Um Framework para Dimensionar o Total Populacional em Redes Compostas por População Rara e Agrupada *

Camila D. da Silva[1] **e Antonio A. de A. Rocha** [1,*]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brasil

camiladomingos@id.uff.br e arocha@ic.uff.br

Abstract. *Dimensioning the total population in networks whose structure is composed of a rare and clustered population is not a trivial task. In these circumstances, it is usual to overlay a grid over the region in which the population is contained and select cells from that grid. A framework called the 2-Layer and 2-Estimator Method - 2L2EM was proposed and it implements the Multiple Capture and Recapture Method - MCRM in layer 1 to obtain the estimates within the cell to be used as input for the Adaptive Cluster Sampling - ACS which is layer 2 in which the total population in the network is estimated. The studies with synthetic and real data reveal that 2L2EM provides relevant estimates in relation to ACS and MCRM.*

Resumo. *Dimensionar o total populacional em redes cuja estrutura é composta por uma população rara e agrupada não é uma tarefa trivial. Neste cenário, é usual sobrepor uma grade à região na qual a população está contida e selecionar células dessa grade. Um framework chamado de Método 2-Camadas e 2-Estimadores - M2C2E foi proposto e ele implementa o Método de Captura e Recaptura Múltipla - MCRM na camada 1 para gerar as estimativas dentro da célula a serem usadas como entrada para a Amostragem Adaptativa por Conglomerados - AAC que é a camada 2 na qual estima-se o total populacional na grade. Os estudos com dados sintéticos e com dados reais revelam que o M2C2E fornece estimativas relevantes em relação à AAC e ao MCRM.*

1. Introdução

A necessidade de quantificar o número de elementos da população é um problema comum em redes de computadores e sistemas distribuídos [Accettura et al. 2015, Peng et al. 2009]. Contudo, saber o total populacional em uma rede pode necessitar de métodos amostrais complexos, em particular quando a estrutura da população venha a ser formada por elementos raros e agrupados. Algumas deficiências em sistemas computacionais podem ser caracterizadas por populações raras e agrupadas, tais como falhas em arquiteturas de sistemas distribuídos. Esse e outros exemplos justificam a importância de se estimar o tamanho de populações raras e agrupadas, visando, por exemplo, solucionar os problemas antes de chegarem a todos os usuários da rede ou mesmo determinar de antemão a magnitude deles.

A Amostragem Adaptativa por Conglomerados - AAC foi proposta por [Thompson 1990] para estimar o total populacional em uma área sobre o contexto de populações raras e agrupadas. O desenho amostral da AAC consiste em sobrepor

*Esse trabalho foi parcialmente apoiado pela FAPESP (processo 2015/24144-7).

uma grade à área que contém a população de interesse e amostrar células dessa grade. Entretanto a AAC leva em consideração a possibilidade de observar todos os elementos de interesse dentro de cada célula amostrada, o que pode não ser viável para todos os cenários. Uma revisão sobre os principais desenvolvimentos na AAC foi apresentada por [Turk and Borkowski 2005], os quais comentaram sobre a condição C, a qual refere-se ao número total de elementos com a característica de interesse dentro da célula, ser difícil ou impossível de determinar em situações reais. Essa condição C tornou-se a principal questão de pesquisa a ser desenvolvida ao longo deste artigo, levando à implementação do framework proposto.

O Método de Captura e Recaptura Múltipla - MCRM foi introduzido por [Schnabel 1938], através do estimador de Schnabel - ES, com a finalidade de estimar o total populacional. O problema que surge quando pensa-se em múltiplas recapturas são dois: (i) nem sempre é possível satisfazer a premissa básica do método de que os elementos da população encontram-se uniformemente distribuídos, durante todo o processo de captura. Por exemplo, quando a população está distribuída de forma desigual na região de estudo; e (ii) não foi possível evidenciar um estudo que defina o critério de parada apropriado para o processo de recaptura.

Neste artigo, é proposto um framework chamado de Método 2-Camadas e 2-Estimadores - M2C2E que soluciona a lacuna da AAC aplicando o estimador de Schnabel do MCRM para estimar o total populacional dentro da célula selecionada, fazendo uso do critério de parada proposto por [Singham 2010], e posteriormente, utiliza-se um dos estimadores usuais da AAC modificado para determinar o total populacional da grade. A vantagem do M2C2E em relação ao AAC está na não obrigatoriedade de conhecer o total populacional dentro de cada célula da grade. O framework proposto foi utilizado para resolver um problema real de estimar o total de táxis no município do Rio de Janeiro com motoristas conectados a um aplicativo de transporte no celular. Embora, o estudo de caso apresentado seja de um tipo específico de aplicação, é possível imaginar que a solução se aplica também a outros aplicativos de celular, além de outros sistemas e aplicações distribuídas, por exemplo, dispositivos de sensores em operação em uma determinada região de monitoramento [Peng et al. 2009].

As principais contribuições do ponto de vista teórico são: (i) adaptação do critério de parada proposto por [Singham 2010] ao contexto do MCRM; (ii) implementação do framework M2C2E o qual visa contornar a lacuna da AAC ao estimar dentro das células; e, (iii) modificação dos estimadores clássicos de Horvitz-Thompson e Hansen-Hurwitz visando o M2C2E. Estas contribuições são descritas na Seção 4 deste artigo. Antes disso, são apresentadas as fundamentações teóricas da AAC e do MCRM, com seus respectivos estimadores de total populacional e os critérios de parada, nas Seções 2 e 3, respectivamente. A validação com dados sintéticos e o estudo de caso com dados reais gerados a partir do acesso dos motoristas de táxi a um aplicativo são apresentados na Seção 5. Por fim, as considerações finais são apresentadas na Seção 6.

