Avaliação de Técnicas de Agrupamento na Amostragem de Tráfego na Internet

Stenio Fernandes[1] , Carlos Kamienski[1,2], Denio Mariz[1,2] e Djamel Sadok[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 7851 – Cidade Universitária – 50.732-970 – Recife – PE

[2]Centro Federal de Educação Tecnológica da Paraíba (CEFET-PB)
Rua 1º de Maio, 720 – 58015-180 – João Pessoa, PB

{stenio, cak, denio, jamel}@gprt.ufpe.br

Abstract. *Gathering information about Internet traffic profile is a key activity for network management and operation. Such tasks surely help identifying applications behavior and detecting traffic abnormality. Most routers provide data concerning flow information summaries which are widely available for analysis. For traffic analysis, one important step is finding groups that share common properties. In this paper, we apply statistical clustering techniques in traffic analysis of Internet flows. We evaluate two partitioning clustering methods, namely CLARA and K-Means, and validate their outcomes by using them as thresholds for stratified sampling.*

Resumo. *A obtenção de informações sobre o comportamento do tráfego na Internet é fundamental para atividades de gerenciamento e operação, como identificação de aplicações e detecção de anomalias. A maioria dos roteadores é capaz de gerar informações de fluxos de pacotes para sumarizar o tráfego na rede. Uma parte significativa do trabalho de análise de fluxos de tráfego é encontrar grupos de interesse que compartilham características semelhantes. Neste artigo, são utilizadas técnicas estatísticas de análise de agrupamento no estudo do tráfego na Internet. São avaliados os algoritmos de agrupamento exclusivo CLARA e K-Means. A relevância dos grupos encontrados pelos algoritmos é validada através da sua utilização como limites de estratificação para amostragem.*

1. Introdução

Durante muito tempo, a maioria dos esforços em pesquisa e desenvolvimento relacionados à arquitetura da Internet se concentraram em desenvolver protocolos e tecnologias para permitir que a rede se tornasse uma plataforma robusta e flexível para o florescimento de diversas aplicações. O interesse em compreender o comportamento do tráfego no interior da Internet é algo que somente teve um grande impulso quando os perfis de tráfego começaram a mudar do tradicional dominado pela Web para novas aplicações como P2P, códigos maliciosos (vírus, vermes) e tráfego multimídia. Com isto, atividades de gerenciamento e operação das redes IP se tornaram de vital importância para o devido planejamento de capacidade, monitoramento de desempenho, identificação de aplicações e detecção de anomalias [10]. De um modo geral, pode-se afirmar que a medição e análise de tráfego na Internet são essenciais para o desenvolvimento de várias estratégias vitais para os

provedores e grandes redes corporativas [11], como prevenção de incidentes de segurança, identificação de modelos de negócio e de cobrança e eliminação de tráfego não desejado (ex: P2P ou vermes).

Em geral, o tratamento do tráfego em redes de alta velocidade, em termos de observação de pacotes, gera requisitos extremos em termos de medição e armazenamento, bem como análises complexas. Em conseqüência, tornou-se uma prática usual a utilização de registros de fluxos, que representam conjuntos de pacotes com campos em comum, como endereços IP e portas. A maioria dos roteadores é capaz de gerar informações de fluxos de pacotes para sumarizar o tráfego na rede. Uma parte significativa do trabalho de análise de fluxos de tráfego é encontrar grupos de interesse que compartilham características semelhantes [8]. Neste contexto, técnicas de análise de agrupamento (Cluster Analysis) podem ser de grande utilidade, pois tentam encontrar estruturas ou padrões nos fluxos sem conhecimento *a priori* ou mesmo suposições iniciais sobre suas propriedades.

Neste artigo, são utilizadas técnicas estatísticas de análise de agrupamento no estudo do tráfego na Internet. São avaliados os algoritmos de agrupamento exclusivo CLARA e K-Means. A relevância dos grupos encontrados pelos algoritmos é validada através da sua utilização como limites de estratificação para amostragem. Diversas estratégias de amostragem têm recentemente sido propostas como forma de otimizar o processo de seleção de fluxos. A técnica de amostragem estratificada ótima tem apresentado bons resultados na diminuição do volume de dados e posterior recuperação das propriedades essenciais do tráfego [7][6]. No entanto, a determinação do limite dos estratos é geralmente empírica, ditada pelo conhecimento dos especialistas na área. Mudanças de perfil de tráfego podem causar o método de estratificação fazer agrupamentos errados e portanto invalidar os resultados das inferências. A contribuição deste artigo se concentra no fato de através das técnicas de agrupamento, permitir que os limites nos estratos sejam determinados pelos próprios registros dos fluxos, portanto gerando uma adaptação dinâmica que independe do conhecimento prévio e configuração estática de limites. Em outras palavras, nossa proposta de utilização de técnicas estatísticas de agrupamento contribui para um melhor entendimento dos perfis de tráfego, bem como auxilia na implantação de sistemas de medição, coleta e análise de tráfego com necessidade mínima de parametrização.

Os resultados apresentados neste artigo mostram que os algoritmos CLARA e K-Means são adequados para a realização de estratificação baseada na métrica da duração dos fluxos. O método CLARA em geral apresentou resultados superiores ao K-Means, devido à sua capacidade de distribuir melhor os fluxos pelos estratos.

Na seqüência, a seção 2 apresenta alguns trabalhos relacionados sobre técnicas de agrupamento e técnicas de amostragem em monitoramento de tráfego. A seção 3 apresenta a fundamentação teórica usada neste trabalho de análise de tráfego, as seções 4, 5 e 6 exibem os principais resultados obtidos, enquanto a seção 7 apresenta as conclusões e os possíveis trabalhos futuros.

2. Trabalhos Relacionados

A eficiência e validação dos métodos estatísticos de análise de agrupamento tem sido comprovadas através de aplicações em diversas áreas de conhecimento. Diversas propostas para melhoria da eficiência computacional de tais métodos são apresentadas, principalmente quando os conjuntos de dados são considerados de grande volume. Além disso, diversas

aplicações para as técnicas de agrupamento foram propostas recentemente. Por exemplo, em [14] Xian et al. usam uma variação do algoritmo K-Means para aumentar a taxa de acertos em sistemas de detecções de anomalias, ao mesmo tempo em que tentam reduzir o número de alarme falsos. Em [9], Laiho et al. usam técnicas de agrupamento e redes neurais em dados de redes celulares 3G. Sua proposta parte do princípio que em tais redes, os dados são compostos de centenas de variáveis diferentes por célula. Isto torna complexa, por exemplo, a análise de identificação de células com comportamentos similares.

Especificamente tratando de análise de tráfego na Internet, o trabalho de Hernández-Campos et al. [4] apresenta um modelo abstrato de comunicação na Internet e desenvolve uma metodologia para agrupamento de conexões TCP em grupos baseados nos padrões de uso da rede. Além disso, os autores desenvolveram técnicas de visualização para melhor interpretação dos resultados do agrupamento. Existem diversas diferenças entre o trabalho de Hernandez-Campos e a nossa proposta. A principal refere-se à metodologia usada para o agrupamento. Para obter grupos de comportamento, os autores usam as técnicas hierárquicas (aglomerativas e divisivas), enquanto nosso trabalho utiliza técnicas de agrupamento por partição. Outro ponto de diferenciação refere-se à granularidade da informação de tráfego, visto que eles trabalham com observação de pacotes enquanto nós observamos os fluxos já consolidados.

