o impacto do número de estados em cadeias de Markov ocultas é mostrada uma comparação entre os Q-Q plots obtidos pelos diferentes números de estados utilizados nas cadeias ocultas (Figuras 6.7 e 6.8). O aumento no número de estados na Cadeia Oculta Simples não altera de maneira visível os resultados do Q-Q plot para mais de 3 estados ocultos. Por outro lado, para Cadeias Ocultas Hierárquicas existe uma melhora visível nos resultados a medida que o número de estados aumenta. Baseado nestes resultados, considera-se que o número ótimo de estados para modelar a distribuição da rajada de perda é de 3 estados para a Cadeia Oculta Simples e 7 estados para a Cadeia Oculta Hierárquica.

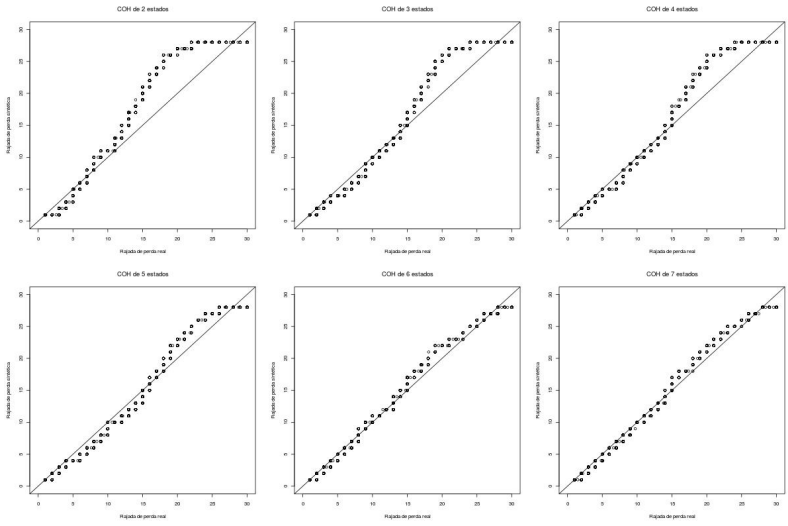


Figura 6.7: Comparação de Q-Q plots da rajada de perda de Cadeia Oculta Hierárquica variando entre 2 e 7 estados

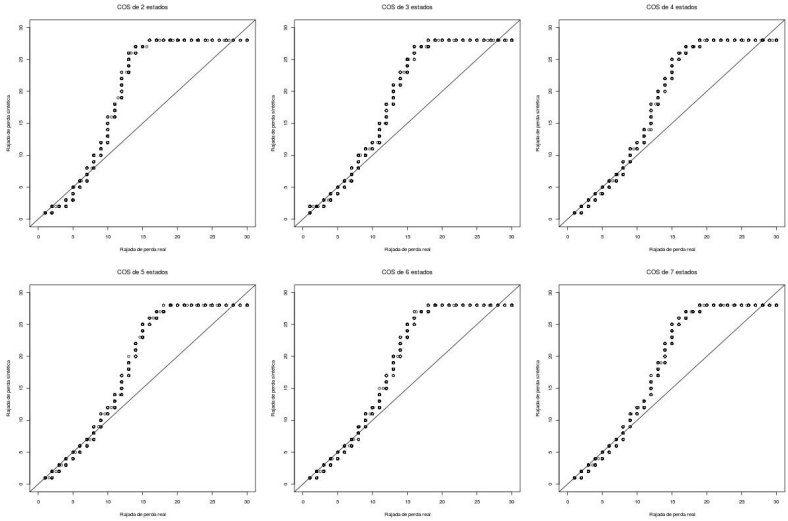


Figura 6.8: Comparação de Q-Q plots da rajada de perda de Cadeia Oculta Simples variando entre 2 e 7 estados

A Figura 6.9 mostra uma comparação entre os Q-Q plots obtidos a partir do modelo de Gilbert simplificado, da Cadeia Oculta Hierárquica de 7 estados com Estrutura Simples e da Cadeia Oculta Simples de 3 estados com Estrutura Simples. Quando observados os Q-Q plots a Cadeia Oculta Hierárquica exibe os melhores resultados, apresentando uma boa aproximação da distribuição real. A Cadeia Oculta Simples modela de maneira menos precisa a distribuição para valores de rajada de perda maiores que 10, enquanto o modelo de Gilbert simplificado exibe uma distribuição significativamente diferente da obtida pelas medições residenciais.

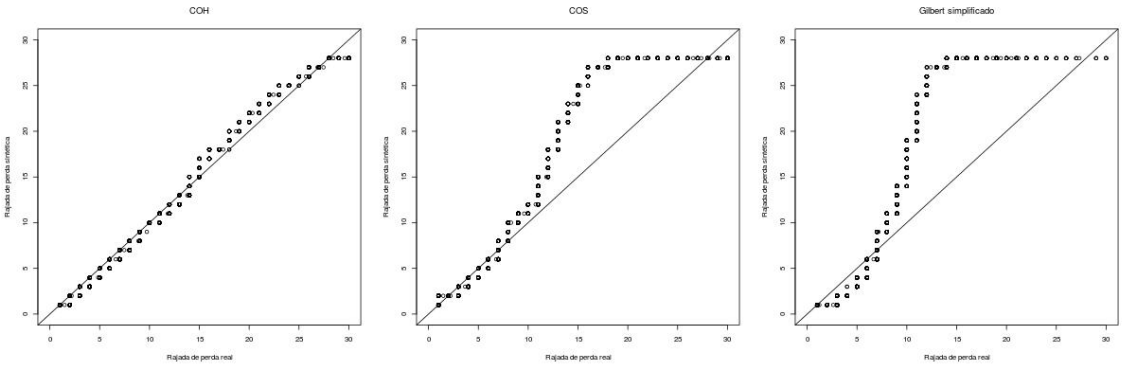
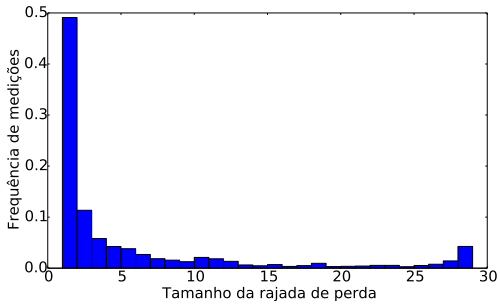


Figura 6.9: Comparação de Q-Q plots da rajada de perda de cada modelo

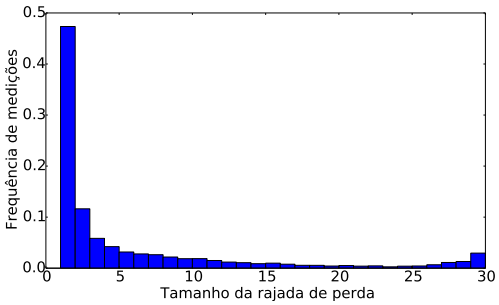
A comparação dos histogramas da distribuição da rajada de perda real e sintética, onde são considerados apenas perdas de tamanho maior do que 0, também indica que a Cadeia Oculta Hierárquica de 7 estados é o modelo mais preciso na modelagem desta métrica, como mostra a Figura 6.10. A Cadeia Oculta Simples de 3 estados ocultos também modela com razoável precisão a distribuição da rajada de perda, enquanto a cadeia de Gilbert simplificada não consegue modelar a rajada de perda de maneira satisfatória. A Tabela 6.1 mostra as distâncias entre as distribuições da rajada de perda real e sintética utilizando a distância de Jensen-Shannon. Os resultados comprovam que a COH é mais precisa na modelagem da rajada de perda e que o modelo de Gilbert simplificado possui a distribuição de rajada de perda mais distante da real.

Tabela 6.1: Distância de Jensen-Shannon para distribuição do tamanho de rajada de perda

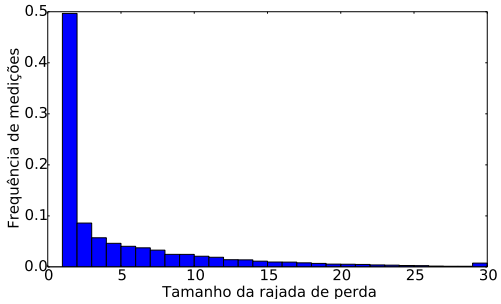
	Distância de Jensen-Shannon
COH	$1,7 \cdot 10^{-2}$
COS	$2,6 \cdot 10^{-2}$
Gilbert simplificado	$9,0 \cdot 10^{-2}$



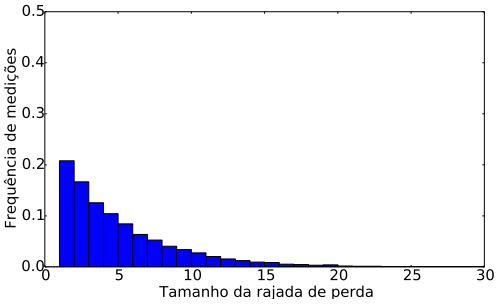
(a) Rajada de perda real



(b) Rajada de perda gerada por COH



(c) Rajada de perda gerado por COS



(d) Rajada de perda gerado por Gilbert

Figura 6.10: Comparação da distribuição do tamanho da rajada de perda real e sintético

Por fim, analisa-se quão precisos são os modelos quando se considera a distribuição do intervalo entre perdas gerado. A Estrutura Modificada mais uma vez não apresentou ganho significativo de desempenho e por isso seus resultados foram omitidos. Para avaliar o impacto do número de estados comparamos a distribuição do intervalo entre perdas real e sintético para diferentes números de estados das cadeias ocultas. Os resultados são mostrados nas Figuras 6.11 e 6.12. É possível observar que para um número de estados maior do que 3 não é visível uma melhoria significativa quando consideramos a Cadeia Oculta Simples, o mesmo acontecendo para Cadeias Ocultas Hierárquicas com mais de 5 estados. Assim, 3 e 5 estados são definidos como o número de estados ótimo para a COS e a COH respectivamente.

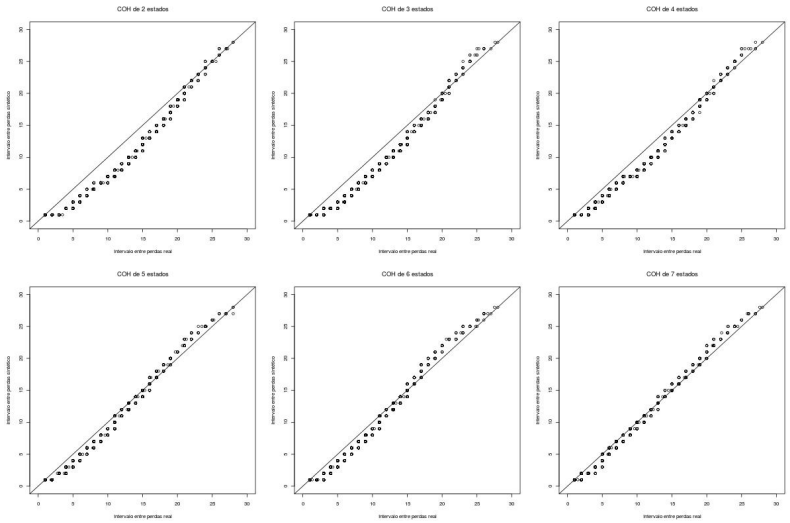


Figura 6.11: Comparação de Q-Q plots do intervalo entre perdas para COH variando entre 2 e 7 estados

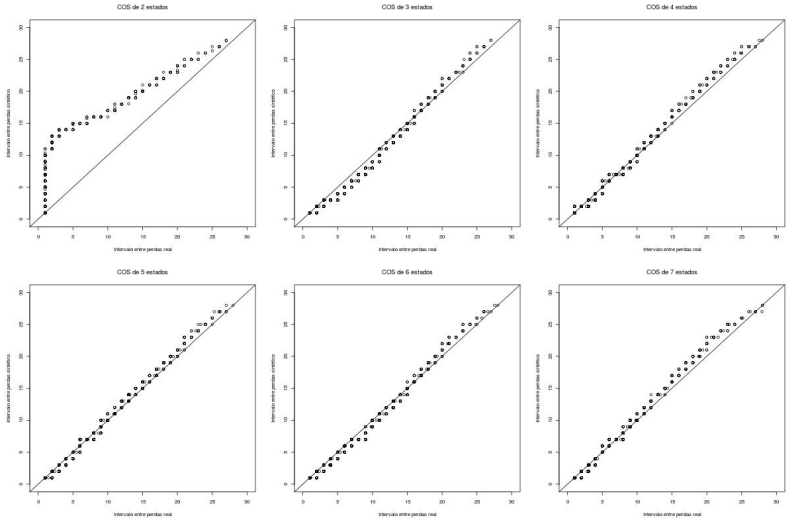


Figura 6.12: Comparação de Q-Q plots do intervalo entre perdas para COS variando entre 2 e 7 estados

A seguir é realizada a comparação entre Q-Q plots do intervalo entre perdas para os diferentes modelos considerados (Figura 6.13). Observa-se que para esta métrica não existe diferença visível entre a Cadeia Oculta Simples e a Cadeia Oculta Hierárquica quando considerados os Q-Q plots. O modelo de Gilbert simplificado, por outro lado, não parece capaz de modelar de maneira precisa a distribuição do intervalo entre as perdas medido nas redes de acesso residencial.

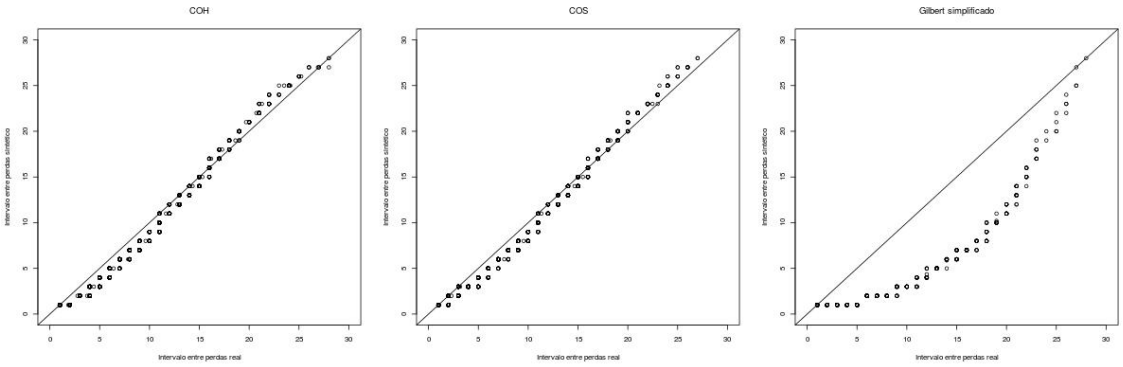


Figura 6.13: Comparação de Q-Q plots do intervalo entre perdas de cada modelo

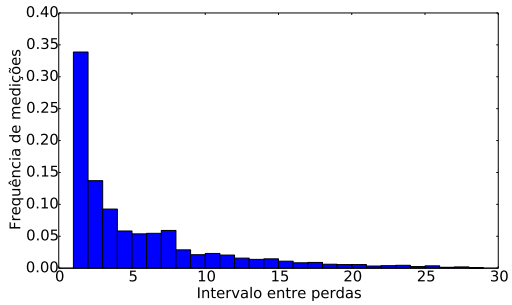
Ao analisar os histogramas gerados pelos dados reais e sintéticos, presentes na Figura 6.14, verificamos que a Cadeia Oculta Hierárquica gerada não captura com precisão a frequência de medições com intervalos entre perdas de tamanho 1 gerados pelo modelo (aproximadamente 0,2 para a COH e 0,35 para os dados reais). Esta diferença é menor quando observamos o resultado gerado pela Cadeia Oculta Simples. Por outro lado a precisão da COH é melhor se considerarmos a fração de intervalos de tamanho 2. No entanto, os dois modelos representam razoavelmente a distribuição do intervalo entre perdas para os exemplos considerados. A cadeia de Gilbert não modela de maneira satisfatória o intervalo entre perdas real. A Tabela 6.2 contém o valor da distância de Jensen-Shannon entre a distribuição real do intervalo entre perdas e a distribuição gerada por cada modelo. Os resultados comprovam que o modelo de Gilbert simplificado apresenta uma distribuição do intervalo entre perdas distante da real, enquanto as cadeias de Markov ocultas apresentam melhores distribuições, com o modelo de COS obtendo a menor distância em relação a distribuição real.

Tabela 6.2: Distância de Jensen-Shannon para distribuição do intervalo entre perdas

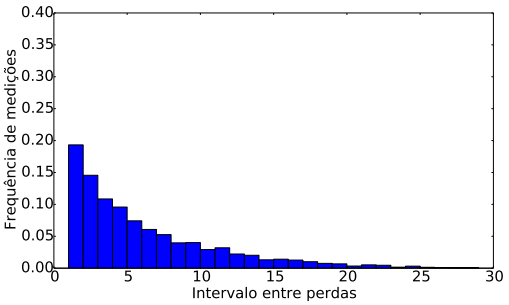
	Distância de Jensen-Shannon
COH	$1,8 \cdot 10^{-2}$
COS	$6,3 \cdot 10^{-3}$
Gilbert simplificado	$1,0 \cdot 10^{-1}$

6.3.1 Interpretação de modelos obtidos

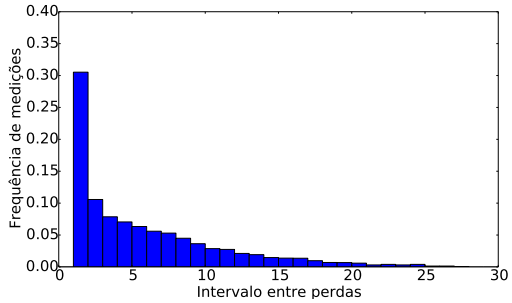
As Figuras 6.15 e 6.16 mostram as cadeias ocultas de 3 estados obtidas após o processo de treinamento utilizando a Estrutura Simples. A partir destes modelos é possível interpretar que cada estado oculto representa um estado da rede com um determinado nível de congestionamento.