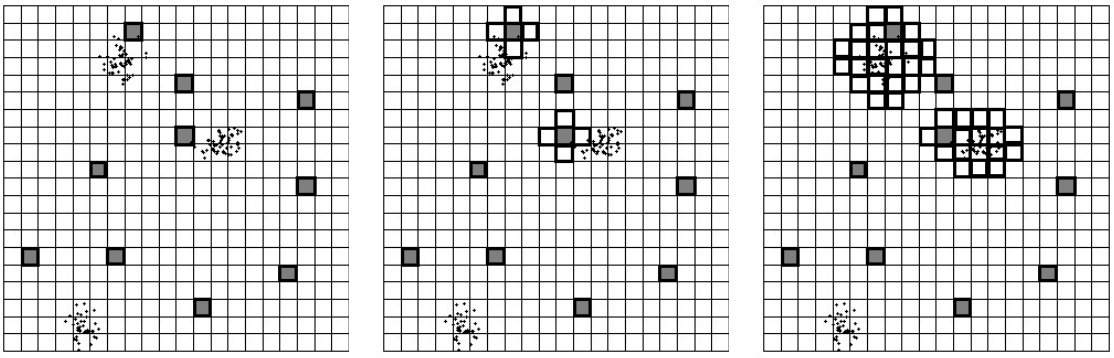
2. Amostragem Adaptativa por Conglomerados

A Amostragem Adaptativa por Conglomerados - AAC foi proposta por [Thompson 1990] e é considerada eficiente para populações raras e agrupadas. A AAC inicia-se com uma amostra ao acaso de z_1 unidades da grade, as quais são chamadas de células e todas têm

a mesma probabilidade de serem incluídas na amostra. No momento em que um ou mais elementos da população são encontrados dentro da célula sorteada, a metodologia da AAC orienta contabilizar todos os elementos dela e das células que compartilham lados com a célula i (direita, esquerda, em cima e embaixo), até chegar às células que não contenham nenhum elemento da população de interesse. A rede é o conjunto formado por células nas quais os elementos de interesse podem ser encontrados. As bordas são células que não contêm elementos de interesse, portanto elas não atendem à condição de interesse, mas são vizinhas de células que satisfazem à condição. O conglomerado é a união das células de rede e das células de borda.

A Figura 1 ilustra a metodologia da AAC, assumindo que os elementos de interesse foram sobrepostos por uma grade. Os passos posteriores são apresentados nas seguintes etapas:

1. Extrair uma amostra aleatória simples com reposição - AASc ou sem reposição - AASs de z_1 células da grade. Essa etapa está representada na Figura 1(a) pelos $z_1 = 10$ quadrados em cinza;
2. Observa-se dentre as células amostradas no passo anterior as que atendem a uma condição C, de forma que $C = \{y|y_i > 0\}$, onde y_i é o número de elementos de interesse na célula i, ou seja, é o parâmetro populacional. No caso favorável, as células vizinhas que compartilham lado com a célula i serão adicionadas à amostra. Note que, na Figura 1(a) duas células satisfizeram a condição C, então apenas as células vizinhas a elas foram observadas, como mostra a Figura 1(b);
3. O processo de observar células vizinhas continuará até chegar às células que não satisfazem a condição de interesse, conforme ilustrado na Figura 1(c). Portanto, a AAC utiliza sua própria estrutura para finalizar o processo, pois o critério de parada está na chegada às células de borda.



(a) Etapa 1 **(b)** Etapa 2 **(c)** Etapa 3

Figura 1. Ilustração da metodologia da AAC em uma população rara e agrupada gerada e sobreposta por uma grade com 400 células.

Seja n a variável que representa o número total de células visitadas ao final do processo da AAC. Uma célula i pode ser incluída na amostra: (i) se ela for selecionada na amostra inicial z_1; (ii) se ela for qualquer célula da rede; ou, (iii) se ela for qualquer célula de borda. É possível que dois conglomerados compartilhem uma ou mais células de borda. Ou ainda, tem-se que a seleção de células na Etapa 1 podem levar à inclusão de uma mesma rede na amostra final. Seja m_i o número de células na rede em que a célula i

pertence incluindo i, no caso em que i não atende ao critério de interesse, então $m_i = 1$ será chamada de uma rede de tamanho 1. Seja a_i o número total de células na rede em que i é uma célula de borda, no caso de i satisfazer a condição de interesse, então $a_i = 0$.

Considere a probabilidade de seleção da célula i partindo de qualquer uma das z_1 células iniciais como sendo a razão entre o número de células que, se selecionadas, levam a i estar na amostra e o número de células na grade. Tal probabilidade é dada por:

$$p_i = \frac{m_i + a_i}{N}, \tag{1}$$

em que N é o número total de células construídas na grade em $\mathbb{R}^2$. Também importante é a probabilidade de inclusão da célula i na amostra através da AASs, dada por:

$$\pi_i = 1 - P(\{i \text{ não estar incluída entre as } n\}) = 1 - \left[\begin{pmatrix} N - m_i - a_i \\ n \end{pmatrix} \Big/ \begin{pmatrix} N \\ n \end{pmatrix} \right].$$

Se as células iniciais z_1 forem selecionadas através de uma amostragem aleatória simples com reposição na qual todas as células têm a mesma probabilidade de serem selecionadas, a probabilidade de seleção será a mesma conforme descrito na Equação 1. Contudo, a probabilidade de inclusão da célula i dentro da AASc é dada pela seguinte expressão:

$$\pi_i = 1 - \frac{(N - m_i - a_i)^n}{N^n} = 1 - \left(1 - \frac{m_i + a_i}{N}\right)^n = 1 - (1 - p_i)^n.$$

A AAC leva em consideração, para desenvolver sua metodologia de estimação, esquemas probabilísticos desiguais. Isso ocorre quando o processo deixa de adicionar células através de uma AAS e passa a considerar as células vizinhas com maior probabilidade para entrarem na amostra do que as demais.