Em relação à aplicação das técnicas de agrupamento em amostragem de tráfego na Internet, a abordagem mais próxima à nossa é o trabalho desenvolvido por Duffield et al [2]. Os autores apresentam o uso da amostragem dependente do tamanho dos objetos (a saber, registros de fluxos), ao invés do uso da amostragem uniforme, com objetivo de diminuir o volume de tráfego coletado. Em tal esquema, conhecido como Amostragem por Limiar ou Amostragem Inteligente (Threshold ou Smart Sampling), um objeto de tamanho x é selecionado de acordo com uma função de probabilidade $p_z(x) = \min\{1, x/z\}$. Portanto, fluxos de tamanho menor que um limiar z, são selecionados com probabilidade x/z. Por outro lado, fluxos com tamanho igual ou maior que z, são sempre selecionados. Os autores demonstram que através desta técnica é possível controlar o tamanho da amostra e a variância dos estimadores (e.g., a soma dos tamanhos dos objetos amostrados).

3. Fundamentação Teórica

3.1. Análise de Agrupamento – CA (Cluster Analysis)

Análise de agrupamento (CA) é um procedimento fundamental ao se deparar com problemas relacionados à extração de informação a partir de dados multivariados. Em outras palavras, tais métodos tentam encontrar estruturas ou padrões nos conjuntos de dados sem conhecimento *a priori* ou mesmo suposições iniciais sobre suas propriedades. CA tenta organizar objetos em grupos com forte similaridade entre eles e alta dissimilaridade em relação aos outros grupos identificados. Este processo de classificação e divisão dos dados resulta em agrupamentos intrínsecos de acordo com alguma métrica de distância definida. O conjunto de técnicas de agrupamento pode ser bastante útil quando se está interessado em encontrar os objetos mais representativos para cada grupo homogêneo (i.e., num processo de amostragem) ou encontrar objetos não usuais (i.e., valores de exceção).

Considerando que se podem usar algoritmos e critérios para análise de agrupamento, o resultado da classificação pode diferir em algumas situações. Em outras palavras, uma seleção diferente de parâmetros ou categoria do algoritmo, pode resultar em diferentes

grupos. Neste caso, uma decisão prudente para um agrupamento apropriado seria experimentar diversos algoritmos de classificação ou parametrização, com o objetivo de obter classificações aceitáveis.

Um elemento chave em qualquer algoritmo é a métrica de distância que determina a proximidade entre os pares de dados. Uma abordagem genérica para dados multidimensionais, é utilizar uma medida bastante conhecida como a métrica de Minkowski, que tem a seguinte formulação:

$$d_p\left(x_i, x_j\right) = \left(\sum_{k=1}^{d} \left| x_{i,k} - x_{j,k} \right|^p \right)^{1/p}$$

onde *d* é o grau de dimensão dos dados. Observe que a distância Euclidiana e a distância Manhatan são casos especiais da formulação acima mencionada, quando *p = 2* e *p=1*, respectivamente.

Em relação às categorias dos algoritmos de CA, devido à sua grande variedade, não existe uma classificação canônica. A principal razão é que os algoritmos de CA variam de acordo com o tipo de dado de entrada, o critério de agrupamento na definição de similaridade ou os conceitos nos quais eles se baseiam (p.ex., estatística ou lógica nebulosa). Na realidade, alguns autores afirmam que tais algoritmos em alguns casos permeiam diversas categorias. Neste trabalho, usamos uma classificação mais simples, onde os algoritmos de agrupamento podem ser ou hierárquicos ou por partição. Contudo, existem diversas outras categorias, tais como agrupamento baseado em densidade, probabilístico, monotético ou politético, etc.

K-Means é um dos mais simples e populares algoritmos de CA. Em linhas gerais, o procedimento segue uma maneira direta de classificar dados em um certo número de grupos. A idéia principal é definir *k centróides*, um para cada grupo. O algoritmo K-Means atribui cada elemento do conjunto de dados ao grupo cujo *centróide* é o mais perto. A partir deste ponto, ele reavalia os centros (i.e., o algoritmo recalcula *k* novos *centróides*) e continua a atribuição até que os centros parem de mudar.

O algoritmo K-Means tenta minimizar uma função objetivo, neste caso uma função de erro quadrático conhecida como a variância total intra-cluster. A função objetivo é descrita como segue:

$$J = \sum_{j=1}^{k} \sum_{i \in S_i} \left\| x_i^{(j)} - c_j \right\|^p$$

onde $\left\| x_i^{(j)} - c_j \right\|^p$ é uma métrica de distância escolhida entre o ponto $x_i^{(j)}$ e o centro do grupo c_j, é um indicador da distância dos *n* pontos de dados dos respectivos centros dos grupo.

PAM (Partitioning Around Medóides) é uma outra técnica classificada como particional que tem propriedades semelhantes ao K-Means. Seu principal objetivo é a determinação de objetos representativos (i.e., *medóides*) para cada grupo. Em outras palavras, PAM tenta encontrar os objetos mais centralmente localizados dentro dos grupos. A primeira fase do algoritmo seleciona objetos aleatoriamente e os define como *medóides* para cada grupo *c*. Após isto, cada um dos objetos não selecionados é agrupado junto ao

medóide ao qual ele é mais próximo. PAM permuta os *medóides* com outros objetos não selecionados até que todos sejam qualificados como *medóides*. Em outras palavras, PAM tenta repetidamente escolher os melhores centros. A qualidade de cada agrupamento resultante após a troca do *medóide* é avaliada. Claramente, PAM é um algoritmo dispendioso com $O(n^2)$ como ordem de complexidade assimptótica em relação ao uso de memória. Portanto, a utilização do PAM para grandes conjunto de dados torna-se inviável.

CLARA (Clustering LARge Applications) é uma implementação do PAM para lidar com grande volume de dados, uma vez que seu tempo computacional e os requisitos de uso de memória é linear ao número de elementos (i.e., $O(n)$). CLARA é um processo essencialmente com dois estágios. No primeiro, uma amostra com um tamanho pré-definido é retirada do conjunto de dados e distribuída nos *k* grupos, usando PAM. Então, CLARA seleciona *medóides* sucessivos objetivando obter a menor distância entre os objetos da amostra (fase BUILD). Posteriormente, CLARA tenta diminuir a distância média entre os objetos, trocando os elementos mais representativos (processo SWAP). Isto é feito repetidamente até que todos os grupos tenham menor distância média entre os objetos.

Indicamos a leitura da referência [5] para uma descrição mais profunda dos algoritmos K-Means e CLARA.

3.2. Amostragem Estratificada

Na técnica de amostragem estratificada [1], uma população de N unidades é primeiramente dividida em sub-populações de $N_1, N_2, \ldots, N_L$ unidades, respectivamente. Essas sub-populações não se superpõem e, juntas abrangem a totalidade da população de tal modo, que $N_1 + N_2 + \ldots + N_L = N$. As sub-populações são denominadas estratos. Para que se obtenham todos os benefícios da estratificação, os valores de N_h devem ser conhecidos. Depois de determinados os estratos, seleciona-se uma amostra em cada um deles, sendo as seleções feitas separadamente nos diferentes estratos. As grandezas das amostras dentro dos estratos são denominados $n_1, n_2, \ldots, n_L$, respectivamente. Em geral, é possível dividir uma população heterogênea em sub-populações que isoladamente sejam homogêneas. Se todos os estratos são homogêneos, no sentido de que o valor das medidas variem pouco de uma unidade para outra, pode-se obter uma estimativa precisa do valor médio de um estrato qualquer mediante uma pequena amostra desse estrato. Por fim, essas estimativas podem ser combinadas para constituírem uma estimativa precisa do conjunto da população.