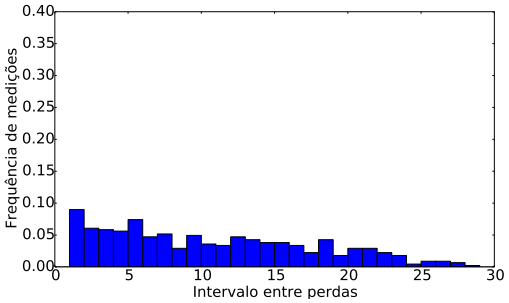
(a) Intervalo entre perdas real



(b) Intervalo entre perdas gerado por COH



(c) Intervalo entre perdas gerado por COS



(d) Intervalo entre perdas gerado por Gilbert

Figura 6.14: Comparação da distribuição do intervalo entre perdas real e sintético

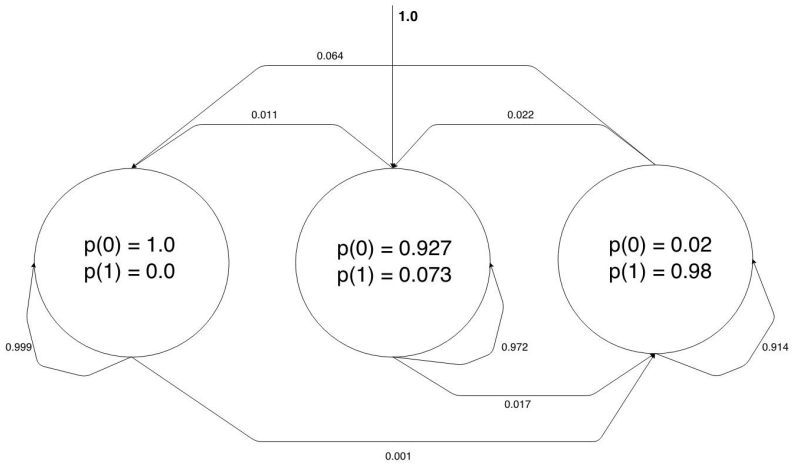


Figura 6.15: COS de 3 estados obtida após treinamento

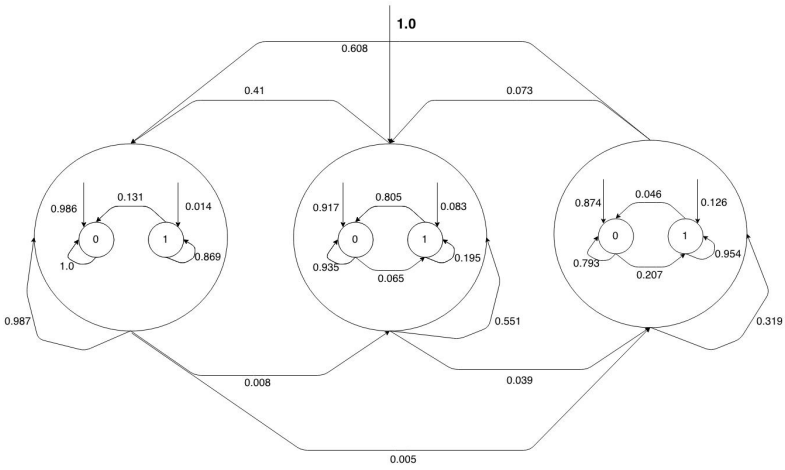


Figura 6.16: COH de 3 estados obtida após treinamento

Quando analisamos o modelo de COS percebemos que o primeiro estado oculto representa períodos em que a rede não possui congestionamento, pois todos os pacotes são transmitidos com sucesso. O segundo estado oculto corresponde a um período de baixo congestionamento, onde poucas perdas são observadas. Por fim, o terceiro estado oculto denota períodos em que a rede está congestionada e gera altas taxas de perda. Percebe-se também que após visitar um determinado estado oculto a rede tende a permancer neste estado.

A análise do modelo de COH também mostra a existência de 3 estados distintos da rede. A estrutura da COH possibilita também a análise do comportamento da rajada de perda. O primeiro estado oculto é interpretado como uma rede pouco congestionada em que poucas perdas ocorrem, uma vez que o estado que gera símbolos 0 é um estado absorvente. No entanto, existe uma baixa probabilidade de ocorrência de perdas em rajadas de tamanho grande no início da sequência gerada. O segundo estado oculto corresponde a rede em um período de congestionamento baixo com rajadas de perda de tamanho baixo. O terceiro estado oculto representa um período de alto congestionamento, no qual rajadas de tamanho grande são geradas. Segundo o modelo, a rede tende a permanecer em um estado com pouco congestionamento. É importante lembrar que cada visita a um estado oculto da COH gera 30 símbolos, diferentemente da COS, em que cada visita gera um único símbolo.

6.4 Conclusões sobre os resultados obtidos

Para todas as métricas consideradas neste capítulo, isto é, taxa de perda de pacotes, distribuição do tamanho de rajadas e distribuição do intervalo entre rajadas, o modelo de Markov oculto hierárquico mostrou ser o mais adequado para representar o processo de perda observado no experimento realizado. Outra conclusão relevante é que mais de 40% das perdas são provenientes de rajadas de tamanho igual ou

superior a 3.

É importante ressaltar que este experimento foi realizado em mais de 600 clientes reais da NET por aproximadamente 6 meses. Do nosso conhecimento nenhum experimento tão detalhado foi reportado na literatura.

Capítulo 7

Conclusões

Nosso trabalho é pioneiro na realização de um estudo do tráfego real gerado por um conjunto de clientes de uma grande operadora no Brasil e do processo de perda de pacotes UDP visto por um grande conjunto de clientes desta mesma operadora. Do nosso conhecimento nenhum trabalho desta natureza existe e só foi possível pela parceria concretizada através do projeto Inova Telecom entre COPPE, NET e TGR, empresa recém-incubada na COPPE.

A coleta de tráfego foi realizada a intervalos pequenos (1 segundo) em relação ao que é comumente coletado entre roteadores (5 minutos). Além disso, a coleta foi realizada em clientes residenciais, não sendo reportado na literatura medições de baixa granularidade neste contexto. Quanto ao processo de perdas, estudamos não somente a taxa de perda, mas o tamanho da rajada de perda e o intervalo entre tais rajadas.

Foram analisados modelos para o tráfego e para o processo de perda em redes de acesso residencial criados a partir de técnicas de aprendizado de máquina. Foram comparados 3 tipos de modelos para o tráfego residencial com granularidade de 1 segundo. Também foram avaliados 3 modelos de perda, analisando o desempenho obtido quando consideramos a distribuição da rajada de perda e do intervalo entre perdas.

Mostramos que o tráfego coletado é *bursty* em várias escalas de tempo e que sua autocorrelação tem decaimento lento. Do nosso conhecimento nenhum trabalho na literatura reporta este comportamento em redes residenciais. Para realizar a modelagem do tráfego é proposta a utilização de uma distribuição mista, composta da função degrau e de uma mistura de distribuições, que modela de maneira satisfatória na maior parte dos casos a distribuição do tráfego com granularidade de 1 segundo, mas que não captura a presença de rajadas de tráfego em diferentes escalas de tempo pois as amostras são independentes entre si. Mostramos que cadeias de Markov ocultas são mais precisas por capturarem dependências temporais. No caso da COS a autocorrelação do modelo decai rapidamente, enquanto o modelo de COH

com poucos símbolos conseguem modelar de maneira satisfatória a distribuição do tráfego e possuem a autocorrelação mais próxima da real dentre os modelos considerados. No entanto, estes modelos precisam ser aperfeiçoados para capturar melhor o decaimento da autocorrelação. Como trabalho futuro nos basearemos em resultados de [34, 45].

Em seguida, é mostrado que o processo de perda é bem modelado por cadeias de Markov ocultas. Os resultados mostram que as cadeias de Markov hierárquicas criadas em [12] são mais precisas na modelagem da rajada de perda. Nossos resultados mostram que 3 estados são suficientes para uma COS modelar o processo de perda. No entanto, percebemos um ganho de desempenho para um número de estados maior do que 3 quando consideramos a distribuição da rajada de perda utilizando uma cadeia de Markov oculta hierárquica. Mostramos também que o modelo de Gilbert simplificado não é suficientemente preciso para modelar o processo de perda.

Possíveis trabalhos futuros incluem uma análise detalhada do impacto da inicialização para os diversos algoritmos de aprendizado de máquina. Por fim, pretendemos modelar outras métricas coletadas em redes residenciais, como latência e capacidade, com o objetivo de criar modelos de redes residenciais.

Nosso trabalho é um primeiro estudo detalhado do desempenho de redes residenciais. Além dos estudos futuros mencionados acima, pretendemos criar um banco de dados com um grande número de clientes. Os modelos criados a partir das métricas servirão para: (a) a emulação da rede para determinar o impacto destes parâmetros na qualidade de experiência vista por clientes residenciais; (b) detecção de problemas de rede com os modelos a partir de coletas futuras de estatísticas; (c) determinação de classes de pontos de acesso com características semelhantes; (d) impacto do tráfego de clientes para estudos de planejamento de capacidade.

Referências Bibliográficas

[1] LEHR, W., BAUER, S., CLARK, D. D. "Measuring Performance when Broadband is the New PSTN", *Performance Evaluation*, v. 3, pp. 411–441, 2013.

[2] BAUER, S., CLARK, D. D., LEHR, W. "Understanding Broadband Speed Measurements". In: *38th Research Conference on Communication, Information and Internet Policy*, 2010.

[3] DE SOUZA E SILVA, E., FIGUEIREDO, D. R., LEÃO, R. M. M. "The TANGRAM-II Integrated Modeling Environment for Computer Systems and Networks". In: *ACM SIGMETRICS Performance Evaluation Review*, v. 36, pp. 64–69, 2009.

[4] OPENWRT. "OpenWrt". 2015. Disponível em: <`https://openwrt.org/`>. Acessado em 11/03/2015.

[5] LELAND, W. E., TAQQU, M. S., WILLINGER, W., et al. "On the self-similar nature of Ethernet traffic (extended version)", *IEEE/ACM Transactions on Networking*, v. 2, pp. 1–15, 1994.

[6] SUNDARESAN, S., DE DONATO, W., N.FEAMSTER, et al. "Broadband Internet Performance: A View From the Gateway". In: *Proceedings of the ACM SIGCOMM 2011*, 2011.

[7] KREIBICH, C., WEAVER, N., NECHAEV, B., et al. "Netalyzr: Illuminating the edge network". In: *IMC'10*, pp. 246–259, 2010.

[8] CANADI, I., BARFORD, P., SOMMERS, J. "Revisiting broadband performance". In: *IMC'12*, pp. 273–286, 2012.

[9] SAMKNOWS. "SamKnows". 2015. Disponível em: <`https://www.samknows.com/`>. Acessado em 23/04/2015.

[10] YAJNIK, M., MOON, S., KUROSE, J., et al. "Measurement and Modelling of the Temporal Dependence in Packet Loss". In: *1999 Proceedings IEEE INFOCOM*, pp. 345–352, 1999.

[11] SALAMATIAN, K., VATON, S. "Hidden Markov Modeling for network communication channels". In: *Proceedings of the ACM SIGMETRICS*, pp. 92–101, 2001.

[12] SILVEIRA, F., DE SOUZA E SILVA, E. "Predicting packet loss statistics with hidden Markov models for FEC control", *Computer Networks*, v. 56, pp. 628–641, 2012.

[13] ELLIS, M., PEZAROS, D. P., KYPRAIOS, T., et al. "A two-level Markov model for packet loss in UDP/IP-based real-time video applications targeting residential users", *Computer Networks*, v. 70, pp. 384–399, 2014.

[14] BARAKAT, C., THIRAN, P., IANNACCONE, G., et al. "Modeling Internet backbone traffic at the flow level", *IEEE Transactions on Signal Processing*, v. 51, pp. 2111–2124, 2003.

[15] REGGANI, A., SCHNEIDER, F., TEIXEIRA, R. "An end-host view on local traffic at home and work". In: *Passive and Active Measurement*, pp. 21–31, 2012.

[16] MAIER, G., FELDMANN, A., PAXSON, V., et al. "On dominant characteristics of residential broadband internet traffic". In: *IMC'09*, pp. 90–102, 2009.

[17] GROVER, S., PARK, M. S., SUNDARESAN, S., et al. "Peeking Behind the NAT: An Empirical Study of Home Networks". In: *IMC'13*, pp. 377–390, 2013.

[18] SIMPSON JR., C. R., REDDY, D., RILEY, G. F. "Empirical Models of End-User Network Behavior from NETI@home Data Analysis", *Simulation*, v. 84, pp. 557–571, 2008.

[19] KIHL, M., ÖDLING, P., LAGERSTEDT, C., et al. "Traffic analysis and characterization of Internet user behavior". In: *2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 224–231, 2010.

[20] AGOSTA, J. M., CHANDRASHEKAR, J., CROVELLA, M., et al. "Mixture Models of Endhost Network Traffic". In: *2013 Proceedings IEEE INFOCOM*, pp. 225–229, 2013.

[21] SIMPSON JR., C. R., RILEY, G. F. "NETI@home: A Distributed Approach to Collecting End-to-End Network Performance Measurements". In: *Passive and Active Measurement*, pp. 168–174, 2004.

[22] RABINER, L. R. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE*, pp. 257–286, 1989.

[23] DE SOUZA E SILVA, E., LEÃO, R. M. M., MUNTZ., R. R. "Performance Evaluation with Hidden Markov Models". In: *Performance Evaluation of Computer and Communication Systems. Milestones and Future Challenges*, pp. 112–128, 2011.

[24] FORNEY, G. D. "The viterbi algorithm". In: *Proceedings of the IEEE*, pp. 268–278, 1973.

[25] BAUM, L. E., PETRIE, T., SOULES, G., et al. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *The Annals of Mathematical Statistics*, v. 41, pp. 164–171, 1970.

[26] DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, v. 39, pp. 1–38, 1977.

[27] GILBERT, E. N. "Capacity of a Burst-Noise Channel", *Bell System Technical Journal*, v. 39, pp. 1253–1265, 1960.

[28] DE VIELMOND, C. C. L. B., LEÃO, R. M. M., DE SOUZA E SILVA, E. "Um modelo HMM hierárquico para usuários interativos acessando um servidor multimídia". In: *In: Simpósio Brasileiro de Redes de Computadores, 2007.*

[29] BISHOP, C. M. *Pattern Recognition and Machine Learning.* 1 ed. New York, Springer, 2006.

[30] MURPHY, K. P. *Machine Learning: A Probabilistic Perspective.* 1 ed. Massachusetts, MIT Press, 2012.

[31] FUGLEDE, B., TOPSOE, F. "Jensen-Shannon divergence and Hilbert space embedding". In: *International Symposium on Information Theory*, 2004.

[32] WOLFF, R. W. "Poisson Arrivals See Time Averages", *Operations Research*, v. 30, n. 2, pp. 223–231, 1982.

[33] SOMMERS, J., BARFORD, P., DUFFIELD, N., et al. "A Geometric Approach to Improving Active Packet Loss Measurement", *IEEE/ACM Transactions on Networking*, v. 16, pp. 307–320, 2008.

[34] DE LUCENA, S. C. *Modelos de Fluido para Tráfego Multimídia e Aplicações.* Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004.

[35] MACQUEEN, J. B. "Some Methods for Classification and Analysis of Multi-Variate Observations". In: *Procedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, pp. 281–297, 1967.

[36] FEUERVERGERL, A. "On some methods of analysis for weather experiments", .

[37] JONES, E., OLIPHANT, T., PETERSON, P. "SciPy". 2015. Disponível em: <http://www.scipy.org/>. Acessado em 11/03/2015.

[38] BENAGLIA, T., CHAUVEAU, D., HUNTER, D., et al. "mixtools: An R Package for Analyzing Finite Mixture Models", *Journal of Statistical Software*, v. 32, pp. 1–29, 2009.

[39] MACDONALD, P., DU, J. "mixdist". 2015. Disponível em: <http://cran.r-project.org/web/packages/mixdist/mixdist.pdf>. Acessado em 09/03/2015.

[40] LI, S., HWANG, C. L. "Queue response to input correlation functions: continuous spectral analysis", *IEEE/ACM Transactions on Networking*, v. 1, pp. 678–692, 1993.

[41] GUSELLA, R. "Characterizing the variability of arrival processes with indexes of dispersion", *IEEE Journal on Selected Areas in Communications*, v. 9, pp. 203–211, 1991.

[42] DA SILVA, A. P. C., VARELA, M., DE SOUZA E SILVA, E., et al. "Quality assessment of interactive voice applications", *Computer Networks*, v. 52, n. 6, pp. 1179–1192, 2008.

[43] FCC. "Raw Data - Measuring Broadband America 2013". 2015. Disponível em: <http://www.fcc.gov/measuring-broadband-america/2014/raw-data-fixed-2013>. Acessado em 02/03/2015.

[44] FCC. "2014 Measuring Broadband America Report - Technical Appendix". 2015. Disponível em: <http://data.fcc.gov/download/measuring-broadband-america/2014/Technical-Appendix-fixed-2014.pdf>. Acessado em 02/03/2015.

[45] ROBERT, S., BOUDEC, J. Y. L. "On a Markov Modulated Chain Exhibiting Self-similarities over Finite Timescale", *Performance Evaluation*, v. 27, pp. 159–173, 1996.

Apêndice A

Resultados de outros clientes

Neste apêndice são apresentados os resultados relacionados a tráfego obtidos pelos outros voluntários.