2.1. Estimador de Horvitz-Thompson

O estimador Horvitz-Thompson - HT é um estimador não tendencioso do total populacional com probabilidades desiguais de seleção e com amostras iniciais selecionadas por uma AAS sem reposição. Segundo [Horvitz and Thompson 1952], a expressão do estimador de Horvitz-Thompson do total populacional é dada por:

$$\widehat{\tau}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i}, \tag{2}$$

na qual y_i é o número total de elementos de interesse na célula i e π_i é a probabilidade de inclusão da célula i na amostra usando a AASs.

2.2. Estimador de Hansen-Hurwitz

Uma alternativa ao caso em que a amostra inicial seja extraída por uma AAS com reposição foi estudada por [Hansen and Hurwitz 1943] e recebeu o nome de estimador Hansen-Hurwitz - HH. Esse estimador é não tendencioso para o total da população e é possível ver sua expressão a seguir:

$$\widehat{\tau}_{HH} = \frac{1}{n} \sum_{i=1}^{N} \frac{y_i}{p_i}, \tag{3}$$

sendo y_i o número total de elementos de interesse na célula i e p_i a probabilidade de seleção da i-ésima célula da grade, para $i = 1, \ldots, N$.

3. Método de Captura e Recaptura Múltipla

O Método de Captura e Recaptura Múltipla - MCRM foi proposto por [Schnabel 1938], visando a extração de um número maior que duas amostras independentes para obter estimativas mais precisas sobre o total populacional. Por outro lado, realizar múltiplas recapturas pode ser uma tarefa bastante difícil por depender da capacidade de acesso à população, por aumentar os custos de pesquisa e pelo aumento do tempo computacional. Além disso, há ainda uma premissa básica no método que prevê uma probabilidade uniforme de capturar e recapturar os elementos em toda região de análise, o que pode não ser válida para uma região muito extensa com população rara e agrupada.

Esse método é composto por um vetor C_1, que representa a captura inicial de elementos e pelo conjunto de vetores $R = \{R_1, R_2, R_3, ..., R_k\}$ que denotam as k recapturas. Define-se, então, o conjunto de vetores $M = \{M_1, M_2, ..., M_k\}$ usados para contabilizar os novos elementos que aparecem a cada recaptura, além das seguintes notações: número total de elementos da população (N); número máximo de recapturas (k); número de amostras coletadas (j), onde $j = \{1, 2, ..., k\}$; número de elementos a cada amostra j (n_j); e, número de elementos recapturados na amostra j dentre os n_j elementos (m_j).

3.1. Estimador de Schnabel

O estimador de Schnabel - ES foi desenvolvido por [Schnabel 1938] e é utilizado no contexto do MCRM para população fechada, ou seja, sem "nascimentos", "mortes", imigração e emigração, com a finalidade de obter a estimativa do total populacional. É possível obter a expressão do estimador de Schnabel por:

$$\widehat{N}_{schn} = \frac{\sum_{j=2}^{k} n_j M_j}{\sum_{j=2}^{k} m_j}, \tag{4}$$

sendo $u_j = n_j - m_j$ e $M_1 = 0$, $M_j = \sum_{j=2}^{k} u_{j-1}$ é número total de elementos novos marcados na população imediatamente antes da próxima amostra ter sido recolhida para $j = \{2, ..., k\}$.

3.2. Critérios de Parada para o MCRM

Os critérios de parada têm sido usados para fornecer auxílio à tomada de decisão. A escolha incorreta do critério de parada pode adicionar viés aos resultados das estimativas como discutido por [Dalal and Mallows 1990], os quais apontaram para o uso de um modelo estocástico ou uma regressão para estimar o número total de falhas em um software ao longo do tempo. No entanto, nenhuma dessas abordagens responderam à questão central de qual era o melhor momento em que o teste deve ser interrompido e adicionaram duas hipóteses sobre critério de parada, são elas: "Se o teste parar muito cedo, muitas falhas permanecem. Portanto, haverá custos de correção e perda de mercado, devido à insatisfação dos clientes; e se o teste continuar até o limite máximo permitido, existe o custo do esforço de teste elevado".

[El Emam and Laitenberger 2001] afirmaram que, o momento de parar as inspeções é um problema e que os estimadores do Método de Captura e Recaptura não são precisos para decidir se devem parar ou voltar a inspecionar. Os autores ainda adicionaram que as organizações precisam definir seus limites de eficácia para suas inspeções

e apresentaram um caso particular aplicado a inspeções de software em que, a decisão de parada da inspeção ocorre quando é excedido um determinado limite superior ou interior. [Smith 1988] também apontou a dificuldade em determinar em qual ocasião as recapturas devem cessar, mas ainda que plotando função de risco e ganhos para a estimativa, não concluiu sobre critério de parada.