A amostragem estratificada pode ser classificada em uniforme, proporcional ou de Bowley e ótima. Na amostragem estratificada uniforme os estratos têm o mesmo tamanho, enquanto que na amostragem proporcional o número de elementos em cada estrato é proporcional ao tamanho do estrato. Por fim, a amostragem estratificada ótima considera além do tamanho do estrato e variabilidade dentro do estrato.

Neste artigo, utiliza-se a amostragem estratificada sem reposição com distribuição ótima. Ou seja, os resultados obtidos consideram o tamanho e a variabilidade do estrato. Suponha que se pretenda utilizar a repartição ótima para um determinado n. A grandeza da amostra, n'_h, no estrato h deve ser

$$n \geq \frac{k^2 \sigma_I^2 N - k^2 \sigma_\sigma (N-1)}{\varepsilon^2 (N-1) + k^2 \sigma_I^2} \qquad (1)$$

onde,

$$\sigma_\sigma^2 = \frac{\sum N_h \sigma_h^2}{\sum N_h} - \left(\frac{\sum N_h \sigma_h}{\sum N_h} \right)^2 \qquad \sigma_I^2 = \frac{\sum N_h \sigma_h^2}{\sum N_h}$$

k : quantil $(1-\alpha)$ da distribuição normal padrão, ε : erro de precisão.

Neyman [1] estabeleceu um critério de distribuição dos elementos da amostra pelos diferentes estratos a partir da condição de ser mínima a variância resultante. De acordo com esse critério o número n_h de elementos do estrato h, em uma amostragem de n elementos será dado pela expressão:

$$n_h = n \frac{N_h \sigma_h}{\sum N_h \sigma_h}$$

onde, n: tamanho da amostra; N_h: total de unidades; σ_h: desvio padrão dentro dos estratos.O propósito é determinar o tamanho n da amostra que se deve extrair para estimar uma característica qualquer desse universo como, por exemplo, o tamanho médio dos fluxos de tráfego ou sua duração média.

4. Metodologia de Coleta de Tráfego e Análise

Neste artigo foram utilizados quatro conjuntos de dados, cada um contendo traces (rastros) de fluxos de tráfego. Adota-se a definição padrão de fluxo, que é o conjunto de pacotes com os mesmos valores dos campos endereço IP de origem e destino, porta (TCP ou UDP) de origem e destino e protocolo. Os traces foram obtidos no Ponto de Presença de Pernambuco (PoP-PE) da Rede Nacional de Pesquisa (RNP), nos dias 15 a 19 de setembro de 2004. A geração dos arquivos de traces contendo os fluxos foi realizada por um capturador de tráfego baseado em pacotes, desenvolvido especialmente para o Grupo de Trabalho em Computação Colaborativa (GT-P2P), da RNP [13]. Foi coletado todo o tráfego de entrada e saída do PoP-PE, que possuía um enlace de 34 Mbps no período em considerado neste estudo. A Tabela 1 apresenta um resumo das principais características dos traces utilizados.

Tabela 1 – Características dos traces utilizados na análise.

Nome	Data	Horário	Volume (GB)	Número de fluxos
Trace 1	15/09/2004	8h às 12h	28,129	7.013.744
Trace 2	17/09/2004	8h às 12h	37,958	7.497.991
Trace 3	18/09/2004	14h às 18h	23,875	4.086.476
Trace 4	19/09/2004	14h às 18h	62,511	6.571.586

A metodologia utilizada para a aplicação das técnicas de agrupamento juntamente com amostragem estratificada foi baseada em 7 passos:

1. Definição das variáveis: foram consideradas as variáveis duração do fluxo em segundos e volume de tráfego do fluxo em bytes. Somente duração foi utilizada para estratificação, enquanto que ambas foram usadas como variáveis de observação.
2. Análise de agrupamentos: Um passo importante na amostragem estratificada é decidir os limites dos estratos. Em [7] esses limites foram decididos manualmente com

aumento exponencial. Neste trabalho os algoritmos CLARA e K-Means foram utilizados para obter grupos a partir dos quais os limites dos estratos para a variável duração foram definidos. Estes resultados estão descritos na seção 5.

3. Definição do número de estratos: A avaliação da estratificação usou quatro diferentes quantidades de estratos, a saber, 2, 4, 6 e 8. Uma vez que a utilização de um número maior de estratos implica em maior complexidade no processamento, o objetivo é encontrar o menor número possível de estratos com o qual se obtém resultados aceitáveis. Em outras palavras, a precisão das métricas descritivas da amostra e da população deve ser alcançada com um número menor de estratos.
4. Definição do tamanho da amostra: Em geral, procura-se obter a maior redução possível no tamanho da amostra, sem comprometer a sua capacidade em representar a população original. Neste trabalho foram avaliados quatro tamanhos de amostras: 0,01%, 0,1%, 1% e 10% do tamanho original da população. Em [7] foi também utilizado o cálculo do tamanho ótimo baseado no método de Neyman, que não se aplica neste trabalho, porque varia para cada número de estratos avaliados.
5. Definição do número de elementos por estrato: Isto deve ser calculado para cada conjunto de dados, para todos os números de estratos e tamanhos de amostras, usando o método de Neyman.
6. Execução de simulação estocástica de Monte Carlo: Consiste em escolher uma amostra para cada estrato usando o software estatístico R [12], repetindo o procedimento 100 vezes (número de réplicas). Em cada replicação, foi coletado o valor de cada métrica e ao final foram calculados a média e o desvio padrão e o intervalo de confiança assintótico ao nível de 99 %.
7. Cálculo e comparação de métricas descritivas da amostra e da população: Média (do tamanho e da duração dos fluxos), desvio padrão e soma foram as métricas observadas. No entanto, somente os resultados para a média são apresentados na seção 6, uma vez que os resultados para soma e desvio padrão não agregam muitas informações complementares, conforme mostrado em [7] e [6].

Além da amostragem estratificada, também foram executadas simulações para as amostragens uniforme e inteligente (ou de limiar), para fins de comparação.

5. Análise de Agrupamentos

Esta seção analisa os grupos gerados pelas técnicas CLARA e K-Means, considerando que as variáveis duração e volume foram fornecidas como entrada para os algoritmos. O estudo foi realizado para os quatro conjuntos de dados disponíveis, mas somente os resultados referentes ao dia 15/09 são mostrados nesta seção, devido à grande similaridade observada entre eles. Para cada algoritmo, foi feito o cálculo do agrupamento considerando quatro quantidades diferentes de grupos: 2, 4, 6 e 8.