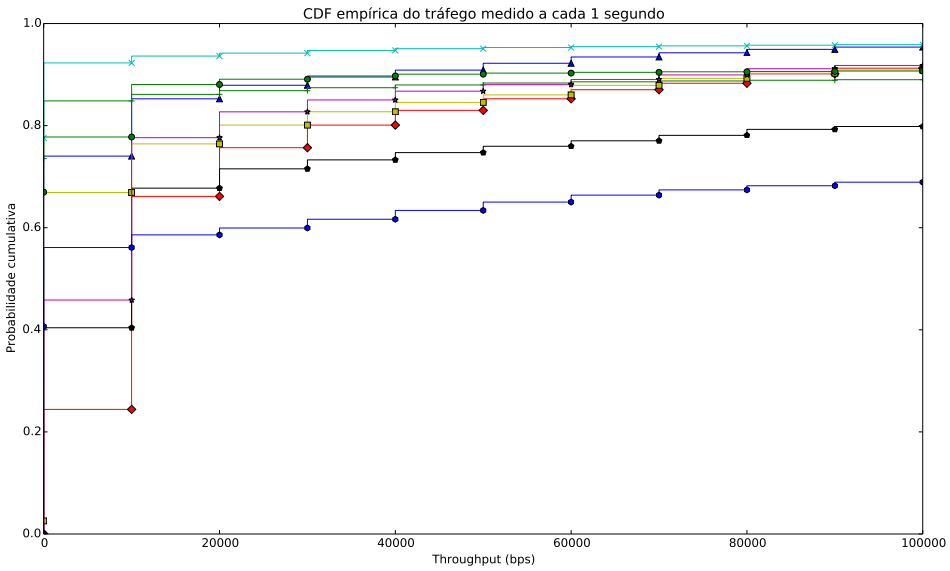


Figura A.1: CDF empírica do tráfego dos voluntários

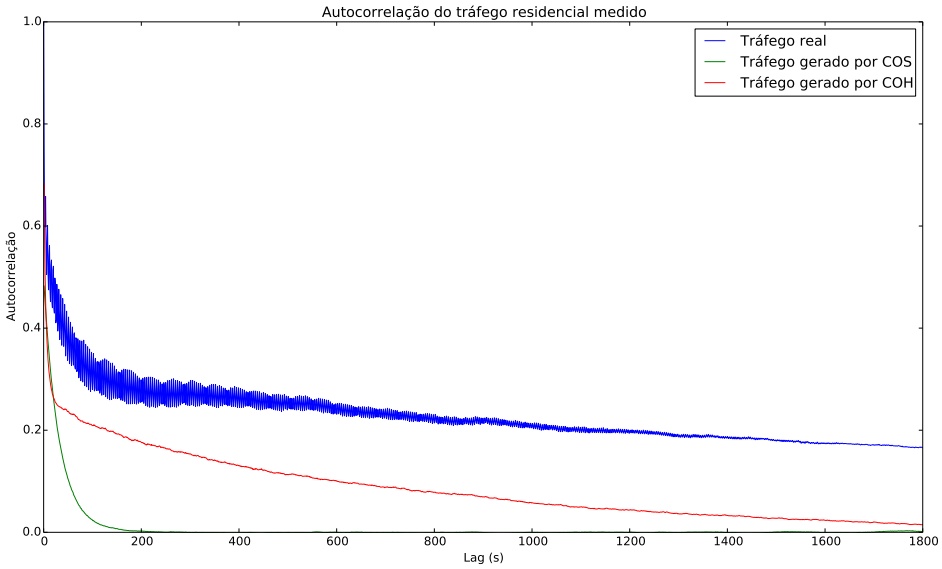


Figura A.2: Autocorrelação do tráfego residencial gerado pelo voluntário 1

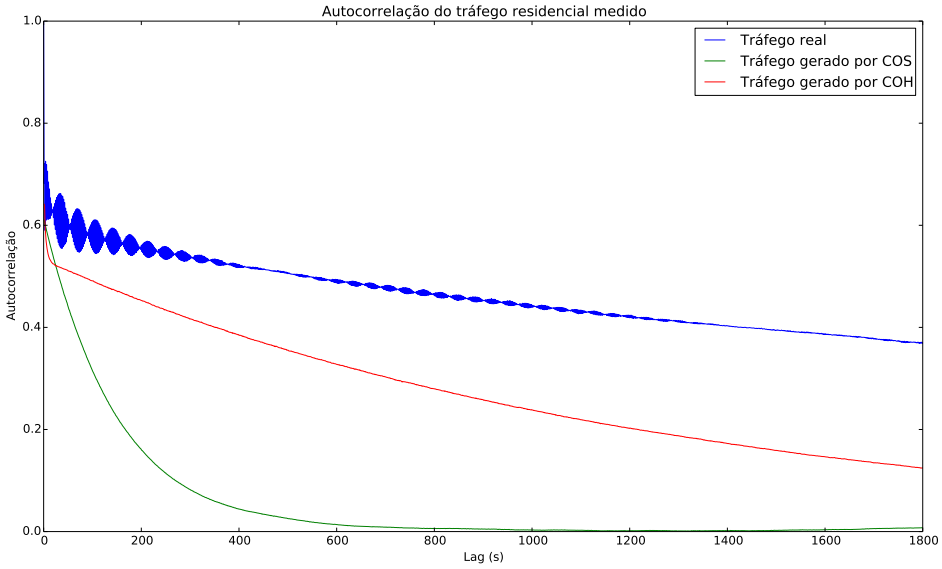


Figura A.3: Autocorrelação do tráfego residencial gerado pelo voluntário 2

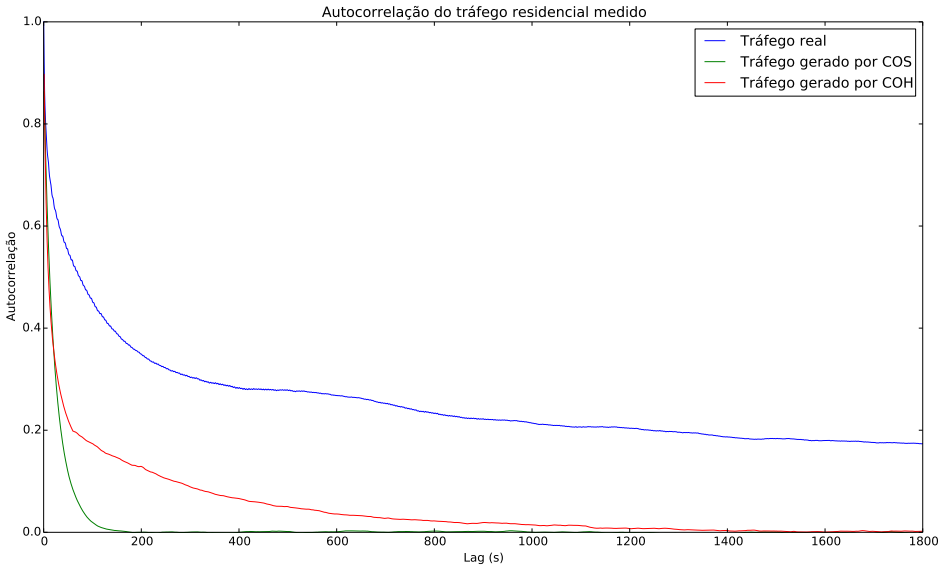


Figura A.4: Autocorrelação do tráfego residencial gerado pelo voluntário 3

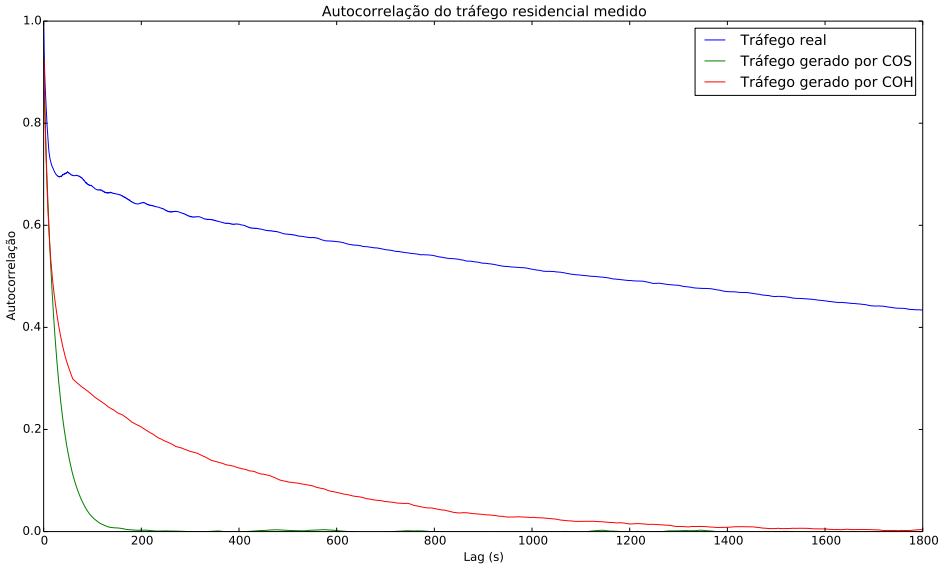


Figura A.5: Autocorrelação do tráfego residencial gerado pelo voluntário 4

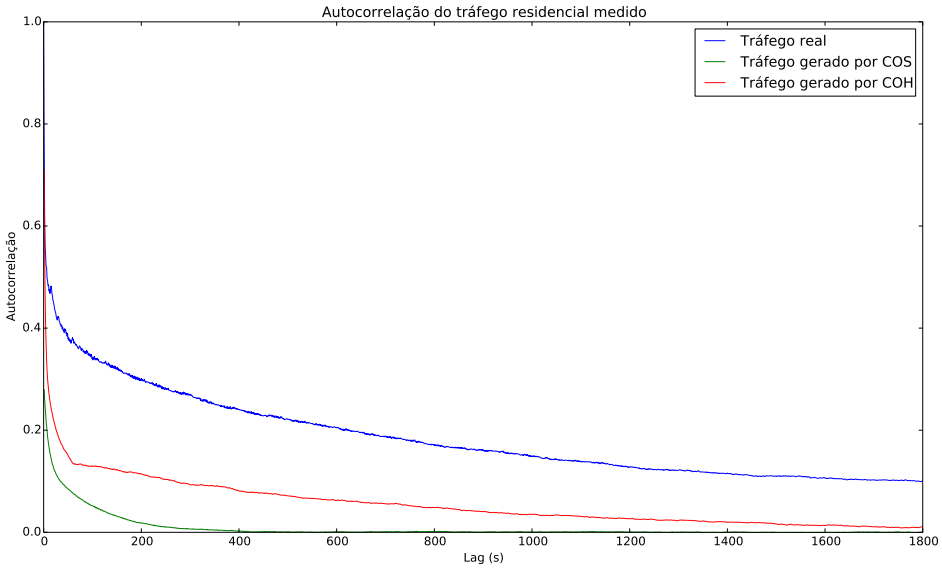


Figura A.6: Autocorrelação do tráfego residencial gerado pelo voluntário 5

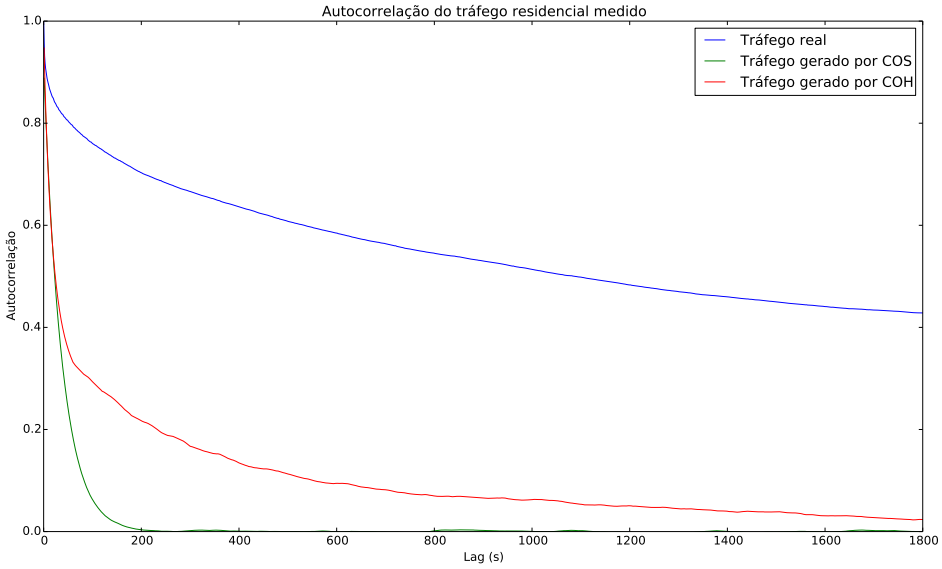


Figura A.7: Autocorrelação do tráfego residencial gerado pelo voluntário 6

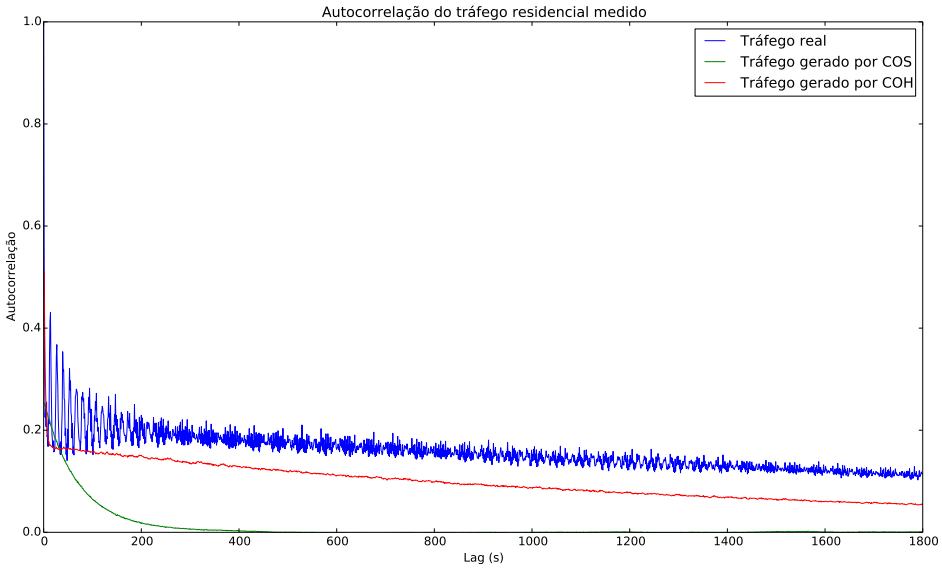


Figura A.8: Autocorrelação do tráfego residencial gerado pelo voluntário 7

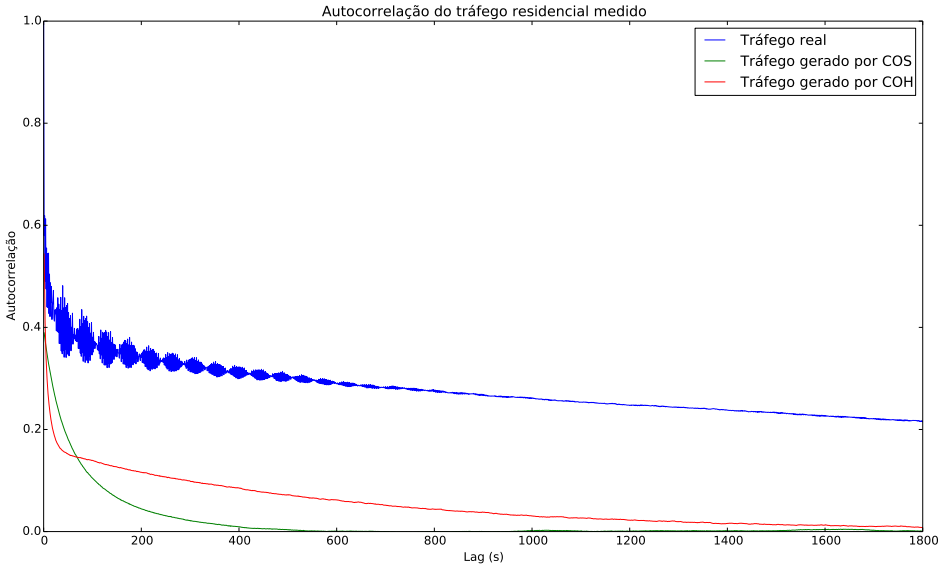


Figura A.9: Autocorrelação do tráfego residencial gerado pelo voluntário 8