Uma abordagem promissora é usar a meia largura do Intervalo de Confiança - IC como critério de parada, conforme proposto em [Singham 2010]. Seja $X = \{X_1, X_2, ..., X_k\}$ uma amostra de tamanho k do total populacional. Seja $\bar{X}$ a estimativa média do total populacional e S_k^2 a sua variância amostral. Considere η como sendo o coeficiente de confiança e assumindo que $t_{\eta,k-1}$ seja $(1+\eta)/2$ o quantil da distribuição T-Student com $k-1$ graus de liberdade. Tem-se que o IC para a média do total populacional é dado por $\bar{X} \pm t_{\eta,k-1}\sqrt{\frac{S_k^2}{k}}$. Define-se $MK_{\eta,k}$ como a meia largura do IC:

$$MK_{\eta,k} = t_{\eta,k-1}\sqrt{\frac{S_k^2}{k}}. \tag{5}$$

Seja δ o valor máximo da meia largura do IC desejado, portanto valores pequenos de δ implicam intervalos de confiança mais estreitos. O critério de parada proposto em [Singham 2010] define que, para um determinado valor de precisão desejado, a parada ocorre quando a meia largura do IC (variável $MK_{\eta,k}$) for menor ou igual a δ. Então, define-se k^* como a iteração na qual ocorre a parada de um procedimento sequencial.

$$k^* = \arg\min_{k>0} MK_{\eta,k} \leq \delta, \tag{6}$$

onde $\arg\min$ é definido como o primeiro valor de $k \in \mathbb{N}$ tal que $MK_{\eta,k} \leq \delta$.

Usando a Inequação 6, observa-se que o critério de parada para o MCRM pode ser definido por k^* como o primeiro valor de recaptura no qual a meia largura do IC atinge o valor máximo para δ. Considere $\widehat{n}^*$ como a estimativa obtida na k^*-ésima recaptura pelo estimador definido na Equação 4. A seguir, são descritas as etapas para implementação do critério de parada com a garantia de confiança η baseada na Inequação 6 para o cenário do MCRM, são elas:

1. Escolha um valor para o coeficiente de confiança η, um valor para o erro δ e faça a primeira captura de elementos.
2. Inicie a recaptura ($k = 1$) e obtenha a estimativa $\widehat{n}_1$.
3. Realize a k-ésima recaptura, para $k = \{2, ..., k^*\}$, obtenha a estimativa $\widehat{n}_k$ e calcule a média, variância e meia largura das estimativas obtidas até k, inclusive.
4. Se $MK_{\eta,k} > \delta$, realize a próxima recaptura $k = k + 1$ na etapa 3. Caso contrário, vá para a etapa 5.
5. Quando $MK_{\eta,k} \leq \delta$, pare e entregue o valor de k^* e a estimativa $\widehat{n}^*$ na k^*-ésima recaptura.

Neste artigo, a adequação da proposta de [Singham 2010] ao contexto do MCRM foi possível devido ao valor de δ poder ser construído a partir da primeira estimativa diferente de zero dos estimadores usuais do MCRM, tal como o estimador de Schnabel:

$$\delta = \widehat{n}_1\varepsilon, \tag{7}$$

onde $\widehat{n}_1$ é a primeira estimativa diferente de zero e ε é erro.

A Equação 7 possibilita utilizar a proposta de [Singham 2010] para diferentes tamanhos de população, pois δ varia de acordo com o produto de uma variável aleatória e não predeterminada ($\widehat{n}_1$) com uma variável predeterminada (ε).

4. Framework M2C2E

A partir da observação de uma premissa na metodologia da AAC na qual todos os elementos dentro da célula deveriam ser encontrados, foi proposto um framework chamado de Método 2-Camadas e 2-Estimadores (ou simplesmente, M2C2E) para estimar o número total de elementos dentro da célula utilizando o MCRM e, posteriormente, essas estimativas são incluídas aos estimadores usuais da AAC modificado. O objetivo é solucionar o caso no qual não é possível encontrar todos os elementos dentro da célula (como requerido da AAC), devido à grande dificuldade de captura, por exemplo, por estarem escondidos ou misturados a outra população com grande número de elementos.

O framework proposto, como o próprio nome sugere, é um processo que contém duas camadas e dois estimadores. Inicia-se de forma semelhante à AAC, sobrepondo uma grade à área a ser estudada cuja variável de interesse esteja presente. Posteriormente, seleciona-se z_1 células da grade e aplica-se o MCRM dentro de cada célula para estimar o número total de elementos dentro da célula i selecionada.

A camada 1 contém as células inicialmente selecionadas usando uma amostragem aleatória simples com ou sem reposição, nelas aplica-se o MCRM e para realizar esse procedimento é necessário usar um estimador do MCRM, por exemplo, o estimador de Schnabel. Vale ressaltar que é preciso definir o critério de parada para obter as estimativas nas células por MCRM. Esse framework, em particular, foi construído a partir do critério proposto por [Singham 2010]. O processo de estimação dentro das células é repetido para as células vizinhas nos casos em que existam elementos de interesse, até chegar às células de borda nas quais o processo de estimação por MCRM será finalizado.

As estimativas da camada 1, para o total populacional dentro da célula, são levadas como entrada para a camada 2, na qual o objetivo é estimar o total populacional na grade. Sendo assim, aplica-se um dos estimadores modificados da AAC (i.e., estimador HT_mod ou HH_mod) conforme descrito na Subseção 4.1. Na camada 2, ocorre o final do processo resultando na estimação do total populacional na grade.

Uma vez que esse framework proposto agrega o MCRM com a AAC, a notação é semelhante a ambos os processos, com exceção para a notação do tamanho da grade, pois na AAC é utilizada a variável N, ao passo que no MCRM N representa o total populacional. Portanto, visto que a AAC finaliza o M2C2E, define-se a variável N como o tamanho da grade e a variável τ como o total populacional na grade no M2C2E. Além disso, tem-se o número total de células selecionadas ao final da AAC (n), o tamanho inicial da amostra de células (z_1) e número de elementos a cada amostra j (n_j), $j = \{1, ..., k\}$.