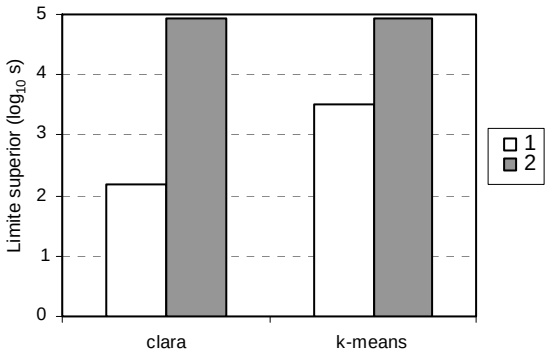
A Tabela 2 apresenta os limites inferiores e superiores gerados por CLARA e K-Means para dois grupos. Pode-se observar que os dois algoritmos produziram grupos não sobrepostos para a duração. Por exemplo, para CLARA, fluxos com até 155 segundos são classificados no grupo 1, enquanto que, acima desse valor, no grupo 2. Para o volume, no entanto, os grupos estão sobrepostos. Por exemplo, K-Means tem o primeiro estrato iniciando em 23 bytes e terminado com 115 MB e o segundo iniciando em 330 bytes e

terminando em 426 MB. Isto significa que um fluxo de tamanho 1 MB pode ser classificado em ambos, o que gera uma ambigüidade para a aplicação em amostragem. Por isso, as simulações de amostragem estratificada apenas utilizaram a duração como variável de estratificação. A título de comparação, o trabalho descrito em [7] apresenta a estratificação para volume e duração, pois os limites dos estratos foram definidos manualmente e as estratificações foram executadas separadamente.

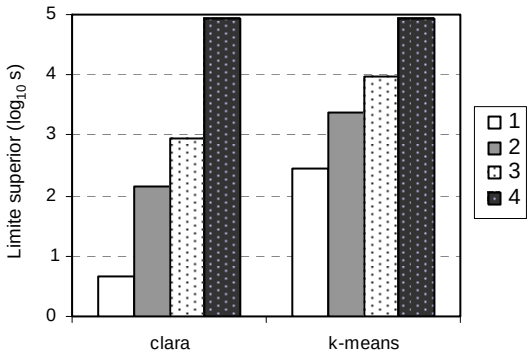
Tabela 2 – Limites inferiores e superiores para 2 grupos (15/09)

Grupo	Duração (segundos)				Volume			
	CLARA		K-Means		CLARA		K-Means	
	Inf.	Sup.	Inf.	Sup.	Inf.	Sup.	Inf.	Sup.
1	0	155	0	3.252	23 B	24 MB	23 B	115 MB
2	155	83.823	3.254	83.823	56 B	426 MB	330 B	426 MB

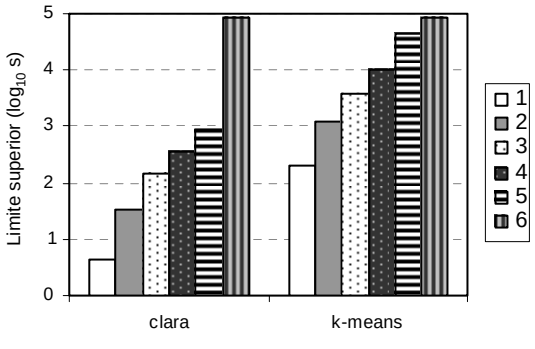
A Figura 1 mostra graficamente os valores dos limites superiores da variável duração para 2, 4, 6 e 8 grupos gerados pelos algoritmos CLARA e K-Means. Todos os gráficos estão em escala logarítmica e os valores do eixo Y representam os expoentes na base 10. É possível observar claramente que os grupos gerados por CLARA têm limites menores que os gerados por K-Means (aproximadamente dois níveis de magnitude). Isto se mantém para todos os números de grupos avaliados e significa que K-Means inclui um número maior de elementos no primeiro grupo, de modo que os demais grupos ficam com tamanhos menores. Como será visto na seção 6.1, este fato tem um impacto nos resultados da estratificação desses algoritmos de agrupamento.

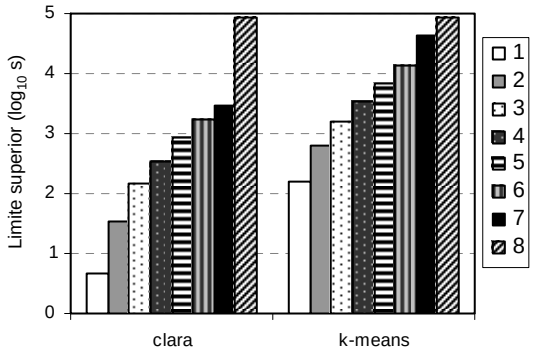


(a) 2 grupos



(b) 4 grupos





(c) 6 grupos (d) 8 grupos

Figura 1 - – Limites superiores dos grupos (15/09)

Os dois algoritmos de agrupamento apresentaram resultados que os tornam adequados para uso na estratificação. Por este motivo, na seção 6, são apresentados os resultados da estratificação com a utilização dos limites superiores dos grupos em lugar dos limites dos estratos, com exceção do último grupo.

6. Experimentos com Amostragem Estratificada

Esta seção mostra os resultados obtidos com a amostragem estratificada, onde os valores nos gráficos representam a média da métrica analisada (a média). Para todos os experimentos foram executadas 100 replicações e os intervalos de confiança assintóticos ao nível de 99% foram calculados. Os valores apresentados nos gráficos referem-se à média das 100 replicações. Nas seções 6.1, 6.2 e 6.3 foi observada a variável direta duração dos fluxos e na seção 6.4 a variável de observação foi o volume.