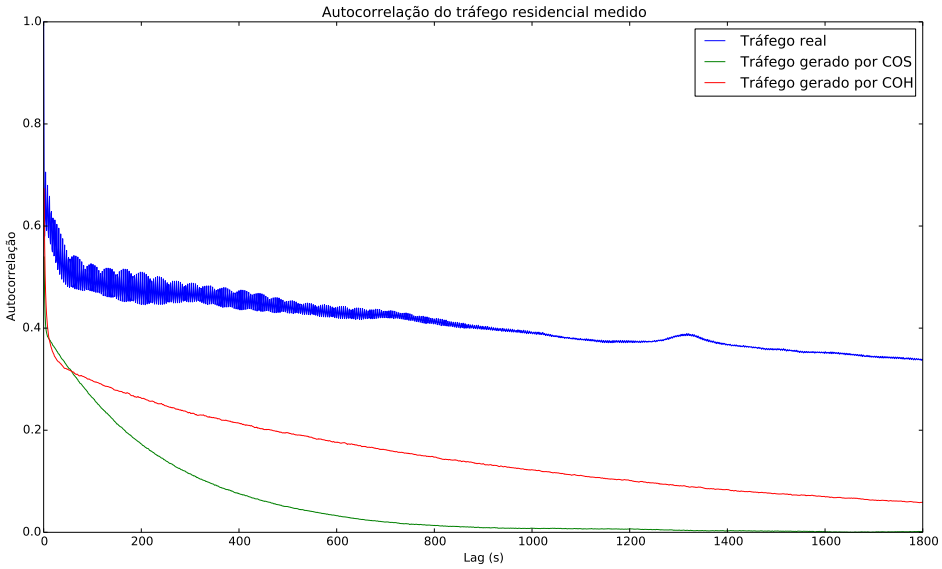


Figura A.10: Autocorrelação do tráfego residencial gerado pelo voluntário 9

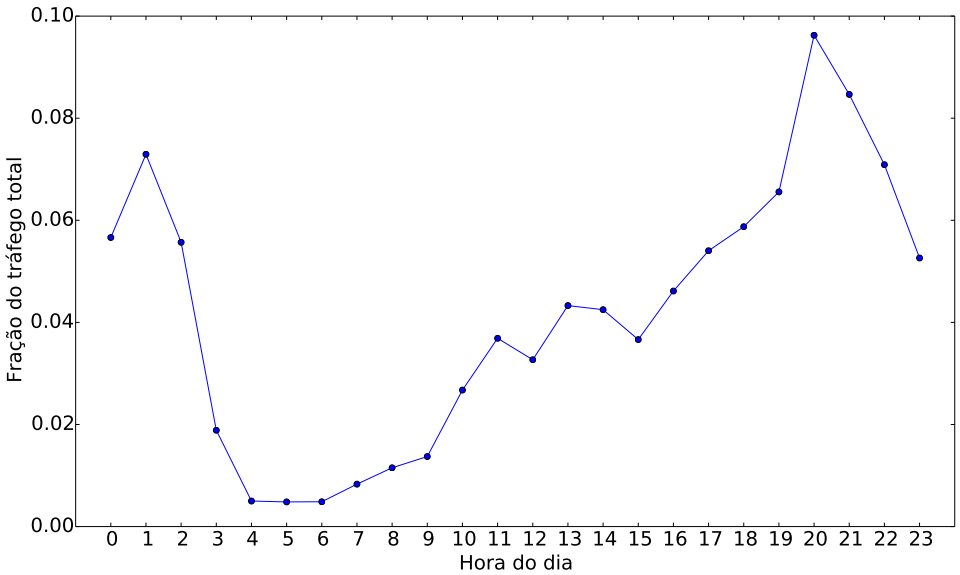


Figura A.11: Fração do tráfego por hora para voluntário 1

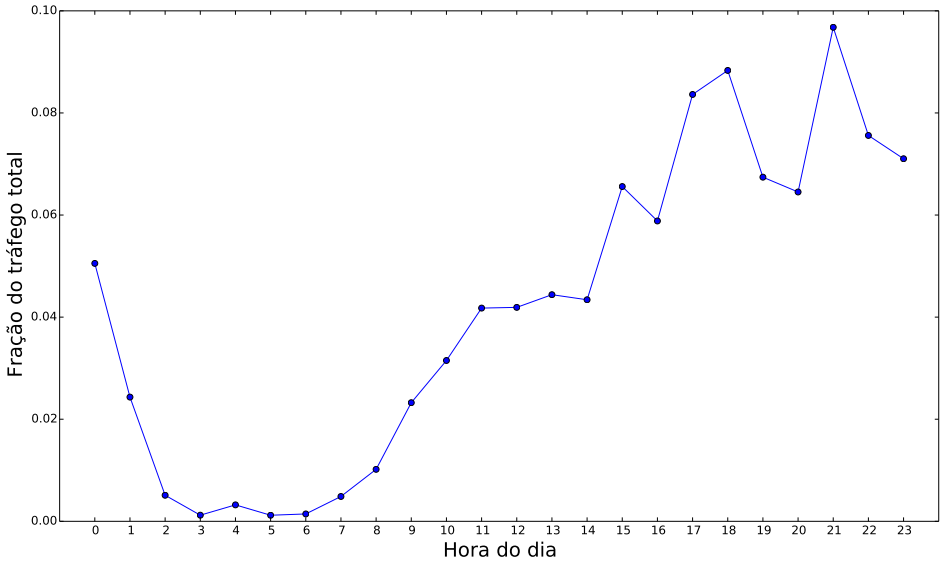


Figura A.12: Fração do tráfego por hora para voluntário 2

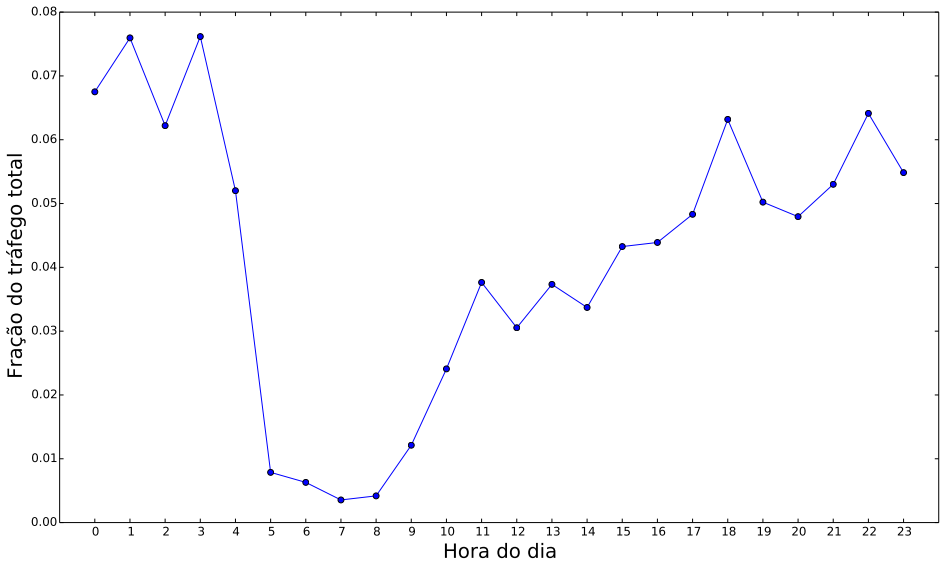


Figura A.13: Fração do tráfego por hora para voluntário 3

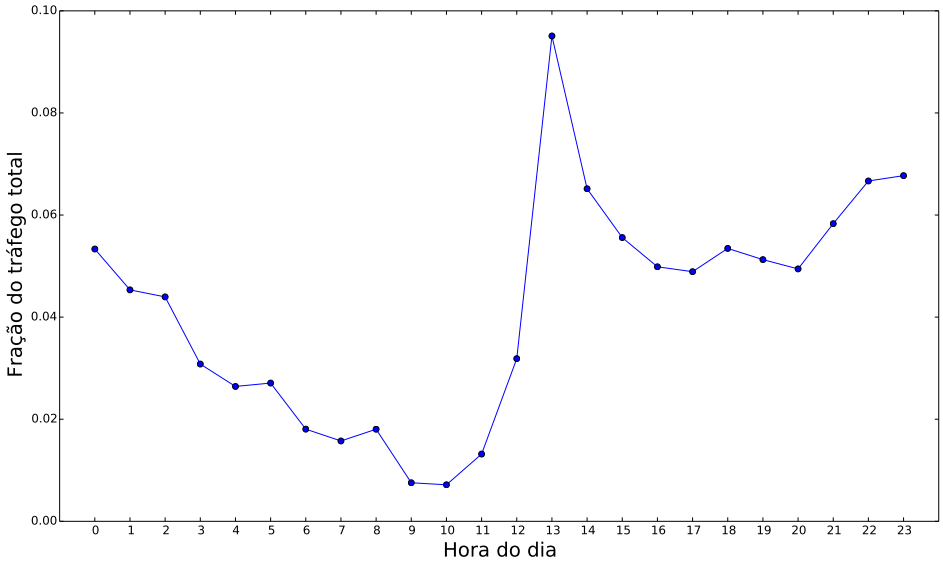


Figura A.14: Fração do tráfego por hora para voluntário 4

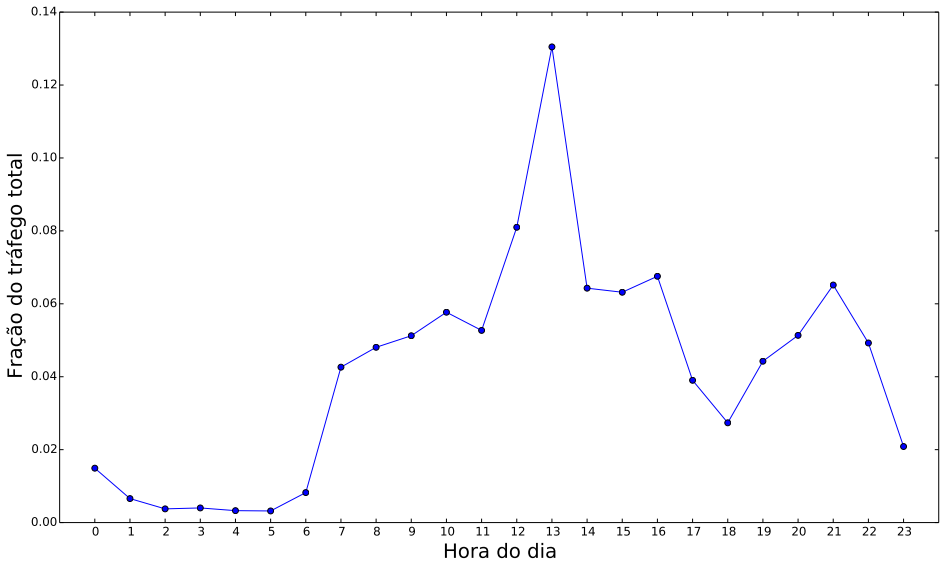


Figura A.15: Fração do tráfego por hora para voluntário 5

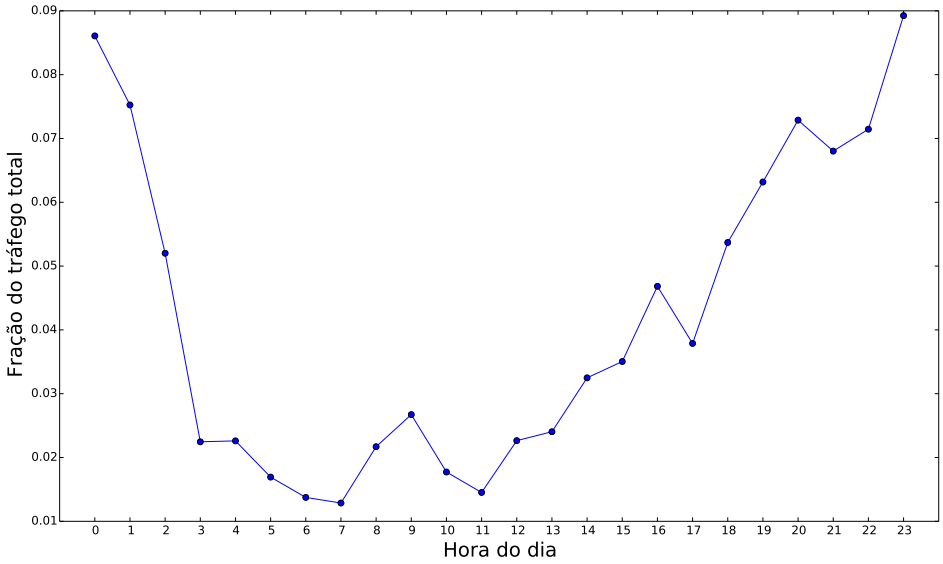


Figura A.16: Fração do tráfego por hora para voluntário 6

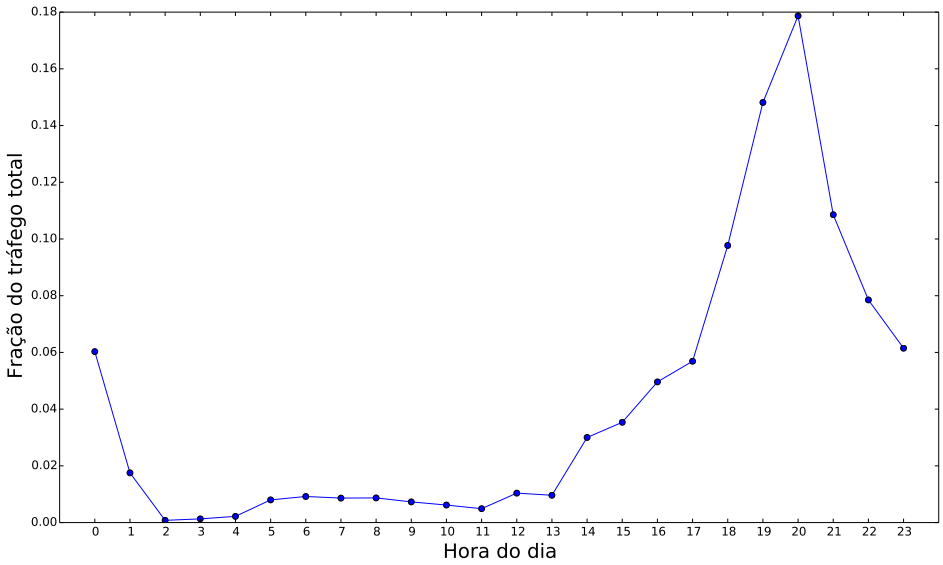


Figura A.17: Fração do tráfego por hora para voluntário 7

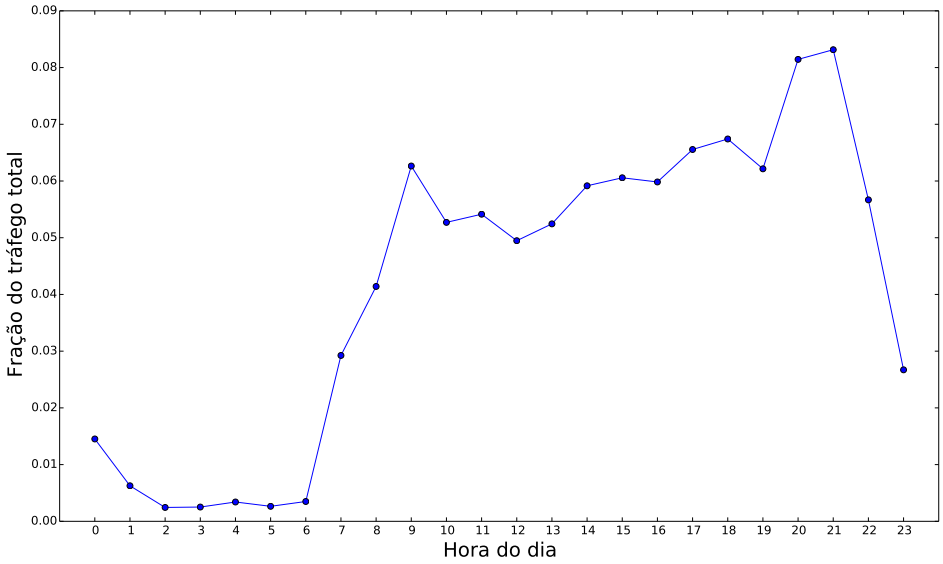


Figura A.18: Fração do tráfego por hora para voluntário 8

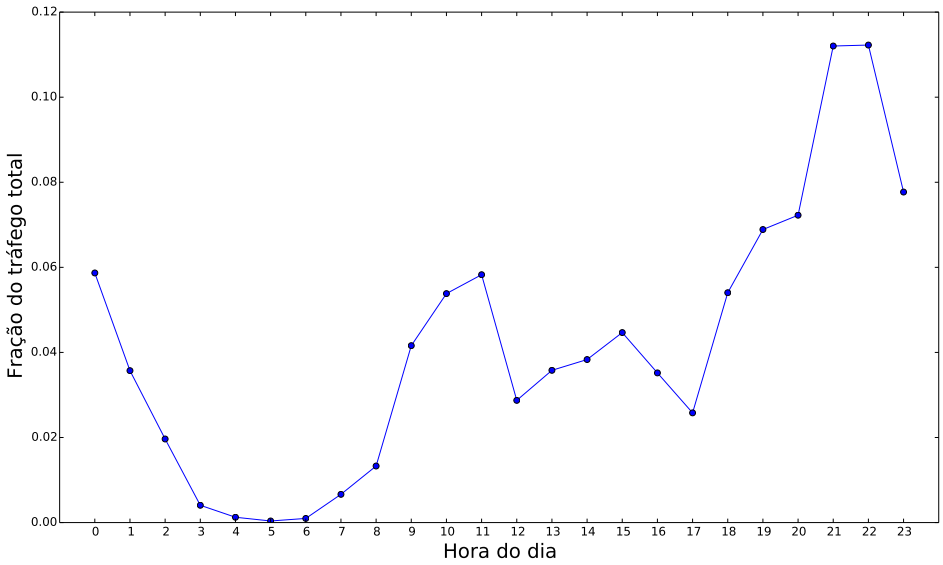


Figura A.19: Fração do tráfego por hora para voluntário 9