4.1. Estimadores Modificados

Seja $\widehat{n}^*_{schn\_i}$ a estimativa de Schnabel na k^*-ésima recaptura para a célula i, onde a variável k^* é obtida pelo critério de parada proposto por [Singham 2010] representado pela Inequação 6. Essa estimativa é fornecida através das informações trazidas pelas amostras coletadas até a k^*-ésima recaptura ao estimador de Schnabel na Equação 4.

Proposição 1: No estimador de Horvitz-Thompson modificado para o total da população ($\widehat{\tau}_{HT\_mod}$), tem-se o parâmetro y_i que significa o número total de elementos na célula i, é substituído por $\widehat{n}^*_{schn\_i}$. De acordo com a Equação 8, $\widehat{\tau}_{HT\_mod}$ será não tendencioso para τ, se e somente $\widehat{n}^*_{schn\_i}$ convergir para y_i.

$$\widehat{\tau}_{HT\_mod} = \sum_{i=1}^{N} \frac{\widehat{n}^*_{schn\_i}}{\pi_i}, \tag{8}$$

onde π_i é a probabilidade de inclusão da célula i na amostra.

Prova: Seja I_i a variável indicadora de inclusão da célula i na amostra que segue uma distribuição Bernoulli com $E(I_i) = \pi_i$. Sendo $\widehat{n}^*_{schn\_i}$ a estimativa de Schnabel no critério de parada para cada célula i, portanto $\widehat{n}^*_{schn\_i}$ é uma constante e I_i como variável aleatória com distribuição Bernoulli:

$$E(\widehat{\tau}_{HT\_mod}) = E\left(\sum_{i=1}^{N} \frac{I_i \widehat{n}^*_{schn\_i}}{\pi_i}\right) = \frac{1}{N\pi_i} E\left(\sum_{i=1}^{N} I_i\right) \sum_{i=1}^{N} \widehat{n}^*_{schn\_i} = \sum_{i=1}^{N} \widehat{n}^*_{schn\_i}.$$

como prova da Proposição 1.

Proposição 2: O estimador de Hansen-Hurwitz modificado para o total da população ($\widehat{\tau}_{HH\_mod}$) é dado pela Equação 9 e será um estimador não tendencioso para τ, se e somente se $\widehat{n}^*_{schn\_i}$ convergir para y_i.

$$\widehat{\tau}_{HH\_mod} = \frac{1}{n} \sum_{i=1}^{N} \frac{\widehat{n}^*_{schn\_i}}{p_i}, \tag{9}$$

sendo p_i a probabilidade de seleção da i-ésima célula da população, para $i = 1, \ldots, N$.

Prova: Seja W_i o número de vezes que a i-ésima célula da grade aparece no estimador. Tem-se que a variável W_i segue uma distribuição Binomial com $E(W_i) = np_i$.

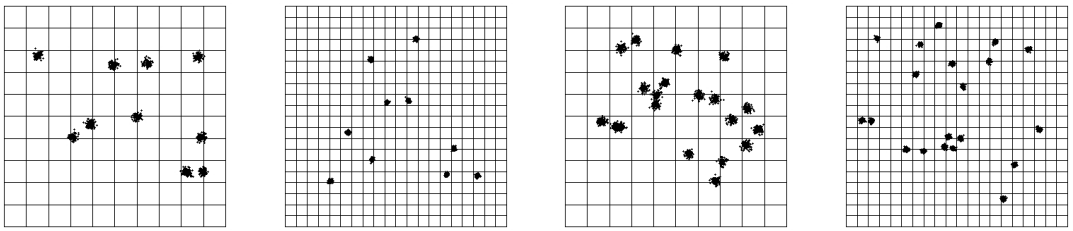
$$E(\widehat{\tau}_{HH\_mod}) = E\left(\frac{1}{n} \sum_{i=1}^{N} \frac{W_i \widehat{n}^*_{schn\_i}}{p_i}\right) = \frac{1}{n \cdot (Np_i)} E\left(\sum_{i=1}^{N} W_i\right) \sum_{i=1}^{N} \widehat{n}^*_{schn\_i} = \sum_{i=1}^{N} \widehat{n}^*_{schn\_i}.$$

como prova da Proposição 2.

4.2. Etapas do M2C2E

Partindo do princípio que os elementos de interesse foram sobrepostos por uma grade e cada célula tem o valor de k^* de forma independente das demais, as etapas do M2C2E são apresentadas a seguir:

1. Selecionar z_1 células da grade a partir de uma AASs ou AASc.
2. Se houver elementos de interesse nas células selecionadas na etapa 1, as células que compartilham lados com elas são observadas, aplicar nelas o estimador Schnabel até chegar à k^*-ésima recaptura e obter as estimativas $\widehat{n}^*_{schn\_i}$.
3. Adicionar as células vizinhas das vizinhas que contêm elementos de interesse, novamente aplicar o estimador Schnabel até chegar à k^*-ésima recaptura e obter as estimativas $\widehat{n}^*_{schn\_i}$ para todas elas até chegar em células vizinhas que não contêm elementos de interesse.
4. Inserir $\widehat{n}^*_{schn\_i}$ ao invés de y_i ao estimador de Horvitz-Thompson na Equação 2 ou ao estimador de Hansen-Hurwitz na Equação 3 e retornar a estimativa para o total na grade usando $\widehat{\tau}_{HT\_mod}$ ou $\widehat{\tau}_{HH\_mod}$ de acordo com as Equações 8 e 9.