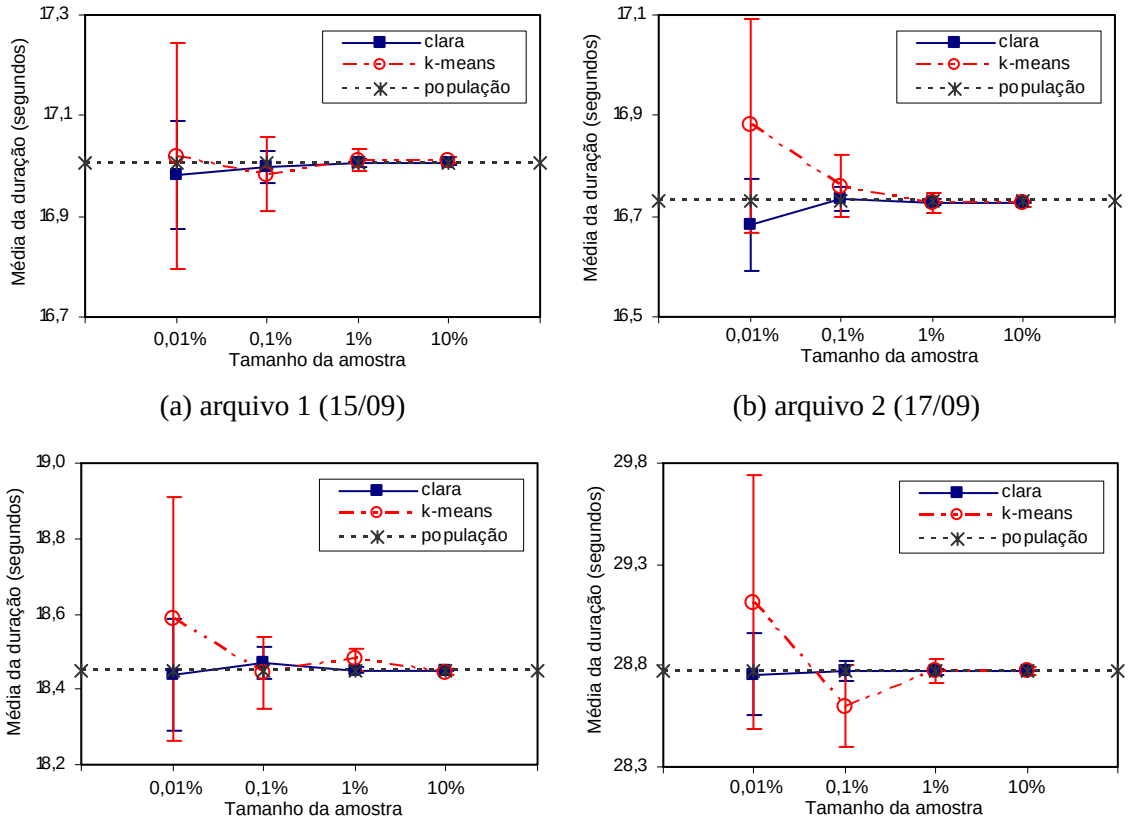
6.1. Análise de Estratificação

Esta seção compara o efeito da estratificação gerada pelos métodos CLARA e K-Means, para diferentes dias e diferentes números de estratos. A Figura 2 apresenta os resultados da média da duração dos fluxos para os quatro arquivos de dados (quatro dias), incluindo os quatro tamanhos amostrais e para oito estratos, considerando os limites dos estratos obtidos a partir dos métodos de agrupamento CLARA e K-Means. Uma análise gráfica mostra que quanto maior o tamanho da amostra, mais preciso é o intervalo de confiança, ou seja, a sua amplitude é menor (a diferença entre o limite superior e inferior). A linha tracejada representa a média real da população calculada para todos os arquivos de fluxos capturados. Pode-se observar que as estratificações produzidas pelos métodos CLARA e K-Means são relevantes, pois a média da população cai dentro dos limites de todos os intervalos de confiança. Isto significa que os intervalos construídos a partir das amostras são representativos e as amostras podem ser usadas para substituir a população para esta variável (a média da duração dos fluxos). Por exemplo, na Figura 2(a) a média da população é 17,005 s, enquanto que a média das amostras para um tamanho amostral de 0,1% é 16,998 ± 0.030 s para CLARA e 16,984 s ± 0,072 s para K-Means.

Este primeiro resultado (Figura 2) enfatiza o compromisso existente entre a escolha de um tamanho de amostra e a sua precisão (incluindo a variabilidade). Sempre que uma maior precisão é necessária, um tamanho maior da amostra deveria ser usado, para aproximar os resultados para população amostrada. Por outro lado, se uma precisão mais baixa pode ser tolerada em algum tipo de análise, mesmo amostras de tamanho 0,01% da população pode ser usada, gerando menores demandas de processamento e armazenamento.

Comparando os resultados obtidos para as técnicas de agrupamento fica explícito que CLARA produz melhores limites de estrato do que K-Means. Isto pode ser observado tanto para a precisão (proximidade da média da população), quanto para a variabilidade (tamanho dos intervalos de confiança). O motivo deste comportamento pode ser explicado por dois importantes fatores. Em primeiro lugar está a influência dos limites superiores gerados por estes métodos de agrupamento. Pela Figura 1 pode-se observar que o limite superior do primeiro grupo de K-Means é cerca de dois níveis de magnitude maior do que

de CLARA, para todos os números de grupos analisados. Isso faz com que a quantidade de elementos do primeiro grupo seja também maior. Por exemplo, para oito grupos no dia 15/09, CLARA produz um limite de 4,5 s e inclui 5.005.776 fluxos, que corresponde a 71,4% do total de fluxos do arquivo. Por outro lado, o limite de K-Means é 156 s, incluindo 6.863.446 fluxos, que representa 97,9% do total de fluxos. O principal motivo para a menor representatividade das amostras geradas pelo K-Means é que sendo o primeiro estrato maior, a sua variância também aumenta e então o método de Neyman identifica a necessidade de uma amostra maior para este estrato. Conseqüentemente, os demais estratos terão amostras menores, o que leva a uma grande tendência nas variáveis calculadas a partir da amostra, conforme a Figura 1 sugere. Em segundo lugar, devido aos estratos superiores conterem uma quantidade menor de fluxos de maior duração, eles têm um maior peso no cálculo da média. O tamanho reduzido desses estratos e sua grande variância fazem com que um grande percentual desses fluxos seja tomado nas amostras. A conjugação desses dois fatores leva a uma menor precisão e maior variabilidade na média das amostras geradas pelo K-Means.



(a) arquivo 1 (15/09)

(b) arquivo 2 (17/09)

(c) arquivo 3 (18/09)

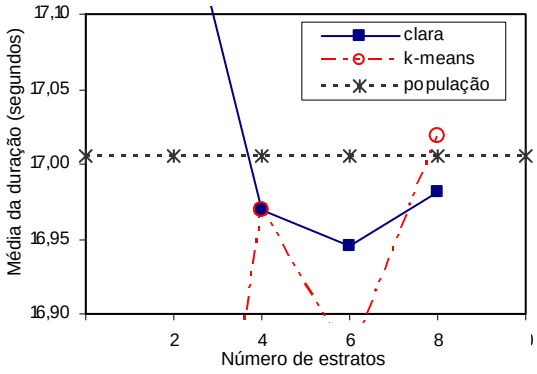
(d) arquivo 4 (19/09)

Figura 2 – Média da duração dos fluxos para diferentes arquivos de dados (8 estratos)

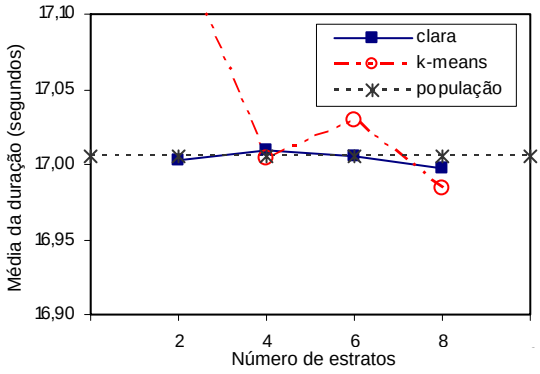
A Figura 2 também mostra que os resultados dos quatro arquivos de dados apresentam comportamento similar. Esta semelhança aparece não somente para a média de um único número de estratos, mas também para outras métricas, como a soma e a variância da duração para vários números de estratos. Por este motivo, devido a restrições de espaço, no resto desta seção apenas os resultados para o arquivo de fluxos do dia 15/09 serão apresentados (a escolha foi aleatória). A sua população é de 7.013.744 fluxos, o que gera

para tamanhos de amostra 0,01%, 0,1%, 1% e 10% um número de fluxos equivalente a 701, 7.013, 70.137 e 701.374, respectivamente. Outro aspecto que será omitido deste ponto em diante são os intervalos de confiança, que apresentaram o mesmo comportamento para todos os demais gráficos.

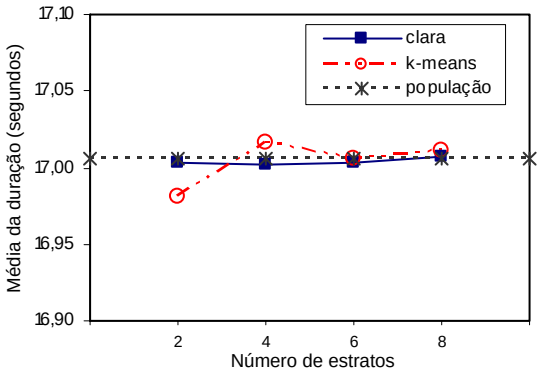
Em seguida, a Figura 3 analisa a estratificação produzida com diferentes números de grupos (2, 4, 6 e 8) gerados pelos métodos CLARA e K-Means. Deve-se notar que os resultados da Figura 2 se referem somente a 8 estratos, que foi o melhor caso encontrado. Duas importantes observações podem ser feitas a partir da Figura 3. Primeiro, tamanho de amostra de 0,01% é inadequado para um número de estratos menor do que 8, pois apresenta alta variabilidade. Segundo, resultados mais confiáveis são obtidos quando o número de estratos é no mínimo 6, pois a média da amostra apresenta maior aproximação da média da população e os intervalos de confiança são menores (não mostrados na Figura 3).



