UNIVERSIDADE FEDERAL FLUMINENSE

CAMILA DOMINGOS DA SILVA

Método de Estimação para Dimensionar o Total Populacional em Redes Compostas por População Rara e Agrupada

NITERÓI

2021

UNIVERSIDADE FEDERAL FLUMINENSE

CAMILA DOMINGOS DA SILVA

Método de Estimação para Dimensionar o Total Populacional em Redes Compostas por População Rara e Agrupada

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientador:
ANTÔNIO AUGUSTO DE ARAGÃO ROCHA

NITERÓI

2021

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

S586m Silva, Camila Domingos da
Método de estimação para dimensionar o total populacional em redes compostas por população rara e agrupada / Camila Domingos da Silva ; Antônio Augusto de Aragão Rocha, orientador. Niterói, 2021.
115 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2021.

DOI: http://dx.doi.org/10.22409/PGC.2021.m.05782849797

1. Amostragem adaptativa por conglomerados. 2. Método de captura e recaptura múltipla. 3. População rara e agrupada. 4. Estimadores. 5. Produção intelectual. I. Rocha, Antônio Augusto de Aragão, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

CAMILA DOMINGOS DA SILVA

Método de Estimação para Dimensionar o Total Populacional em Redes Compostas por População Rara e Agrupada

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Aprovada em 31 de março de 2021.

BANCA EXAMINADORA

Prof. Antônio Augusto de Aragão Rocha - Orientador, UFF

Prof. Daniel Ratton Figueiredo, UFRJ

Prof. Leandro Augusto Frata Fernandes, UFF

Niterói

2021

"We can only see a short distance ahead,
but we can see plenty there that needs to be done."
— Alan Turing.

Agradecimentos

Agradeço a Deus por me fortalecer todos os dias.

A minha querida mãe Maria Clara, pelo amor, pelo apoio e por acreditar que a educação pode transformar a minha vida.

Ao meu orientador, professor Antônio Augusto, por aceitar me orientar na construção deste trabalho. Suas sugestões serviram para a elaboração desta dissertação.

Aos professores Daniel Ratton Figueiredo e Leandro Augusto Frata Fernandes os quais aceitaram ao convite para compor a banca examinadora.

Aos meus professores, pelos conhecimentos transmitidos.

A todos que torcem pela minha vida acadêmica.

Aos autores por mim referenciados.

Resumo

A estimação do total populacional em redes cuja estrutura é composta por uma população rara e agrupada não é uma tarefa trivial, mas o sucesso da pesquisa e a otimização de recursos em rede podem depender do tamanho e da forma como a população apresenta-se distribuída. Neste cenário, é usual sobrepor uma grade a região na qual a população está contida, selecionar células dessa grade, avaliar se existem nelas elementos da população e nos casos favoráveis, adicionar as células vizinhas a elas que contenham a variável de interesse. Esta metodologia é apresentada pela Amostragem Adaptativa por Conglomerados - AAC. Contudo, a AAC leva em consideração a coleta de todos os elementos dentro da célula, o que não é realista para todos os casos e, sendo assim, a AAC também será chamada de Método Ótimo - MO, por utilizar o parâmetro. Visando solucionar tal fato, foi proposto um framework chamado de Método 2-Camadas e 2-Estimadores - M2C2E o qual implementa o Método de Captura e Recaptura Múltipla - MCRM na camada 1 para gerar as estimativas de total populacional dentro da célula a serem usadas como entrada para a AAC que é a camada 2 na qual estima-se o total populacional na grade. Entretanto, para conseguir aplicar o framework proposto foi preciso implementar um critério de parada para o número de recapturas no MCRM com o objetivo de obter estimativas eficientes sem exceder em número de recapturas. Os estudos com dados sintéticos e a aplicação aos dados reais revelam que o M2C2E fornece estimativas relevantes em relação ao MO e vantagens significativas sobre o MCRM separadamente.