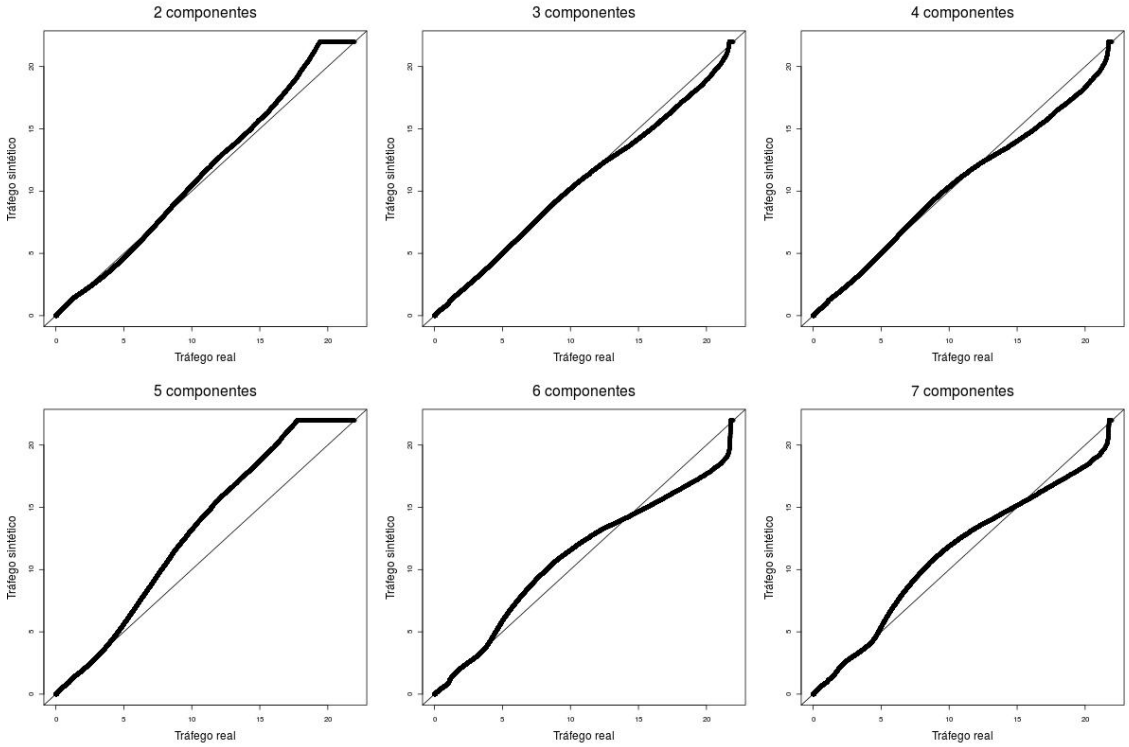


Figura A.20: Comparação entre misturas de Weibull e tráfego real do voluntário 1

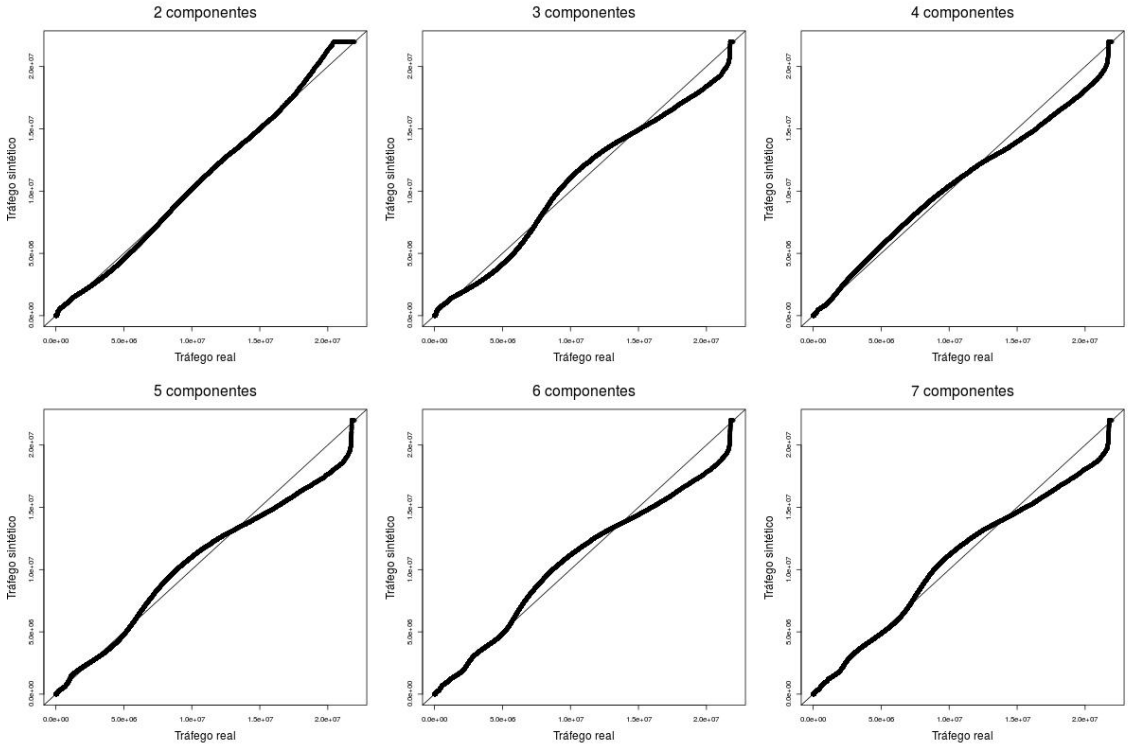


Figura A.21: Comparação entre misturas de Gama e tráfego real do voluntário 1

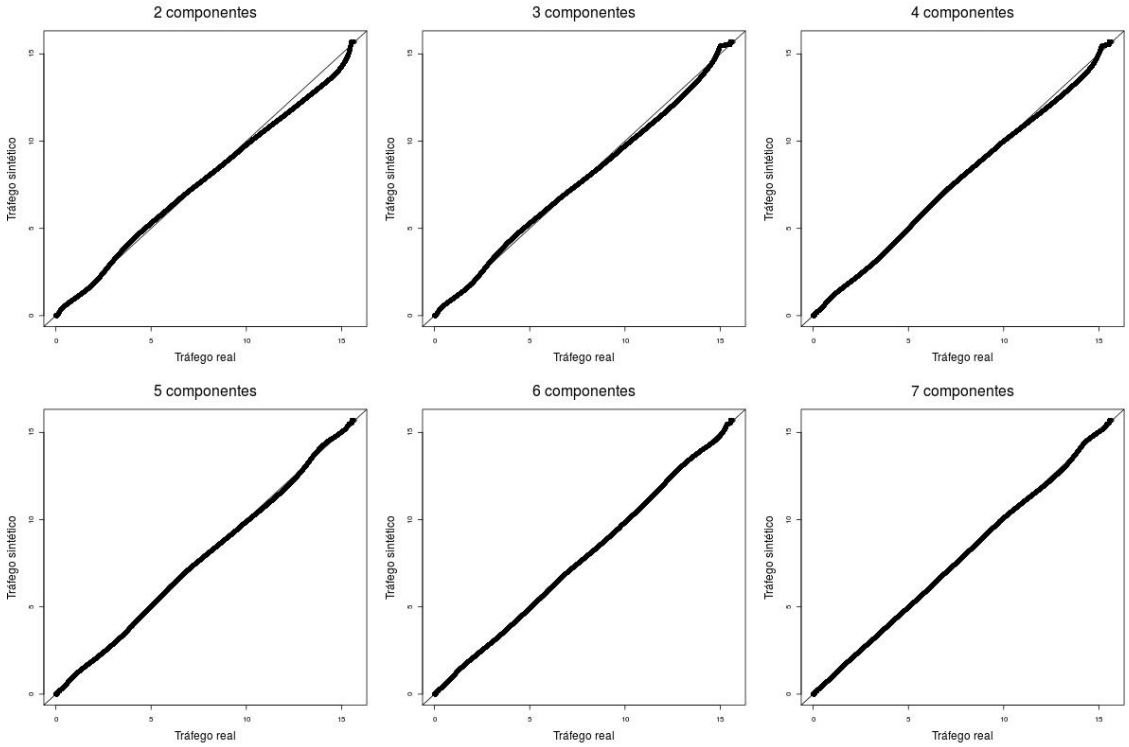


Figura A.22: Comparação entre misturas de Weibull e tráfego real do voluntário 2

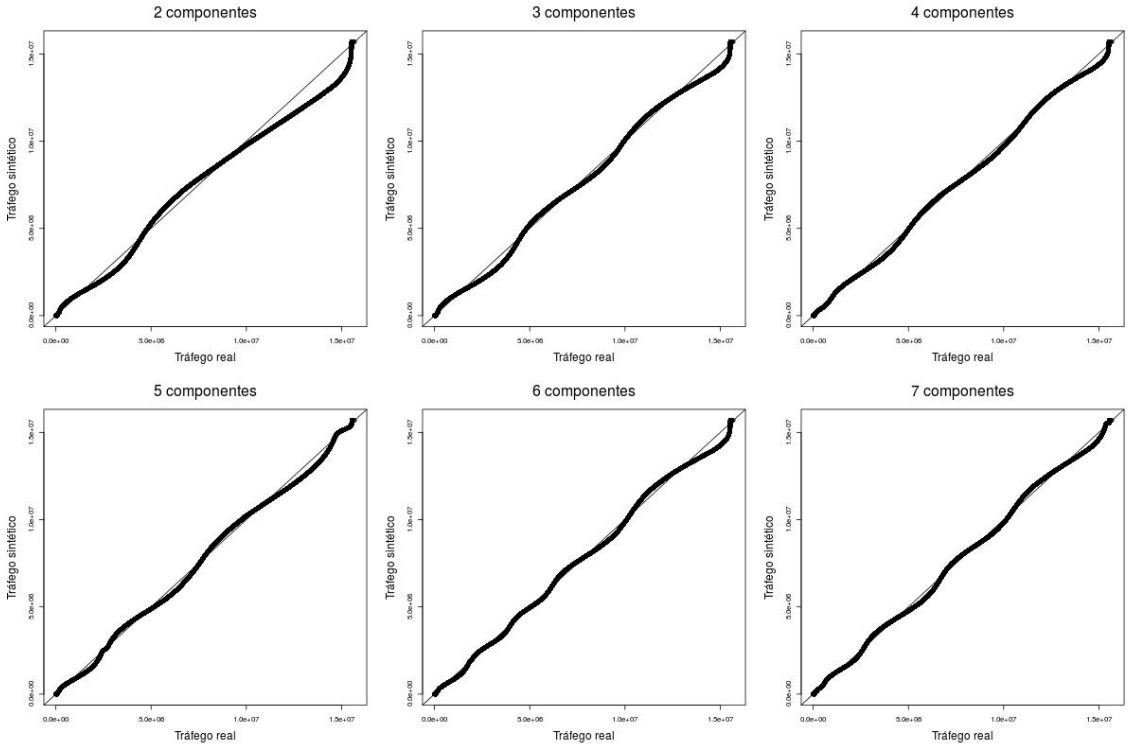


Figura A.23: Comparação entre misturas de Gama e tráfego real do voluntário 2

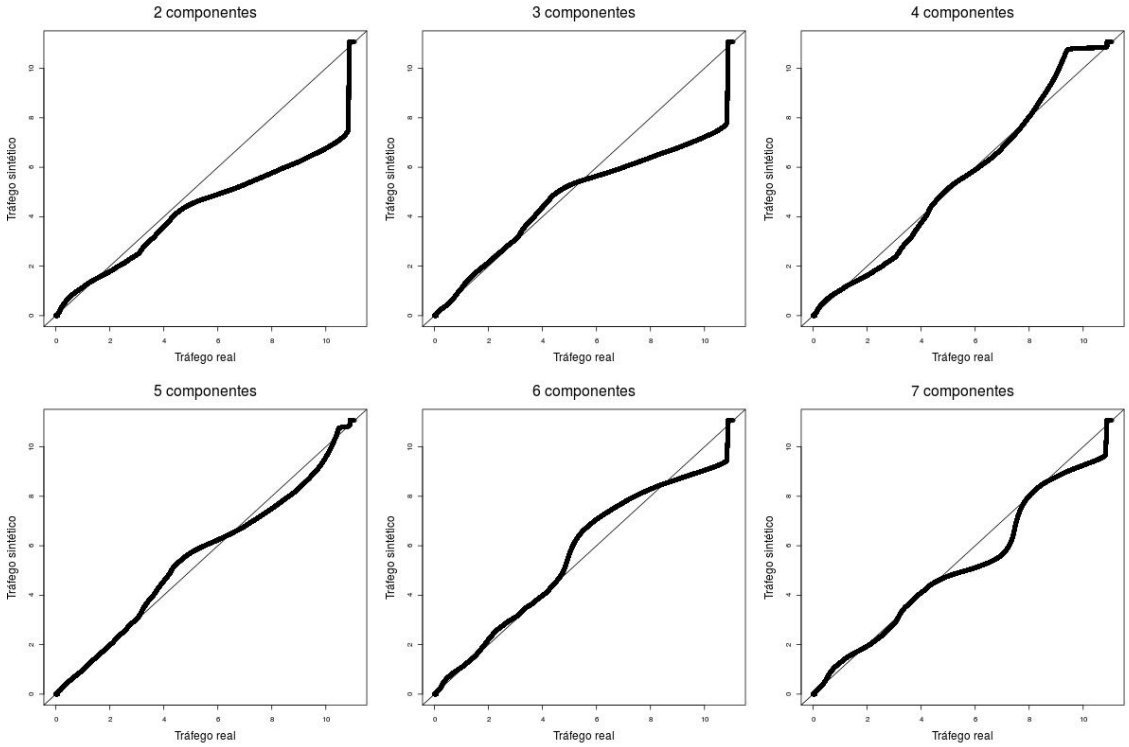


Figura A.24: Comparação entre misturas de Weibull e tráfego real do voluntário 3

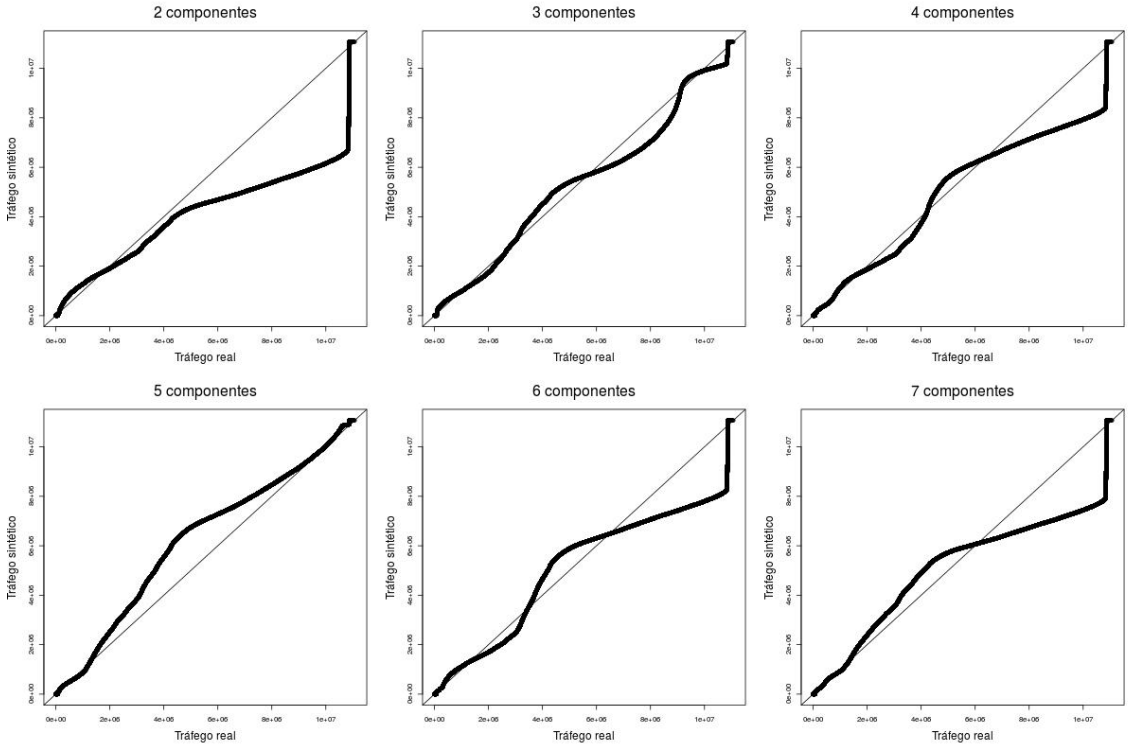


Figura A.25: Comparação entre misturas de Gama e tráfego real do voluntário 3

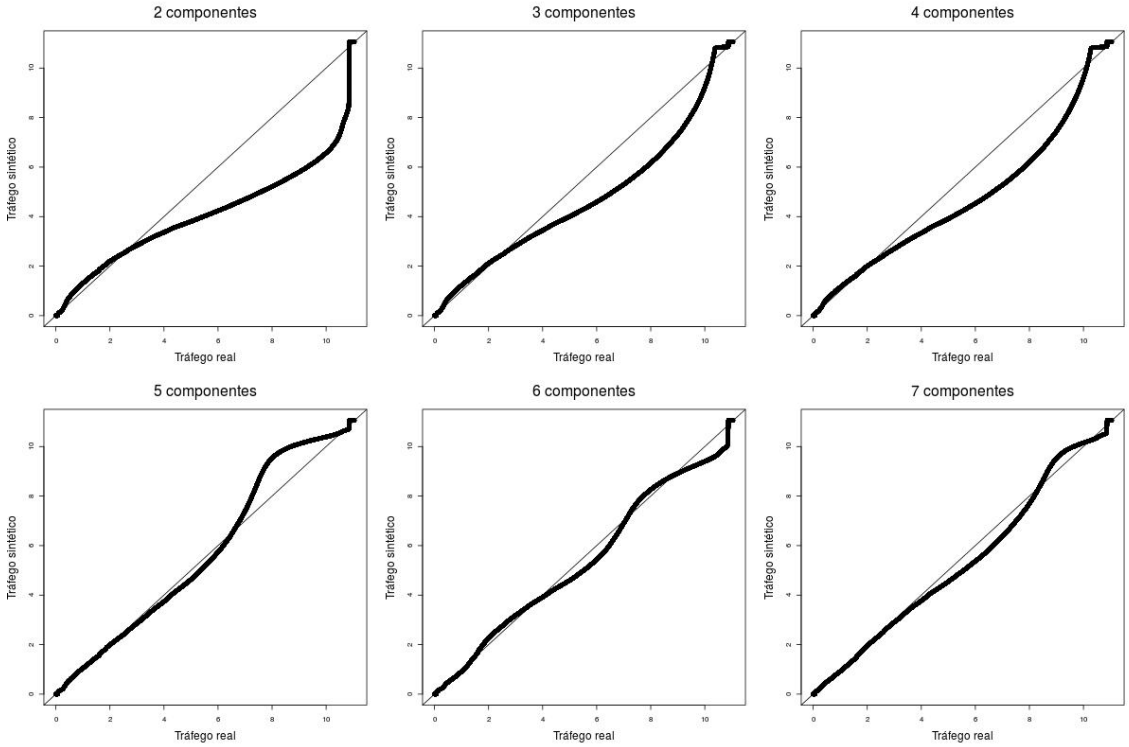


Figura A.26: Comparação entre misturas de Weibull e tráfego real do voluntário 4

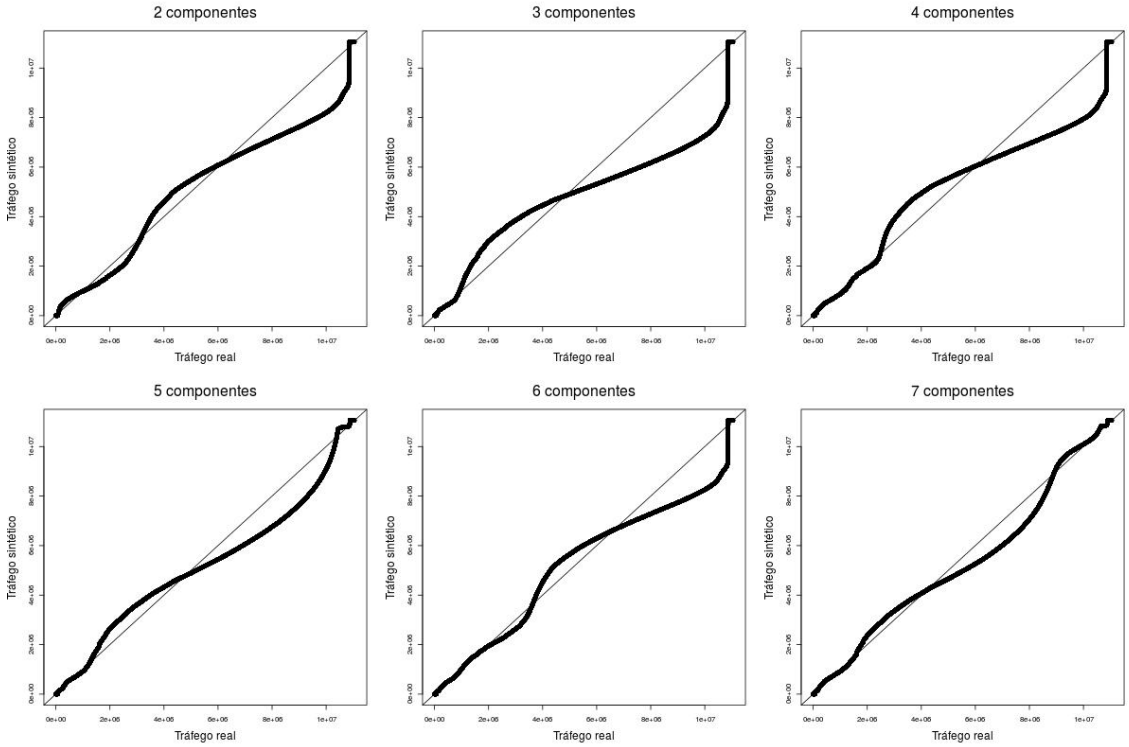


Figura A.27: Comparação entre misturas de Gama e tráfego real do voluntário 4

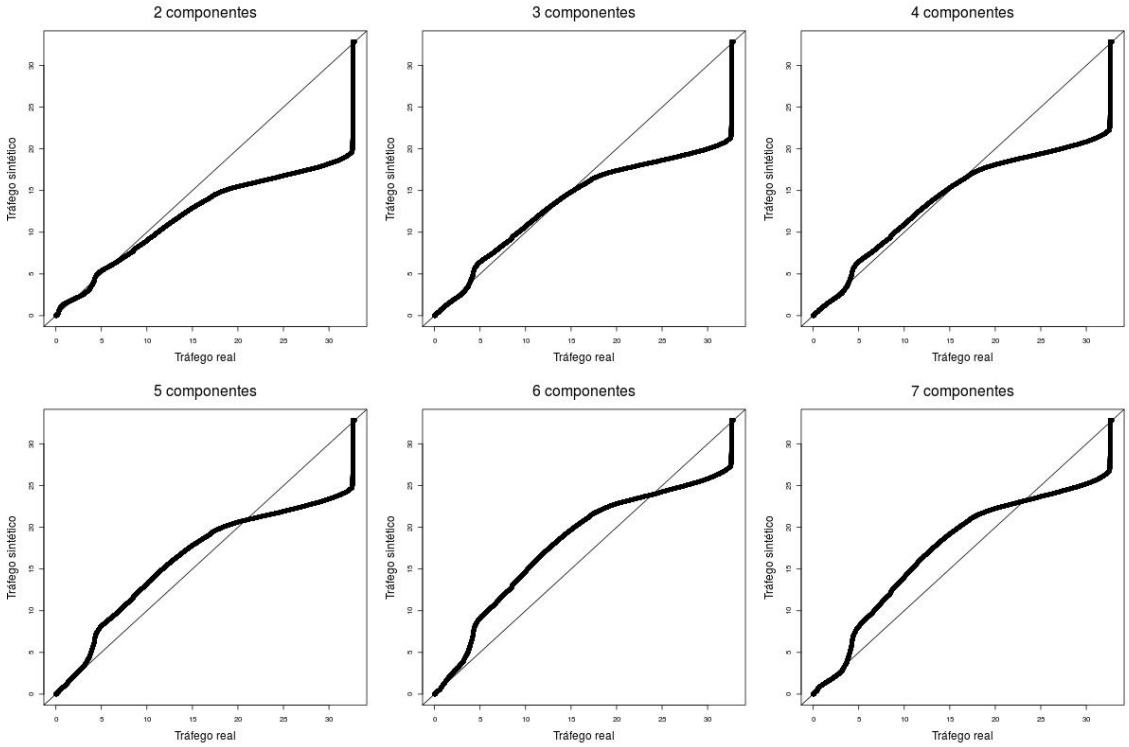