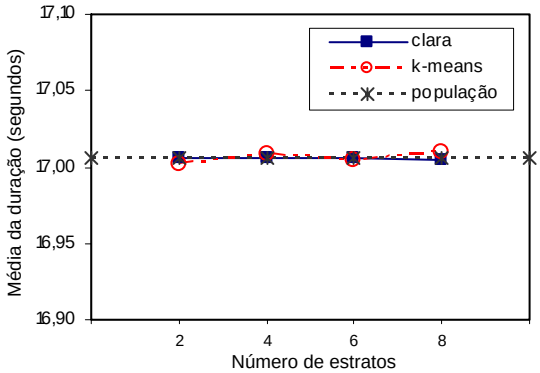
(a) amostra de tamanho 0,01%



(b) amostra de tamanho 0, 1%



(c) amostra de tamanho 1%



(d) amostra de tamanho 10%

Figura 3 – Comparação de diferentes números de estratos

6.2. Ganho de Estratificação

A técnica mais simples de amostragem consiste em retirar um elemento para a amostra a cada N elementos da população. Esta amostragem uniforme apresenta simplicidade tanto no processamento quanto no armazenamento, pois somente as amostras precisam ser armazenadas, enquanto que a amostragem estratificada necessita de processamento e armazenamento adicionais para as informações dos estratos. Isso significa que o uso de qualquer técnica mais complexa somente é justificado caso exista um ganho real perceptível. Por isso, esta seção apresenta o ganho de estratificação, que avalia o

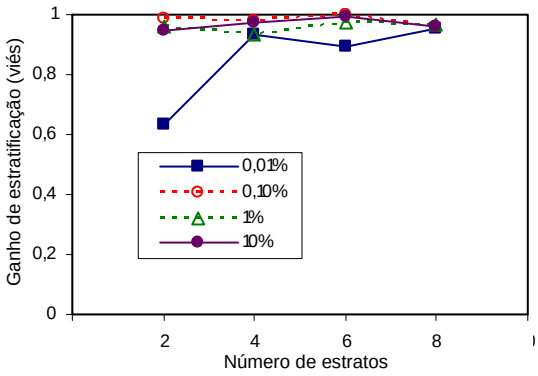
benefício do uso da amostragem estratificada sobre a amostragem uniforme. Depois, são apresentados resultados da avaliação usando o ganho de estratificação para a média e a variância da duração, que em conjunto definem a precisão e confiabilidade das estimativas.

Formalmente, o ganho de estratificação para a média g_{strat} é definido como a diferença normalizada entre o viés da média $\bar{y}$ obtida pela amostragem uniforme ($b_{\bar{y}unif}$) e média obtida pela amostragem estratificada ($b_{\bar{y}strat}$). O ganho de estratificação para a variância é definido de forma similar.

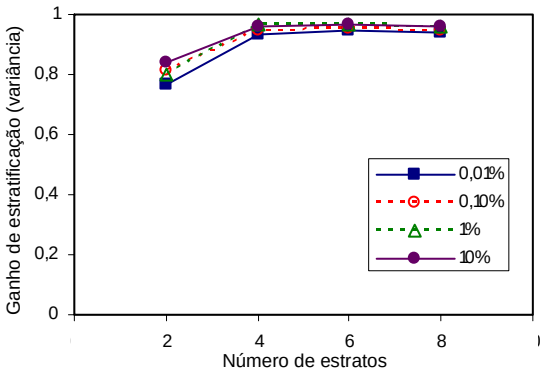
$$g_{strat} = \begin{cases} \dfrac{\left(b_{\bar{y}unif} - b_{\bar{y}strat}\right)}{b_{\bar{y}unif}}, & quando \quad b_{\bar{y}unif} \geq b_{\bar{y}strat} \\ -\dfrac{\left(b_{\bar{y}strat} - b_{\bar{y}unif}\right)}{b_{\bar{y}strat}}, & quando \quad b_{\bar{y}strat} > b_{\bar{y}unif} \end{cases} \qquad (2)$$

Da fórmula (2) pode-se depreender que o ganho de estratificação é próximo a 0 quando os dois vieses têm valores semelhantes (ou seja, não há ganho na estratificação). Ele se aproxima de 1 quando a amostragem estratificada gera um viés consideravelmente menor do que o da uniforme (ou seja, o ganho é grande). Ao contrário, se aproxima de -1 quando o viés da amostragem uniforme é menor que o da estratificada (ou seja, ocorre uma grande perda com a estratificação).

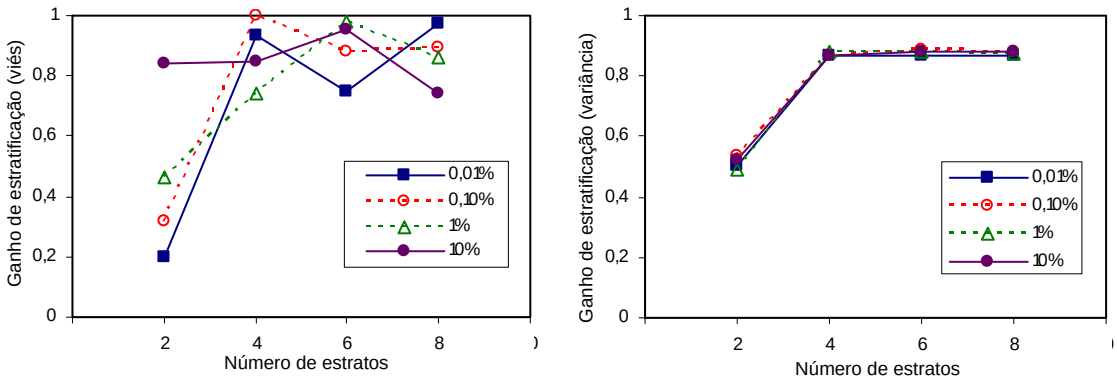
A Figura 4 apresenta o ganho de estratificação para a média e variância para os métodos CLARA e K-Means. A superioridade dos agrupamentos gerados por CLARA para compor os limites dos estratos fica mais uma vez demonstrada nestes gráficos. Enquanto CLARA gera ganhos de estratificação muito próximos a 1 para 4, 6 e 8 estratos, K-Means apresenta maior variabilidade nos ganhos. Para 2 estratos o ganho da média é acanhado para tamanhos de amostra menores que 10% e fica na média em 0,8 para outros números de estrato. Para a variância, o ganho fica na faixa de 0,85 para 4, 6 e 8 estratos. No entanto, é correto afirmar que os dois métodos de agrupamento geram ganhos de estratificação consideráveis, evidenciando os benefícios na utilização da amostragem estratificada, uma vez que todos os ganhos ficaram acima de 1.