Palavras-chave: Método de Captura e Recaptura Múltipla; Amostragem Adaptativa por Conglomerados; População Rara e Agrupada; Estimador de Horvitz-Thompson; Estimador de Hansen-Hurwitz; Estimador de Schnabel.

Abstract

The estimation of the total population in networks whose structure consists of a rare and clustered population is not a trivial task, but the success of research and the optimization of resources in network may depend of size and the distribution of population. In these circumstances, it is usual to overlay a grid over the region in which the population is contained, select cells from that grid, analyze the existence in them of the population elements and in favorable cases, add neighboring cells to it that contain the variable of interest. This methodology is presented by Adaptive Cluster Sampling - ACS. However, ACS considers the collection of all elements within the cell, which is not realistic for all cases, it was called the Optimal Method - OM, because it uses the parameter. In order to solve this fact, a framework called the 2-Layer and 2-Estimator Method - 2L2EM was proposed which implements the Multiple Capture and Recapture Method - MCRM in layer 1 to obtain the total population estimates within the cell to be used as input to ACS which is layer 2 where the total population in the network is estimated. However, in order to be able to apply the proposed framework, it was necessary to implement a stopping criterion for the number of recaptures to the MCRM in order to obtain efficient estimates without exceeding the number of recaptures. The studies with synthetic data and application to real data reveal that 2L2EM provides relevant estimates in relation to the OM and significant advantages over MCRM separately.

Keywords: Multiple Capture and Recapture Method; Adaptive Cluster Sampling; Rare and Clustered Population; Horvitz-Thompson Estimator; Hansen-Hurwitz Estimator; Schnabel Estimator.

Lista de Figuras

1.1 População real de táxis conectados a um aplicativo transporte através de redes móveis às 10:00hs do dia 22 de junho de 2016 no município do Rio de Janeiro representada pelas coordenadas em formato de "x" e sobreposta por uma grade com 1600 células. 4

3.1 Ilustração de célula no plano cartesiano. 10

3.2 Ilustração de células vizinhas no plano cartesiano. 10

3.3 Ilustração de um conglomerado da AAC com 24 células. 11

3.4 Ilustração de vizinhança na AAC através da utilização de arestas. 11

3.5 Exemplo de uma população rara e agrupada sobreposta por uma grade com 400 células. 12

3.6 Ilustração da metodologia da AAC em uma população rara e agrupada sobreposta por uma grade 400 células. 13

4.1 Ilustração da estrutura do estimador de Lincoln-Petersen. 21

4.2 Ilustração do Método de Captura e Recaptura Múltipla. 22

5.1 Ilustração do Método 2-Camadas e 2-Estimadores na camada 1. 30

5.2 Ilustração do Método 2-Camadas e 2-Estimadores na camada 1 e na camada 2. 30

5.3 População sintética sobreposta por uma grade com $N = 100$ células e $\tau = 50$ elementos de interesse. 34

5.4 População sintética sobreposta por uma grade com 100 células com seleção de 5 células iniciais. 34

5.5 População sintética sobreposta por uma grade com 100 células com seleção de 5 células iniciais e com as vizinhas imediatas. 34

5.6 População sintética sobreposta por uma grade com 100 células com seleção de 5 células iniciais e com todas as vizinhas. 35

6.1 Populações sintéticas com 100, 1000 e 10000 elementos. 38

6.2 Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 10$ referente a população sintética com 100 elementos. . 38

6.3 Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 10$ referente a população sintética com 1000 elementos. 39

6.4 Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 10$ referente a população sintética com 10000 elementos. 39

6.5 Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 100$ referente a população sintética com 10000 elementos. 40

6.6 Exemplo de refinamento de grade para populações extremamente pequenas saindo do MCRM grade 1x1 para a AAC grade 10x10. 41

6.7 Populações sintéticas utilizadas para a comparação entre M2C2E e MO. . . 42

6.8 Boxplots com os erros relativos das estimativas médias do M2C2E usando o estimador de Horvitz-Thompson modificado das populações sintéticas $\tau = 1000$ e $N = 100$, $\tau = 1000$ e $N = 400, \tau = 2000$ e $N = 100$, $\tau = 2000$ e $N = 400$. 43