Figura A.28: Comparação entre misturas de Weibull e tráfego real do voluntário 5

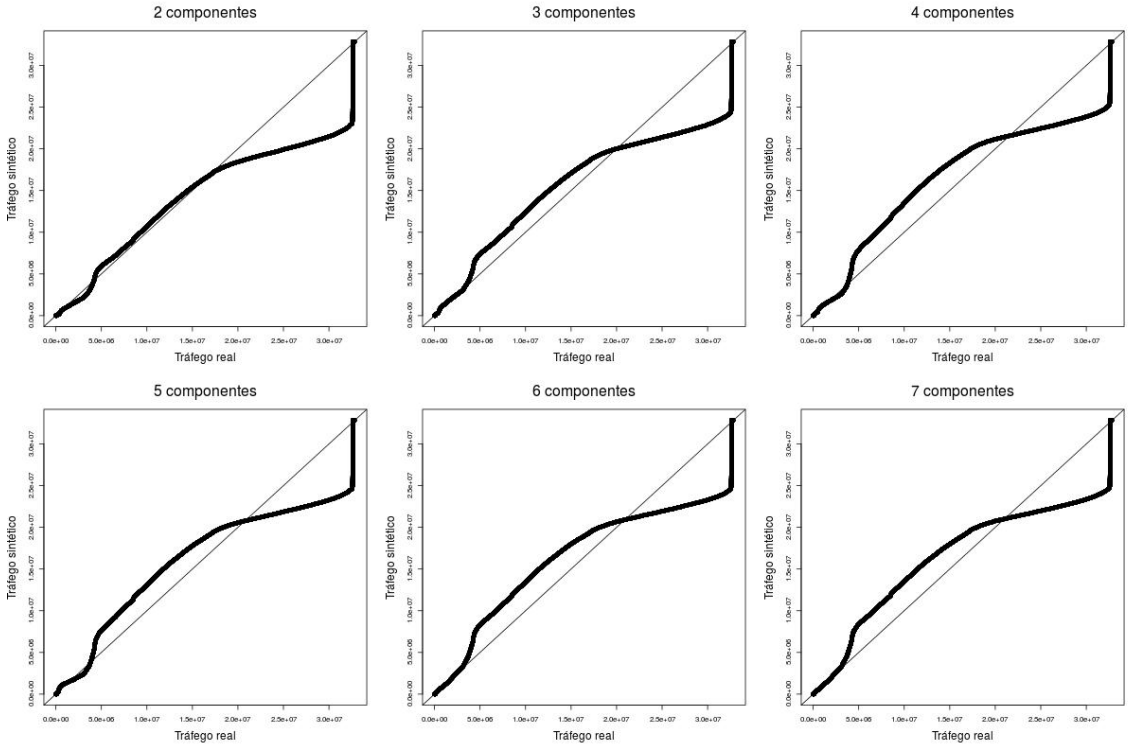


Figura A.29: Comparação entre misturas de Gama e tráfego real do voluntário 5

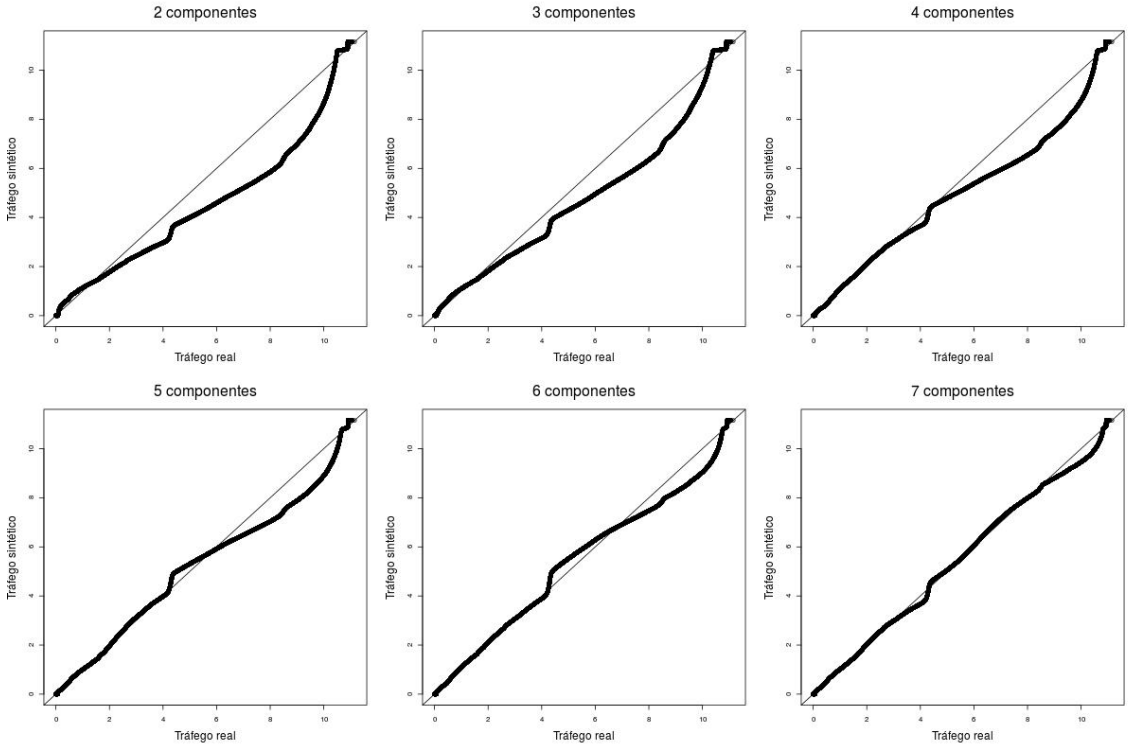


Figura A.30: Comparação entre misturas de Weibull e tráfego real do voluntário 6

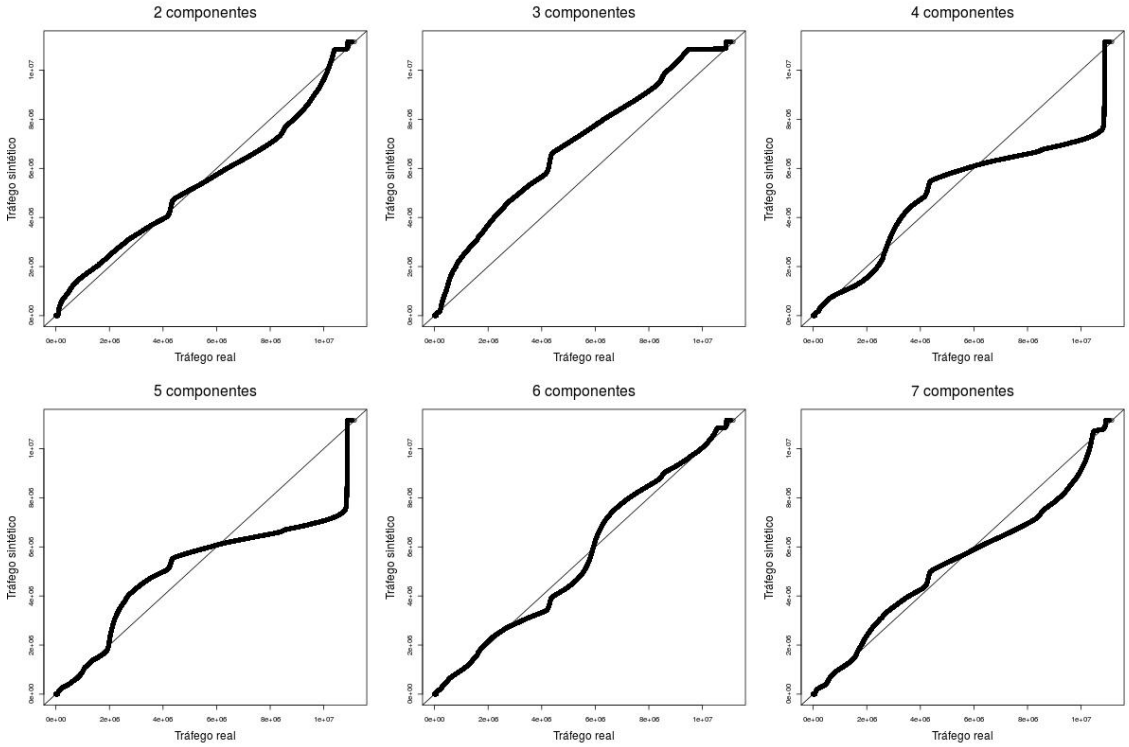


Figura A.31: Comparação entre misturas de Gama e tráfego real do voluntário 6

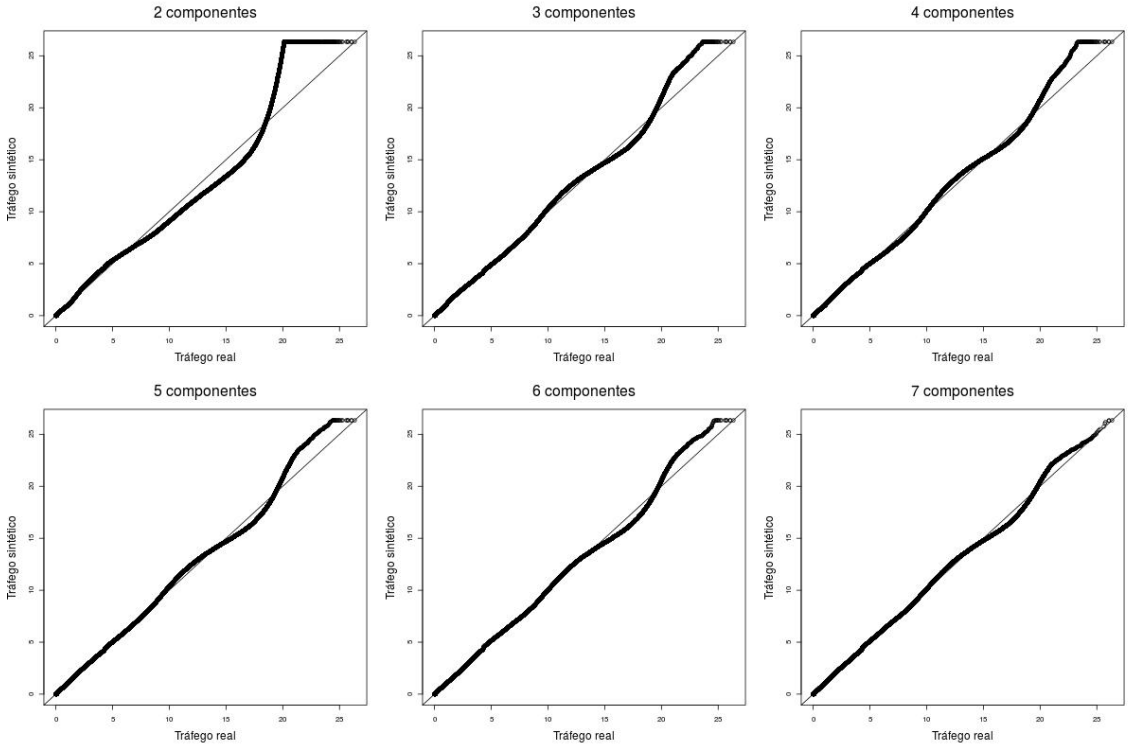


Figura A.32: Comparação entre misturas de Weibull e tráfego real do voluntário 7

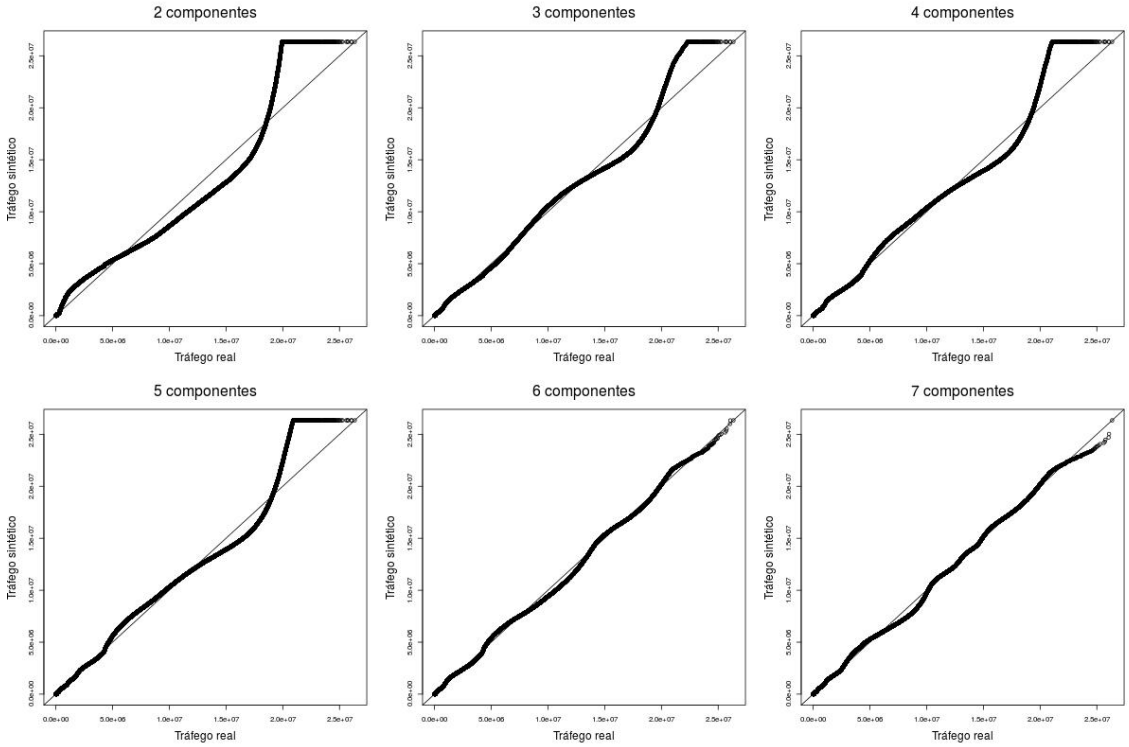


Figura A.33: Comparação entre misturas de Gama e tráfego real do voluntário 7

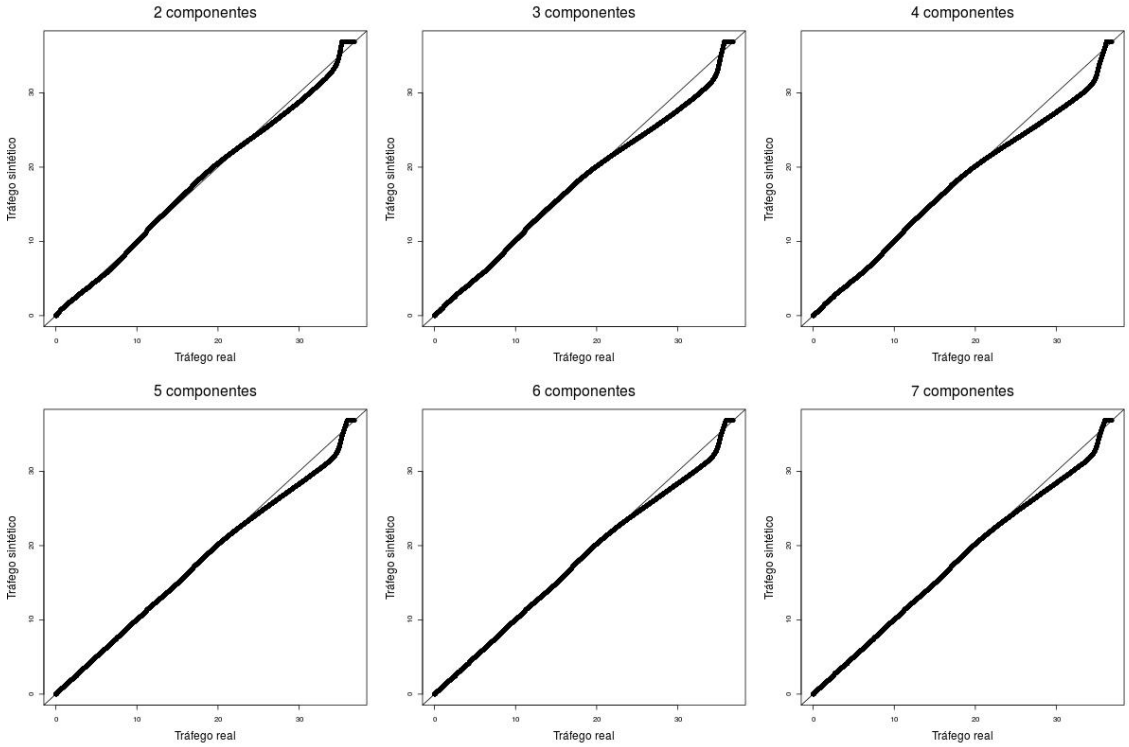


Figura A.34: Comparação entre misturas de Weibull e tráfego real do voluntário 8

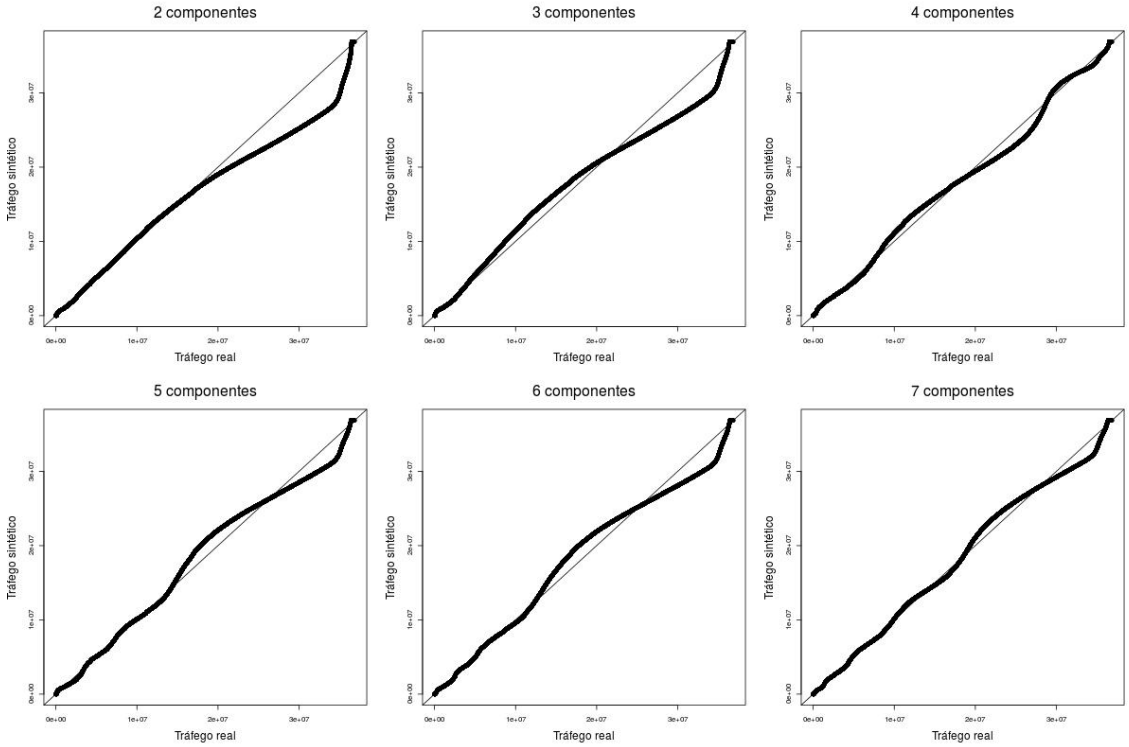


Figura A.35: Comparação entre misturas de Gama e tráfego real do voluntário 8

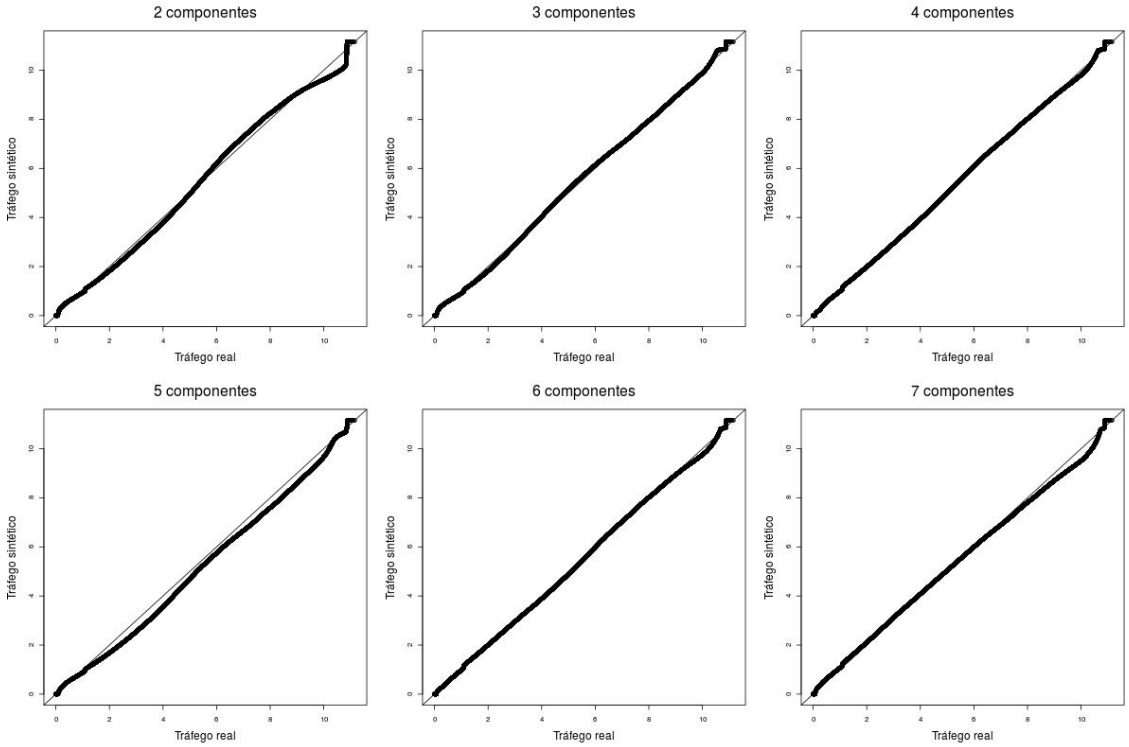


Figura A.36: Comparação entre misturas de Weibull e tráfego real do voluntário 9