(a) Ganho do viés (CLARA)



(b) Ganho da variância (CLARA)



(a) Ganho do viés (K-Means) (b) Ganho da variância (K-Means)

Figura 4 – Ganho de estratificação para CLARA e K-Means

Não há um ganho significativo em aumentar o número de estratos para um valor acima de 8. Isto pode ser observado na tendência das curvas, que em certos casos inclusive apresentam desempenho inferior para 8 estratos (comparado com 6), considerando-se o ganho de estratificação. O ganho de estratificação foi também calculado para a soma (da duração dos fluxos) e os resultados são semelhantes aos da Figura 4.

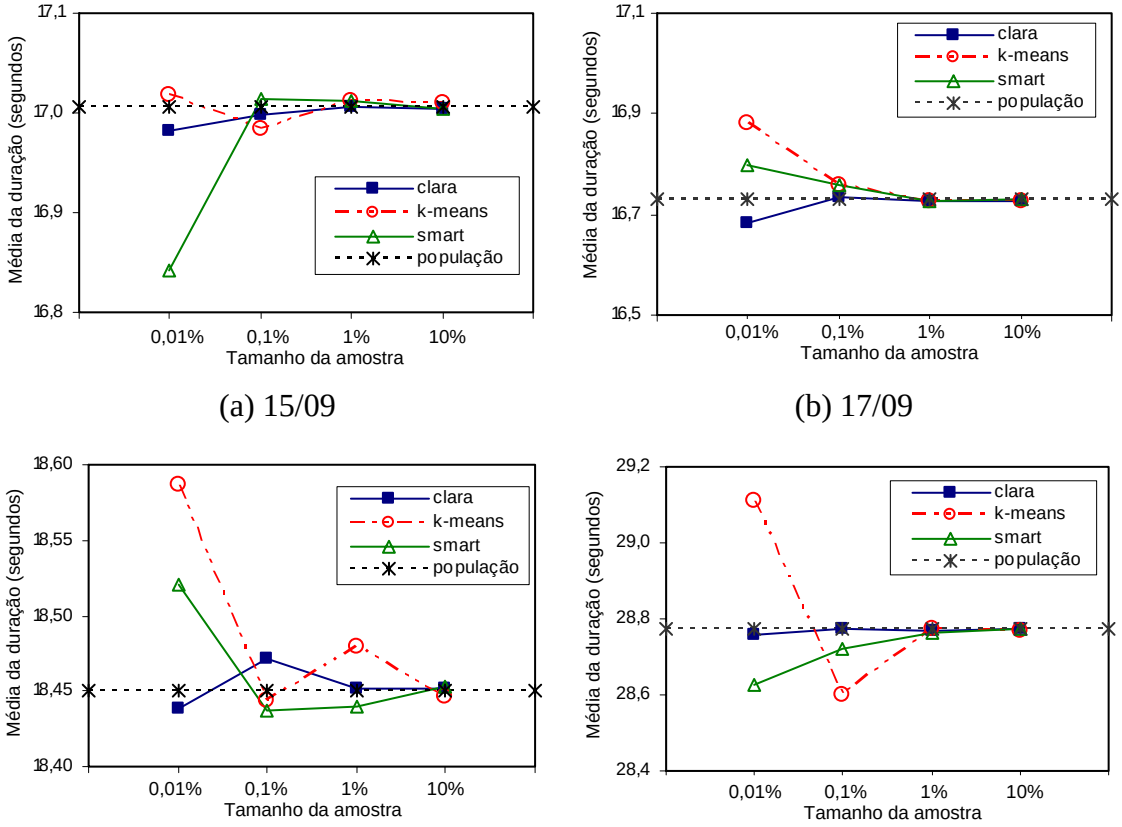
Também foram feitas comparações entre as técnicas de alocação ótima (de Neyman) e proporcional. Em [15] os autores afirmam que para amostragem de pacotes usando o tamanho como variável de estratificação, nenhuma diferença significativa foi observada em favor do método ótimo. Entretanto, neste trabalho o uso da estratificação proporcional obteve resultados insignificantes, devido à natureza de cauda pesada dos fluxos de tráfego (enquanto que pacotes têm um tamanho máximo, em geral de 1500 bytes). Os gráficos comparando amostragem ótima e proporcional não são mostrados devido à limitação de espaço neste artigo.

6.3. Amostragem Inteligente

A Figura 5 mostra os resultados da comparação da amostragem estratificada baseada nos métodos CLARA e K-Means (8 estratos) com o método de amostragem inteligente ou de limiar (aqui chamado de *Smart*) apresentado na seção 2. Três observações importantes podem ser feitas sobre a utilização do Smart. Em primeiro lugar, ele foi capaz de construir intervalos de confiança menores, que é um resultado intuitivo, uma vez que o limiar é escolhido de modo a haver a menor variabilidade nas amostras. Os intervalos não estão mostrados na Figura 5 para obter maior clareza visual nos resultados. Em segundo lugar, nos experimentos realizados o Smart não foi capaz de produzir aproximações da média da população mais precisas que a amostragem estratificada. Principalmente para tamanhos de amostra maiores que 0,01%, o método CLARA obteve precisão igual ou maior que o Smart. Na comparação com o K-Means, o Smart apresentou maior precisão.

Finalmente, para todas as condições avaliadas no trabalho apresentado neste artigo, o processamento do Smart mostrou-se extremamente lento, tanto em termos absolutos como em termos relativos, comparado à amostragem estratificada. Enquanto que para processar um determinado tamanho de amostra (ex: 1%) para um dia e cem replicações com a estratificada são necessários alguns minutos em uma máquina Pentium 4 de 3,2 GHz, o

Smart necessita de algumas horas. Isto inviabiliza a sua utilização em vários cenários, como por exemplo, na amostragem em tempo real feita em roteadores.