(a) Caso 1 **(b)** Caso 2 **(c)** Caso 3 **(d)** Caso 4

Figura 2. Populações sintéticas: caso 1 - $\tau = 1000$ **e** $N = 100$**; caso 2 -** $\tau = 1000$ **e** $N = 400$**; caso 3 -** $\tau = 2000$ **e** $N = 100$**; e caso 4 -** $\tau = 2000$ **e** $N = 400$**.**

5. Avaliação Experimental do Framework

Com a finalidade de analisar o framework proposto, foram realizados experimentos com populações geradas sinteticamente e com dados reais no software RStudio.

5.1. Validação com Dados Sintéticos

Na verificação para saber se o M2C2E pode ser uma solução à premissa da AAC, quatro exemplos de populações raras e agrupadas dispostas em uma região, vistas na Figura 2, foram consideradas. Em particular, na Figura 2, os pontos foram gerados a partir de uma distribuição Normal Bivariada com matriz de variância cuja diagonal principal foi fixada em $\phi = 0,1$ a qual representa o grau de agrupamento entre os pontos.

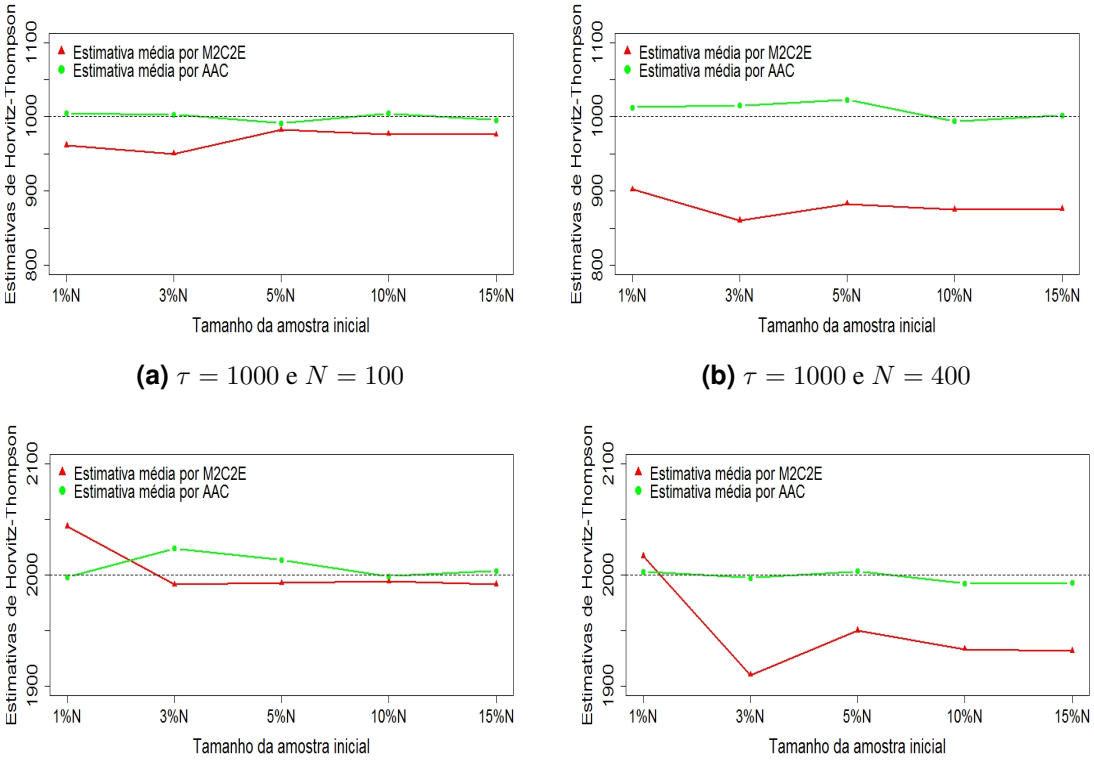
Foram realizadas 10 mil replicações para cada valor de $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$, para 2 diferentes refinamentos de grade $N = 100$ e $N = 400$ e 2 diferentes números de elementos na rede $\tau = 1000$ e $\tau = 2000$, totalizando 400 mil rodadas (200 mil M2C2E e 200 mil AAC). A Figura 3 contém as estimativas médias da AAC (com HT) e do M2C2E (com ES na camada 1 e HT na camada 2) após 10 mil replicações de cada método e para cada z_1 inicial. Nos casos em que $\tau = 1000$ e $N = 100$ (Figura 3(a)) e $\tau = 2000$ e $N = 100$ (Figura 3(c)), as estimativas estão visualmente mais próximas da linha pontilhada que representa o total populacional verdadeiro. Pela natureza do método, os resultados do M2C2E são limitados superiormente pelas estimativas da AAC, uma vez que a AAC tem como premissa a possibilidade de determinar toda população de cada célula.

O teste de hipótese T^2 Hotelling é um teste paramétrico multivariado para testar a igualdade entre os vetores de estimativas médias $\vec{\mu}$. O pré-requisito para realizar esse teste é a normalidade dos dados o qual pode ser verificado pelo Teorema do Limite Central válido a este cenário. Sob as hipóteses $\vec{\mu}_{M2C2E_{HT\_mod}} = \vec{\mu}_{AAC_{HT}}$ versus $\vec{\mu}_{M2C2E_{HT\_mod}} \neq \vec{\mu}_{AAC_{HT}}$, a única população na Figura 2 que não rejeitou a hipótese de igualdade entre os vetores de estimativas médias foi a $\tau = 2000$ e $N = 100$, ao nível de significância de 5%, com p-valor de 0,66. Isto significa que, as estimativas médias da AAC e do M2C2E para essa população não podem ser consideradas diferentes.