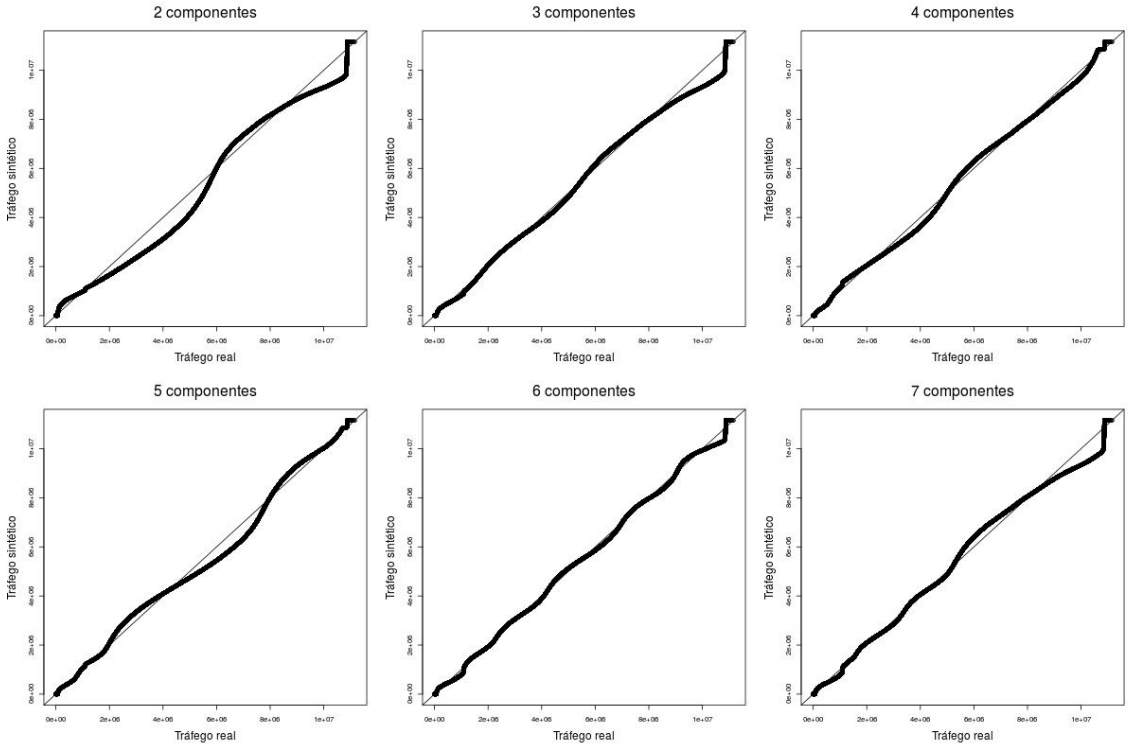


Figura A.37: Comparação entre misturas de Gama e tráfego real do voluntário 9

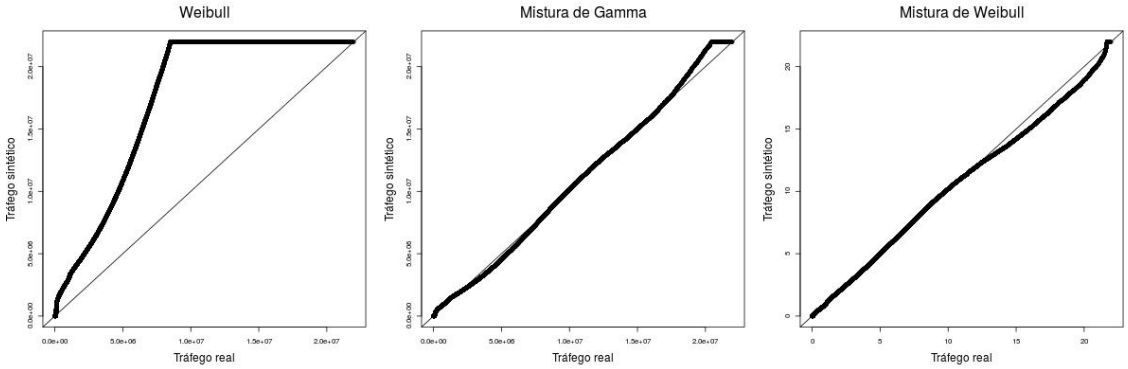


Figura A.38: Comparação entre distribuições mistas e tráfego real gerado por voluntário 1

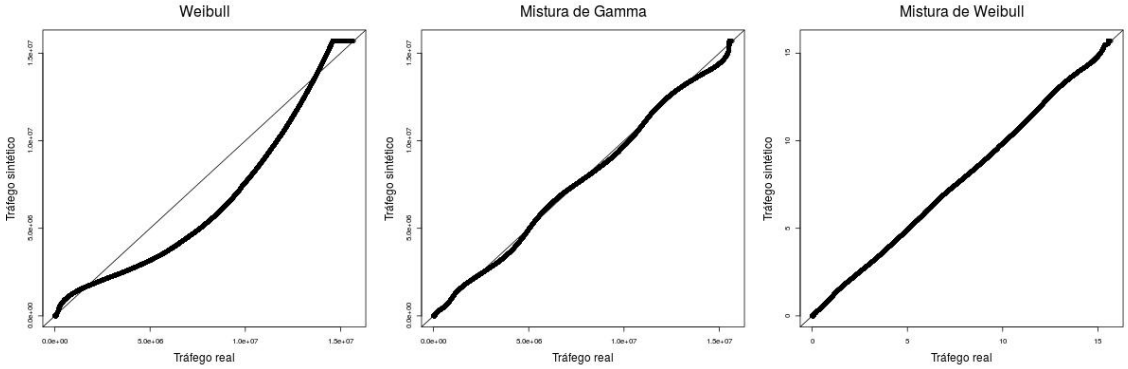


Figura A.39: Comparação entre distribuições mistas e tráfego real gerado por voluntário 2

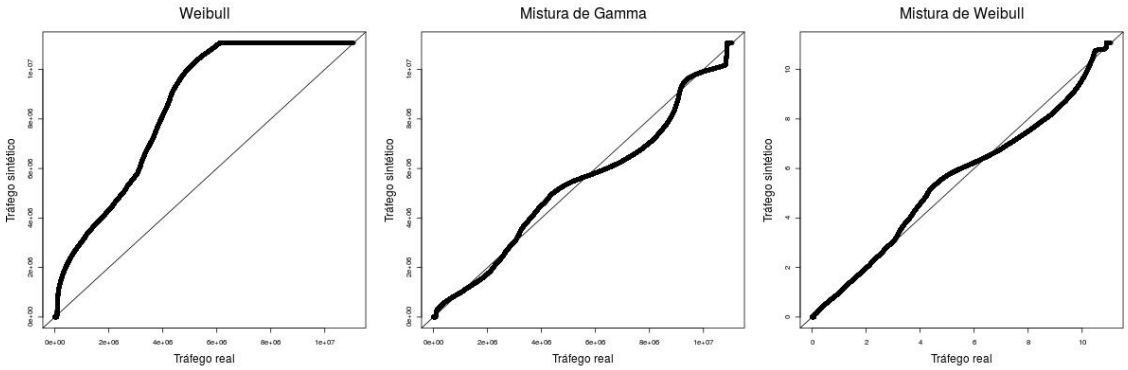


Figura A.40: Comparação entre distribuições mistas e tráfego real gerado por voluntário 3

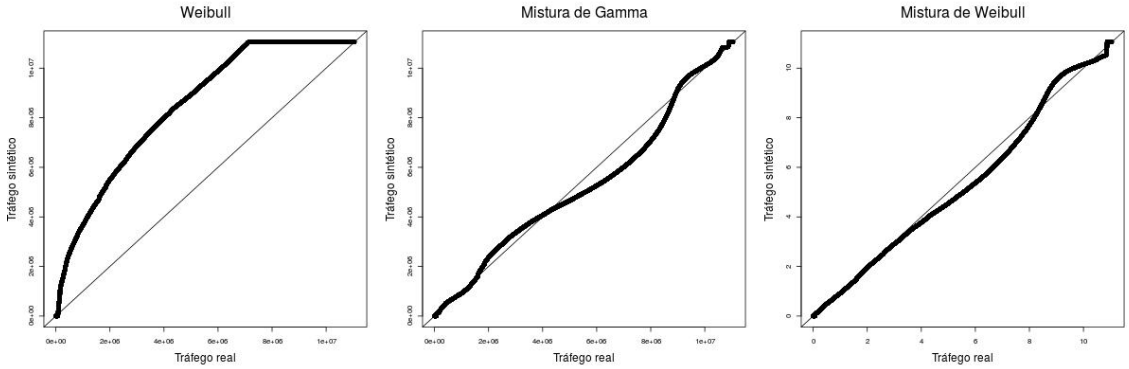


Figura A.41: Comparação entre distribuições mistas e tráfego real gerado por voluntário 4

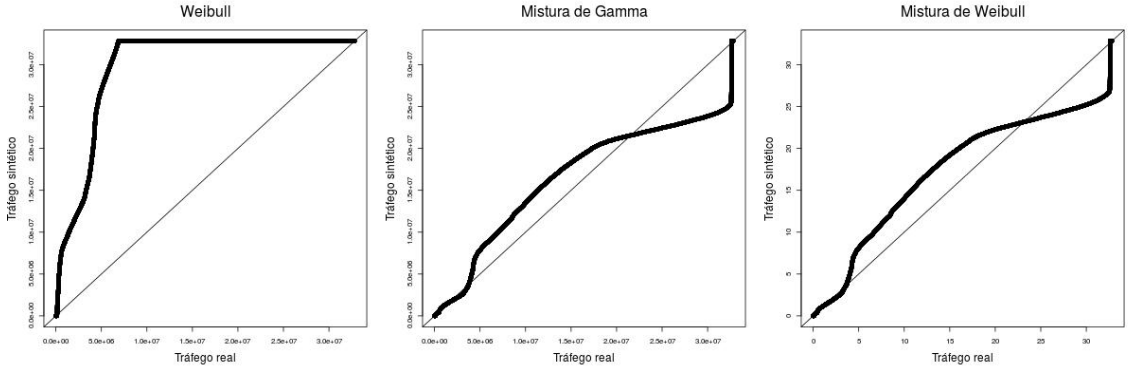


Figura A.42: Comparação entre distribuições mistas e tráfego real gerado por voluntário 5

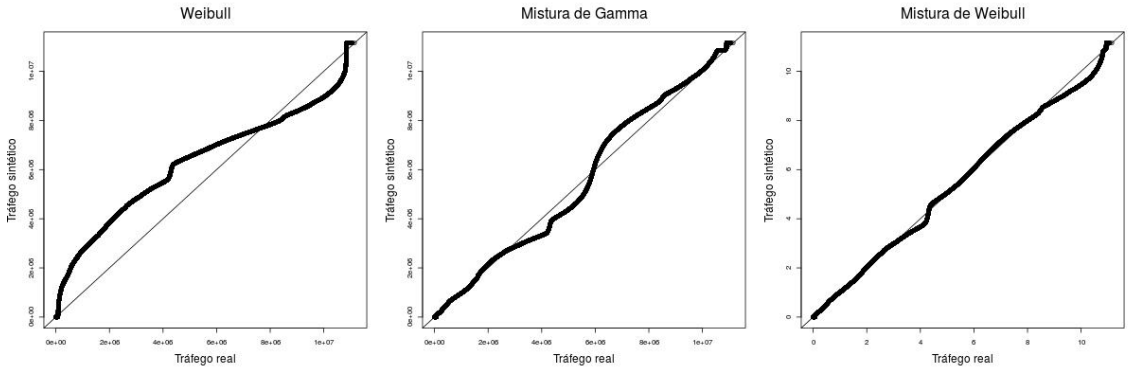


Figura A.43: Comparação entre distribuições mistas e tráfego real gerado por voluntário 6

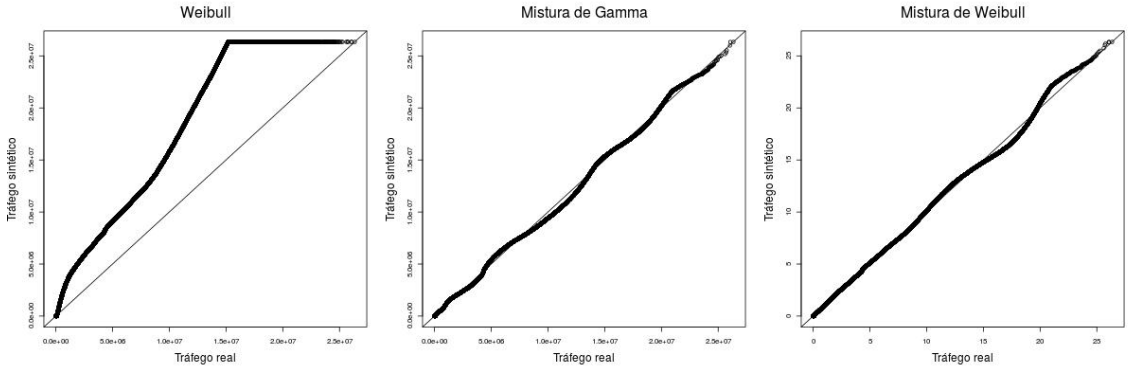


Figura A.44: Comparação entre distribuições mistas e tráfego real gerado por voluntário 7