(a) 15/09 (b) 17/09

(a) 18/09 (b) 19/09

Figura 5 – Comparação das Amostragens Inteligente e Estratificada

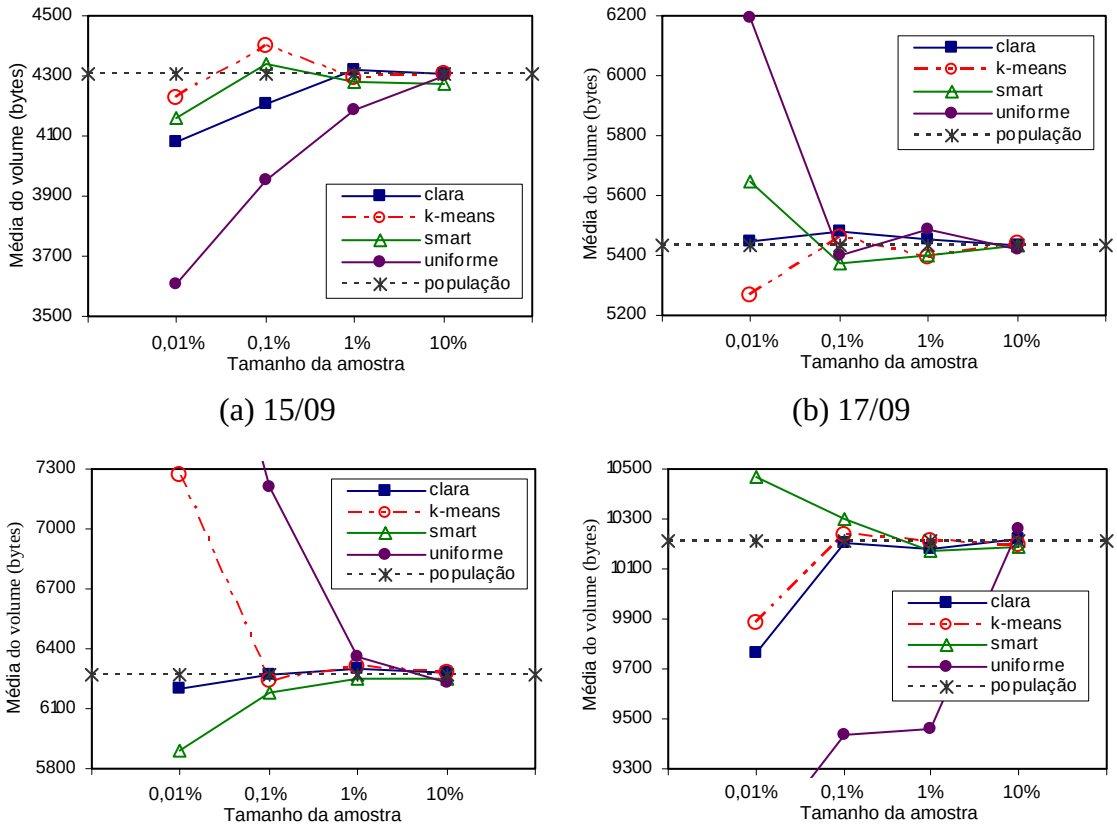
6.4. Estratificação Cruzada

As análises anteriores sempre utilizaram o tempo de duração dos fluxos como variável de observação. No entanto, mesmo que a variação de estratificação seja a duração, é desejável que se possam fazer inferências posteriores sobre outras variáveis, como o volume dos fluxos. Esta seção analisa o efeito da observação de variáveis cruzadas, ou seja, a estratificação pela duração e a observação pelo volume. Outra opção é fazer N estratificações para N variáveis de observação. No entanto, esta abordagem exige praticamente N vezes mais recursos de processamento e armazenamento.

A Figura 6 mostra a estratificação cruzada para a média do volume, comparando os métodos CLARA, K-Means (ambos com 8 estratos) e Smart com a população e a amostragem uniforme. É possível observar que a amostragem uniforme apresenta um resultado pobre, principalmente com tamanhos de amostra menores que 1%. Em geral, apenas com amostra de tamanho 10% a amostragem uniforme é capaz de gerar um viés inferior a 1%. Por outro lado, o pior caso foi observado para amostra 0,01% do dia 18/09, com viés de 41,3%.

Na comparação dos outros métodos, não há uma distinção inequívoca entre os vieses produzidos por eles. O K-Means, que teve resultados inferiores ao CLARA na comparação

da estratificação direta, em alguns casos apresenta os melhores resultados. Uma conclusão importante para os cenários avaliados é a adequação de quaisquer desses três métodos para obter resultados significativos para estratificação cruzada, desde que não sejam utilizadas amostras de tamanho 0,01% da população.



(a) 15/09 (b) 17/09

(a) 18/09 (b) 19/09

Figura 6 – Análise cruzada: estratificação → duração e observação → volume

7. Conclusão

Este trabalho fornece subsídios para aumentar o conhecimento existente em técnicas de tratamento de informações de fluxos na Internet. Particularmente, técnicas de análise de agrupamento e sua aplicação em amostragem estratificada de fluxos de tráfego são apresentadas e comparadas. Os métodos de análise de agrupamento CLARA e K-Means são apresentados e os grupos produzidos por eles em arquivos de registros de fluxos são utilizados como limites para a amostragem estratificada.

Os resultados mostram claramente que os algoritmos CLARA e K-Means podem ser usados eficientemente em amostragem estratificada, embora o CLARA apresente resultados mais precisos, devido à sua capacidade de gerar agrupamentos mais bem distribuídos. Avaliações subseqüentes mostram que existe um compromisso explícito entre tamanho de amostra, número de estratos e precisão nos resultados. Em geral, quanto maiores os requisitos de processamento e armazenamento, maior a precisão obtida. Na comparação com o método de amostragem inteligente (Smart), a amostragem estratificada com CLARA ou K-Means se mostrou equivalente, embora com um tempo de processamento significativamente menor.

Como trabalhos futuros será investigada a viabilidade do uso de técnicas de agrupamento e amostragem estratificada para processamento em tempo real, para ser utilizada em qualquer infra-estrutura de medição. Além disso, novas aplicações para essas técnicas no contexto da análise de fluxos na Internet estão sendo identificadas.

8. Referências

[1] Cocharan, William G., *Sampling Techniques*, 3ª ed. New York: John Willey, 1977.

[2] Duffield, N., Lund, C., and Thorup, M., "Learn more, sample less: control of volume and variance in network measurement", *IEEE Transactions in Information Theory*, vol. 51, no. 5, pp. 1756-1775, 2005.

[3] Duffield, N., Lund, C., and Thorup, M.,"Flow Sampling Under Hard Resource Constraints", *ACM SIGMETRICS, 2004.*

[4] Hernández-Campos, F., Nobel, A.B., Smith, F.D. and Jeffay, K., "Understanding Patterns of TCP Connection Usage with Statistical Clustering", *Proceedings of the 13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, Atlanta, GA, September 2005, pages 35-44.

[5] Jain, A.K., Murty, M.N., and Flynn, P.J., "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, September 1999

[6] Kamienski, C., Fernandes, S. & Sadok, D., "Characterizing Essential Properties of Traffic Flows using Stratified Sampling", submetido para o IEEE JSAC edição especial em amostragem, Outubro 2005.

[7] Kamienski, C., Souza, T., Fernandes, S., Silvestre, G. & Sadok, D., "Caracterizando Propriedades Essenciais do Tráfego de Redes através de técnicas de Amostragem Estratificada", SBRC 2005, Maio 2005.

[8] Kompella, R. R. and Estan, C., "The Power of Slicing in Internet Flow Measurement", *Internet Measurement Conference*, October 2005

[9] Laiho, J.; Raivio, K.; Lehtimaki, P.; Hatonen, K.; Simula, O., "Advanced Analysis Methods for 3G cellular networks," *Wireless Communications, IEEE Transactions on*, vol.4, no.3pp. 930- 942, Maio 2005

[10] Lakhina, A., Crovella, M., and Diot, C., "Mining anomalies using traffic feature distributions", Computer Communication Review, Volume 35, Issue 4 (October 2005), Pages: 217 – 228, 2005.

[11] Papayanaki, K., Taft, N., and Diot, C. "Impact of Flow Dynamics on Traffic Engineering Design Principles". *IEEE INFOCOM 2004*. 7-11 March 2004. Hong-Kong.

[12] R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, *http://www.r-project.org.*

[13] Silvestre, G., Kamienski, C., Fernandes, S., Mariz, D., and Sadok, D., "Análise de Tráfego Peer-to-Peer baseada na Carga Útil dos Pacotes", Relatório Técnico – GPRT/UFPE, 2005.

[14] Xian, Ji-Qing, Lang, Feng-Hua, and Tang, Xian-Lun, "A Novel Intrusion Detection Method Based on Clonal Selection Clustering Algorithm", *Proc. of the 4th International Conference on Machine Learning and Cybernetics*, Guangzhou, 18-21 Agosto 2005.

[15] Zseby, T., "Stratification Strategies for Sampling-based Non-intrusive Measurements of One-way Delay", *Passive and Active Measurement Workshop Proceedings*, April 2003.