Tal fato indica que o M2C2E ajusta-se melhor em grade com menor número de células e maiores valores de τ. Ao utilizar o MCRM conjuntamente ao critério de parada proposto por [Singham 2010] com erros de 2% e 5% para as populações nas Figuras 2(a) e 2(c), a parada ocorreu na 9ª recaptura, retornando a estimativa de $\widehat{n}^*_{schn} = 921$ elementos sendo que $N = 1000$ e na 15ª recaptura com estimativa de Schnabel $\widehat{n}^*_{schn} = 1901$ elementos, onde $N = 2000$ elementos.

5.2. Estudo de Caso com Dados Reais

Os dados reais utilizados para validação do M2C2E conta com 110369 coordenadas de táxis ao longo do dia 22 de junho de 2016 fornecidas pela empresa prestadora de



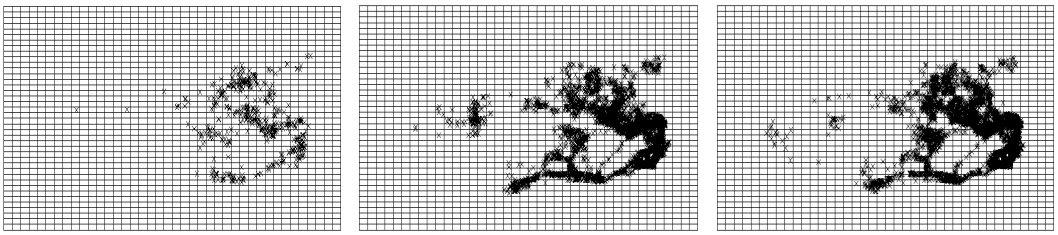
(a) $\tau = 1000$ e $N = 100$ **(b)** $\tau = 1000$ e $N = 400$

(c) $\tau = 2000$ e $N = 100$ **(d)** $\tau = 2000$ e $N = 400$

Figura 3. Estimativas médias do tamanho das populações sintéticas para diferentes valores de τ e N usando a AAC e o M2C2E.

serviço de táxis por aplicativo no município do Rio de Janeiro. Uma vez que existem concentrações de táxis em região com maior renda e em locais como aeroportos, shoppings e rodoviárias, i.e., a presença de táxis varia em relação à localização do município.

Portanto, procura-se trazer um exemplo de população rara e agrupada, em um sistema distribuído, conectada por uma rede móvel a um aplicativo de transporte. Além de observar que os táxis são distribuídos de forma desigual em determinados bairros pelas causas mencionadas, suas frequências variam em função do horário do dia, conforme ilustrado na Figura 4 na qual os "x" representam as coordenadas geográficas dos táxis no respectivo horário.



(a) 08:00hs **(b)** 16:00hs **(c)** 24:00hs

Figura 4. Distribuição espacial da população de táxis no dia 22 de junho de 2016 no município do Rio de Janeiro sobreposta por grade 40x40.

Foram utilizadas duas configurações com $n_1 = n_2 = ... = n_{k^*} = 100$ para determinar o total populacional de táxis por hora: a primeira configuração visa implementar o estimador de Schnabel na camada 1 e o estimador HT modificado na camada 2, variando $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$; e a segunda configuração em relação à primeira muda o estimador utilizado na camada 2, tem-se o estimador de Schnabel na camada 1 e o estimador HH modificado na camada 2.

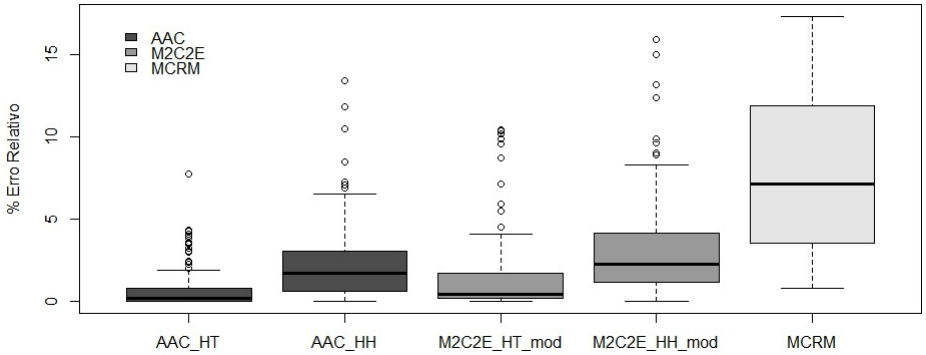


Figura 5. Boxplots dos erros relativos referentes às estimativas média da AAC com HT e com HH, do M2C2E com HT_mod e com HH_mod e MCRM.

Na Figura 5, é possível ver que o MCRM tem a maior amplitude interquartílica e concentração em valores mais elevados de erro relativo das estimativas médias de 100 replicações em comparação com os demais métodos. Por outro lado, os erros relativos da estimativas médias no M2C2E em comparação com a AAC estão próximos e em comparação com o MCRM estão bem inferiores.