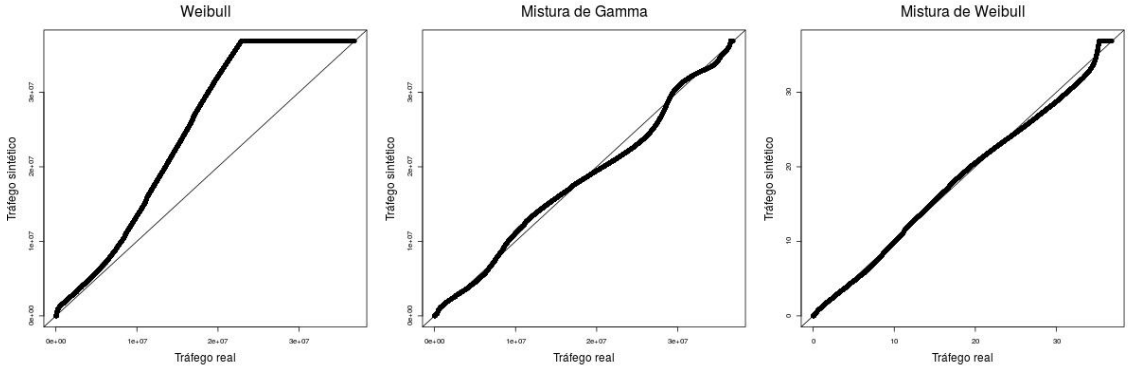


Figura A.45: Comparação entre distribuições mistas e tráfego real gerado por voluntário 8

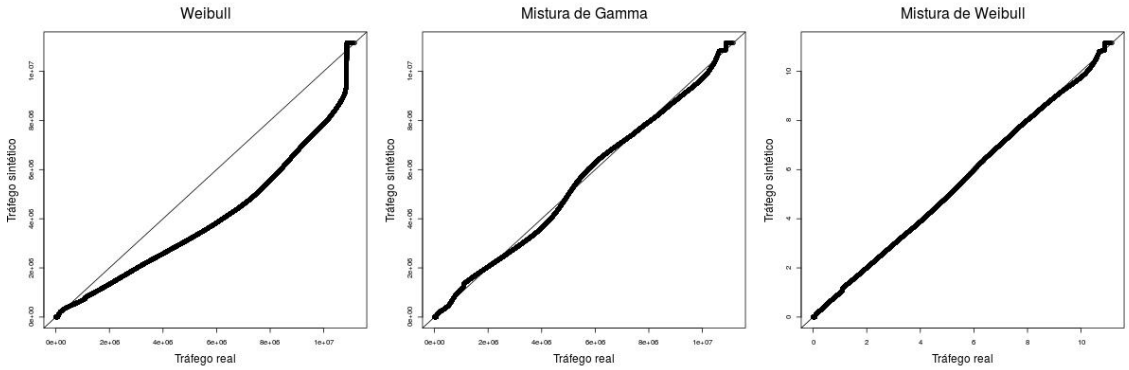


Figura A.46: Comparação entre distribuições mistas e tráfego real gerado por voluntário 9

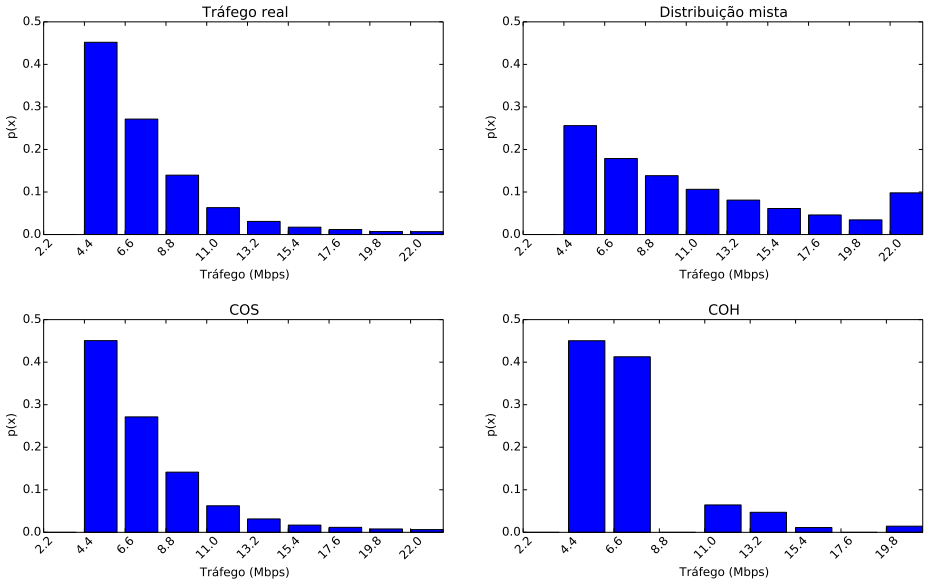


Figura A.47: Comparação entre distribuições condicionais de tráfego do voluntário 1

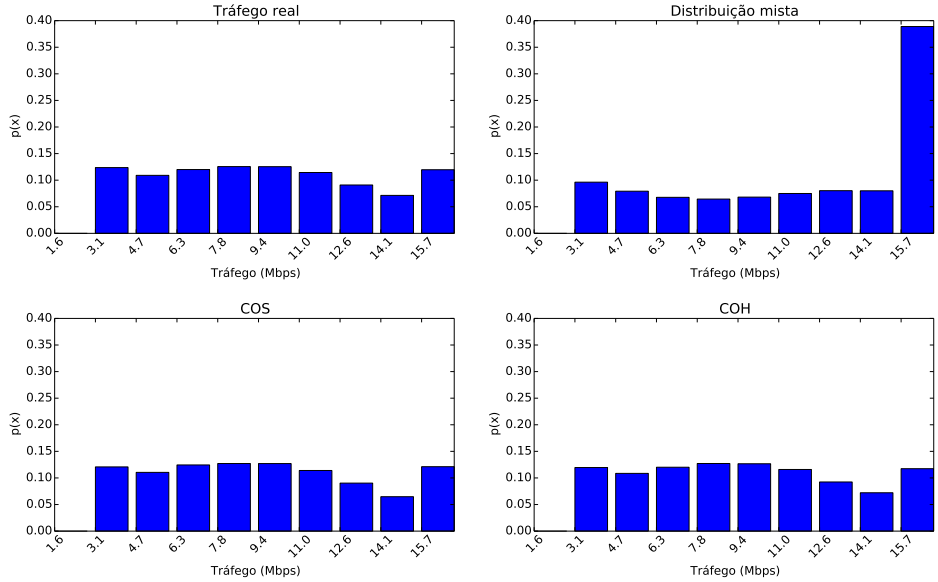


Figura A.48: Comparação entre distribuições condicionais de tráfego do voluntário 2

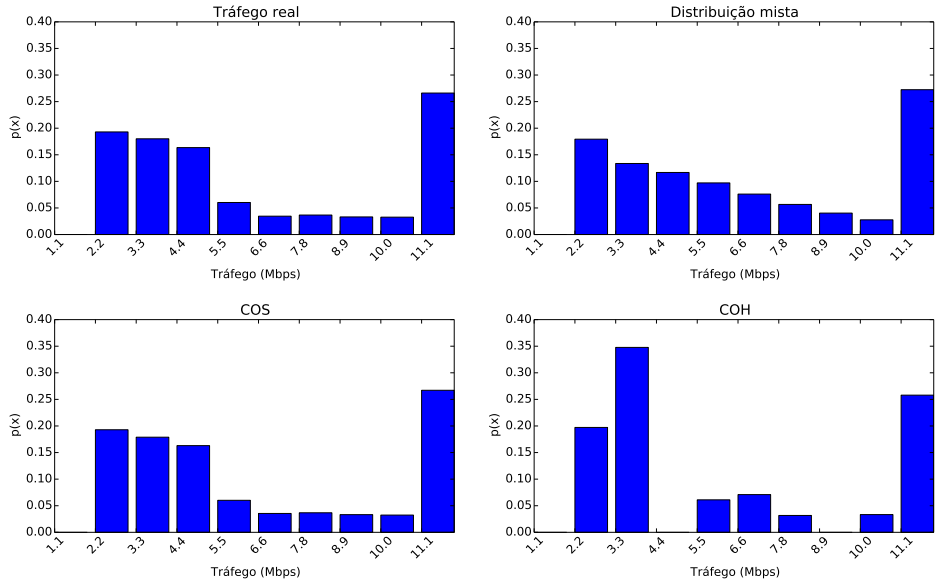


Figura A.49: Comparação entre distribuições condicionais de tráfego do voluntário 3

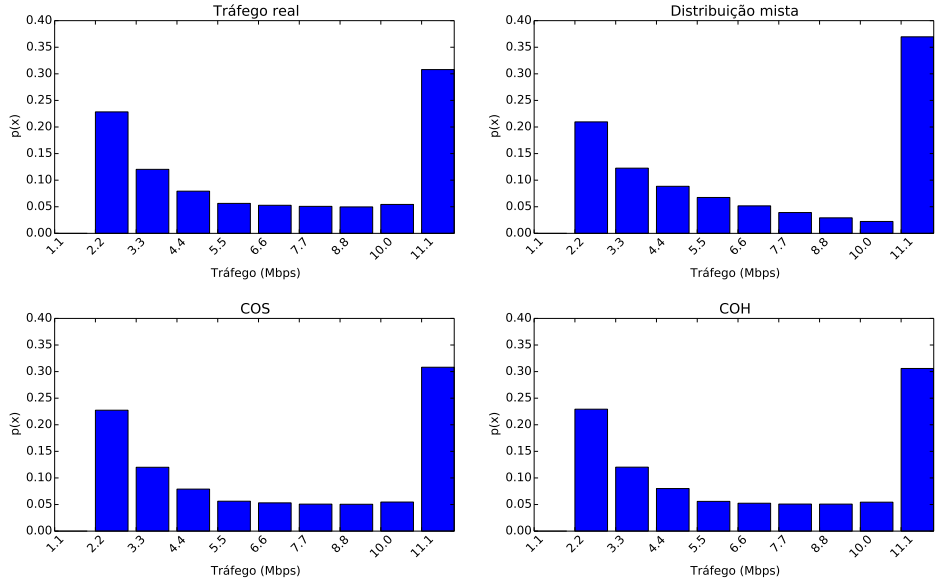


Figura A.50: Comparação entre distribuições condicionais de tráfego do voluntário 4

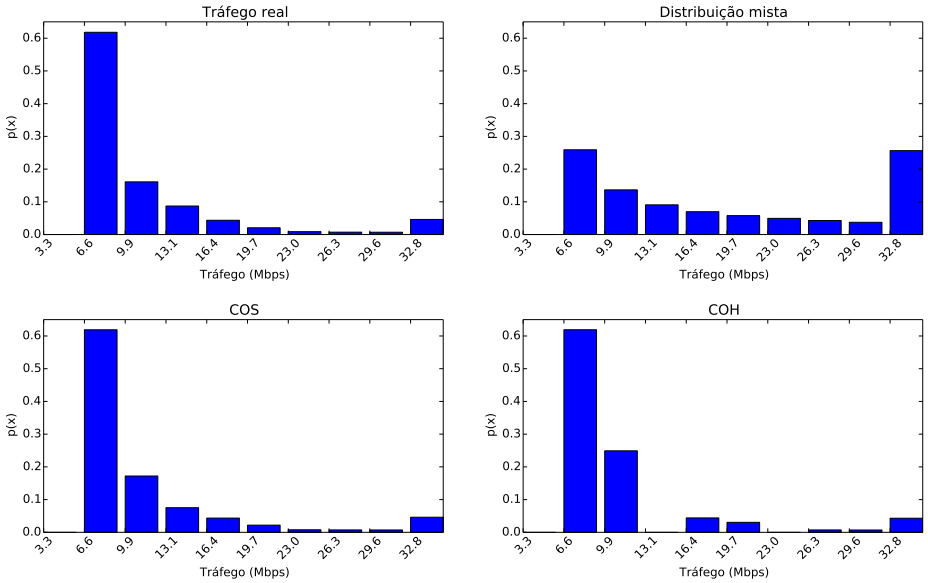


Figura A.51: Comparação entre distribuições condicionais de tráfego do voluntário 5

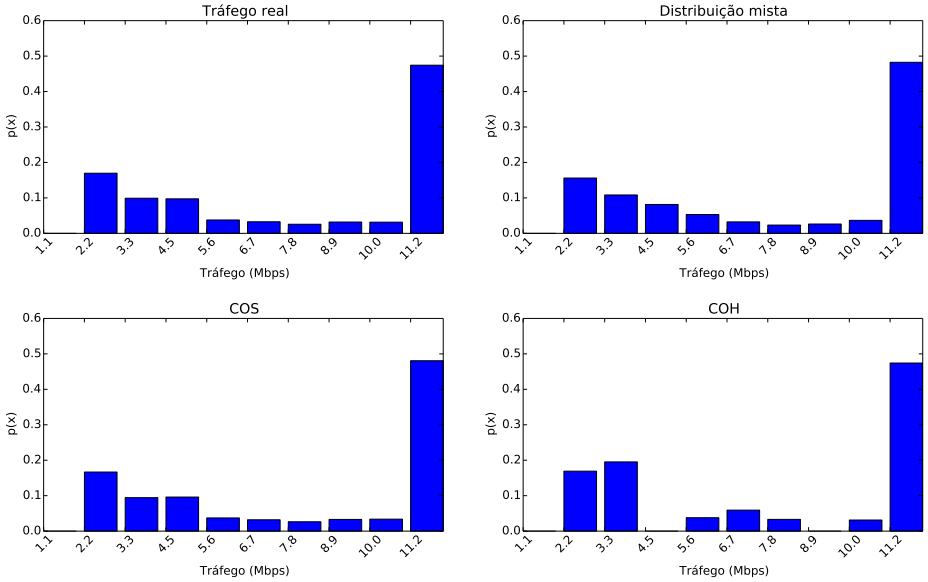


Figura A.52: Comparação entre distribuições condicionais de tráfego do voluntário 6

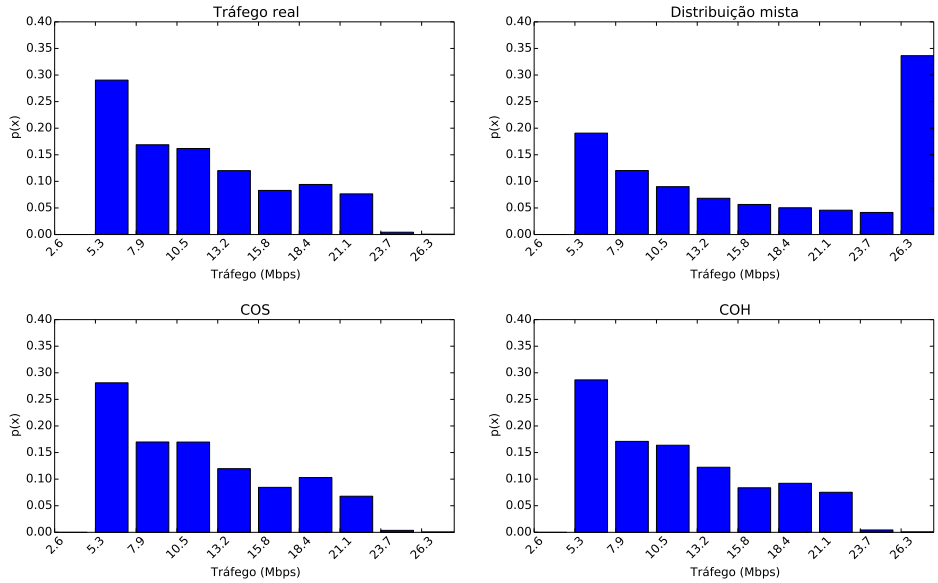


Figura A.53: Comparação entre distribuições condicionais de tráfego do voluntário 7

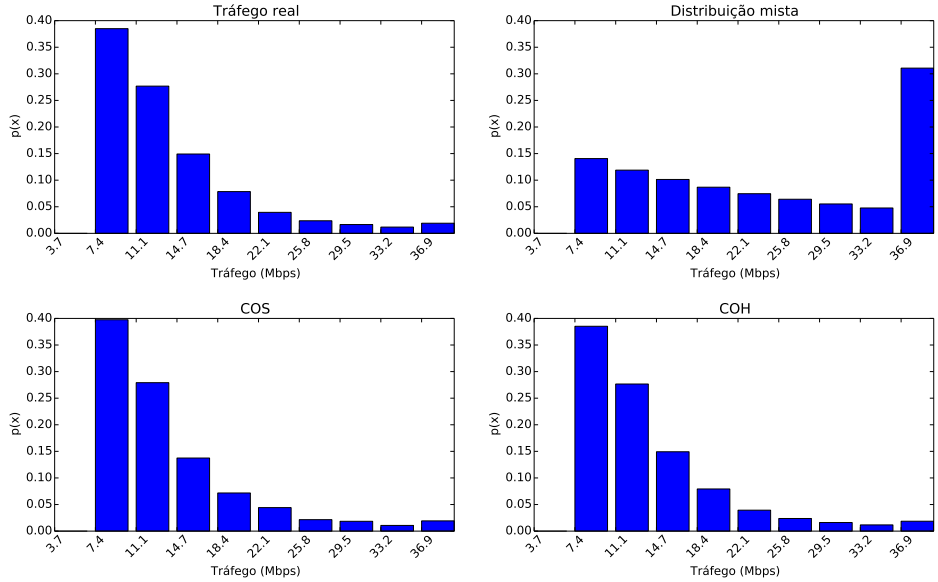


Figura A.54: Comparação entre distribuições condicionais de tráfego do voluntário 8

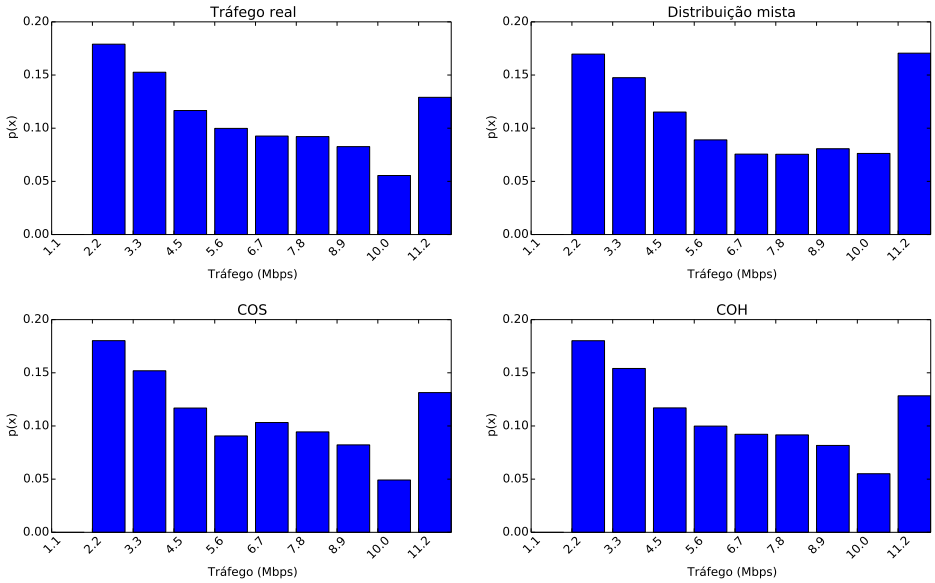


Figura A.55: Comparação entre distribuições condicionais de tráfego do voluntário 9