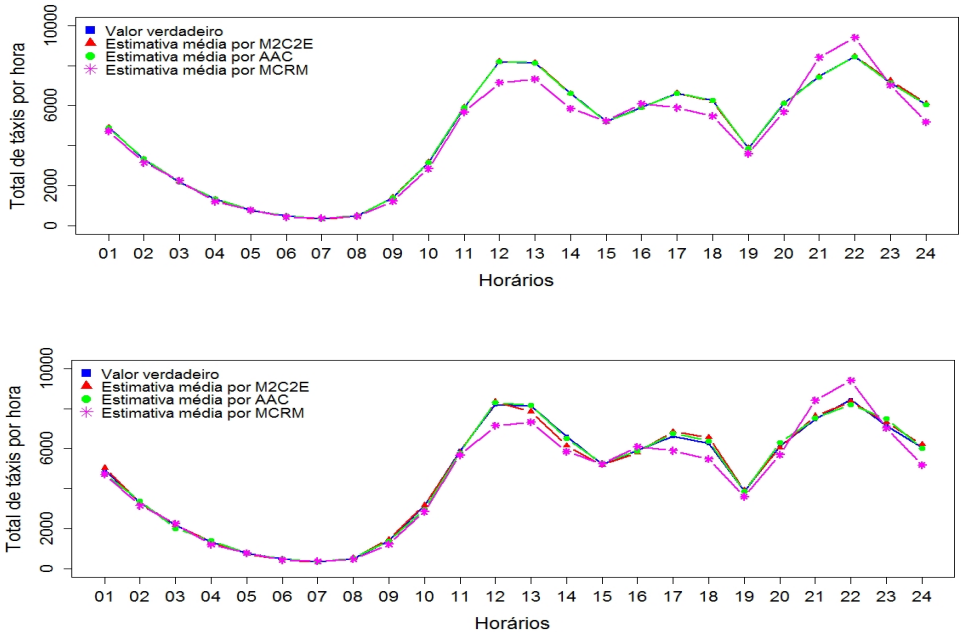


Figura 6. Estimativas médias para número total de táxis por AAC, M2C2E e MCRM nas configurações 1 (em cima) e 2 (embaixo) com $z_1 = 5\%N$.

A Figura 6 apresenta as estimativas médias após 100 replicações com $z_1 = 5\%N$ para ambas as configurações, mas também foram realizadas com $z_1 = \{1\%N, 3\%N, 10\%N, 15\%N\}$. Os vetores com as estimativas médias da AAC e do método M2C2E com todos os valores de z_1 foram testados através do teste de hipótese T^2 Hotelling, considerando as estimativas médias obtidas pela AAC e pelo M2C2E independente do tamanho inicial z_1. Na configuração 1, os horários 3, 7, 11, 14, 16, 22 e 23, ao nível de significância de 5%, há evidências para rejeitar que $\vec{\mu}_{M2C2E_{HT\_mod}} = \vec{\mu}_{AAC_{HT}}$ com p-valores iguais a 0,01; 0,00; 0,00; 0,00; 0,00; 0,01 e 0,00, respectivamente. Portanto, para os demais horários não há evidências para rejeitar a mesma hipótese. Por outro lado, na configuração 2, apenas nos horários 7 e 24 com p-valor de 0,01 em ambos, ao nível de significância de 5%, há evidências para rejeitar que $\vec{\mu}_{M2C2E_{HH\_mod}} = \vec{\mu}_{AAC_{HH}}$, para os demais horários não há evidências para rejeição.

6. Considerações Finais

Este artigo visa quantificar o total populacional em redes formadas por população rara e agrupada. O framework M2C2E, que implementa o MCRM com critério de parada proposto na camada 1 para estimar o tamanho populacional da célula e a AAC na camada 2 para determinar o total populacional da rede, foi apresentado.

Os estudos com dados sintéticos mostraram que o M2C2E apresenta-se melhor do que o MCRM e se aproxima da AAC, quando o total populacional aumenta e o tamanho da grade diminui. A validação com dados reais evidenciam as estimativas relevantes obtidas pelo framework proposto e o teste de hipótese para comparação dos métodos com os dados reais ao nível de significância de 5% revelam que a utilização do estimador de HH_mod aproxima as estimativas médias do M2C2E a AAC mais do que o estimador de HT_mod.

Os resultados permitem concluir que o M2C2E apresenta estimativas a serem usadas como uma solução à lacuna da AAC e tem vantagens significativas sobre o MCRM implementado separadamente em relação aos erros das estimativas médias e à região de estudo, visto que o M2C2E: (i) não precisa observar toda a região igual ao MCRM, mas apenas algumas células; e, (ii) não precisa conhecer toda população de cada célula, como pressuposto pela AAC, pois estima apenas a partir da amostra de alguns elementos.

Referências

Accettura, N., Neglia, G., and Grieco, L. A. (2015). The capture-recapture approach for population estimation in computer networks. *Computer Networks*, 89:107–122.

Dalal, S. R. and Mallows, C. L. (1990). Some graphical aids for deciding when to stop testing software. *IEEE Journal on Selected Areas in Communications*, 8(2):169–175.

El Emam, K. and Laitenberger, O. (2001). Evaluating capture-recapture models with two inspectors. *IEEE Transactions on Software Engineering*, 27(9):851–864.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Peng, S., Li, S., et al. (2009). Estimation of a population size in large-scale wireless sensor networks. *Journal of Computer Science and Technology*, 24(5):987–997.

Schnabel, Z. E. (1938). The estimation of the total fish population of a lake. *The American Mathematical Monthly*, 45(6):348–352.

Singham, D. I. (2010). *Analysis of Sequential Stopping Rules for Simulation Experiments*. PhD thesis, University of California, Berkeley, USA.

Smith, P. J. (1988). Bayesian methods for multiple capture-recapture surveys. *Biometrics*, pages 1177–1189.

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85(412):1050–1059.

Turk, P. and Borkowski, J. J. (2005). A review of adaptive cluster sampling: 1990–2003. *Environmental and Ecological Statistics*, 12(1):55–94.