6.9 Boxplots com as eficiências entre o M2C2E e o MO das populações sintéticas $\tau = 1000$ e $N = 100$, $\tau = 1000$ e $N = 400, \tau = 2000$ e $N = 100$, $\tau = 2000$ e $N = 400$. 43

6.10 Estimativas médias referente ao número total das populações sintéticas $\tau = 1000$ e $N = 100$, $\tau = 1000$ e $N = 400$, $\tau = 2000$ e $N = 100$, $\tau = 2000$ e $N = 400$ usando o Método Ótimo - MO e o Método 2-Camadas e 2-Estimadores - M2C2E. 44

6.11 População sintética com 1000 elementos e estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 100$. . . . 46

6.12 População sintética com 2000 elementos e estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 100$. . . . 47

6.13 Boxplots com 50 replicações para cada número de recaptura dos dados sintéticos de 1000 elementos com $n_1 = n_2 = ... = 100$. 49

6.14 Boxplots com 50 replicações para cada número de recaptura dos dados sintéticos de 2000 elementos com $n_1 = n_2 = ... = 100$. 49

7.1 Contorno do município do Rio de Janeiro sobreposto por uma grade com 256 células. 51

7.2 Gráfico de barras do número total de táxis no dia 22 de junho de 2016 no município do Rio de Janeiro por hora. 52

7.3 Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da madrugada entre 01:00hs - 06:00hs no município do Rio de Janeiro sobreposta por grade 40x40. 54

7.4 Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da manhã entre 07:00hs - 12:00hs no município do Rio de Janeiro sobreposta por grade 40x40. 55

7.5 Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da tarde entre 13:00hs - 18:00hs no município do Rio de Janeiro sobreposta por grade 40x40. 56

7.6 Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da noite entre 19:00hs - 24:00hs no município do Rio de Janeiro sobreposta por grade 40x40. 57

7.7 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 1\%N$. 71

7.8 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 3\%N$. 72

7.9 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 5\%N$. 72

7.10 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 10\%N$. 73

7.11 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 15\%N$. 73

7.12 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 1\%N$. 74

7.13 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 3\%N$. 74

7.14 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 5\%N$. 75

7.15 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 10\%N$. 75

7.16 Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 15\%N$. 76

7.17 Boxplots dos erros relativos referentes as estimativas médias do método ótimo usando o estimador de Horvitz-Thompson - MO com HT e estimador de Hansen-Hurwitz - MO com HH, do framework proposto usando o estimador de Horvitz-Thompson - M2C2E com HT_mod e estimador de Hansen-Hurwitz - M2C2E com HH_mod e MCRM. 77

7.18 Boxplots do número médio de células visitadas do método ótimo - MO e do framework proposto - M2C2E, por número de células visitadas inicialmente $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$. 77

Lista de Tabelas

2.1 Estimadores utilizados no Método de Captura e Recaptura Simples e Múltipla apresentados em trabalhos anteriores. 7

2.2 Estimadores utilizados na Amostragem Adaptativa por Conglomerados apresentados em trabalhos anteriores. 8

4.1 Notações específicas para Método Captura e Recaptura Múltipla. 22

4.2 Estruturação das variáveis referente ao estimador de Schnabel. 24

5.1 Notações para o M2C2E. 31

6.1 Resultado da implementação considerando o valor fixo $\phi = 0,1$ com $n_1 = n_2 = ... = n_j = 10$ e variando os valores de τ e N. 45

7.1 Configuração 1 - Resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 59

7.2 Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 60

7.3 Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 61

7.4 Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 62

7.5 Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 63

7.6 Resultados do MCRM com $n_1 = n_2 = ... = n_j = 100$ para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016. 64

7.7 Configuração 2 - Resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 65

7.8 Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 66

7.9 Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 67

7.10 Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 68

7.11 Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos. 69

7.12 Estatística descritiva dos erros relativos referentes as estimativas médias do Método Ótimo, do Método 2-Camadas e 2-Estimadores e do Método de Captura e Recaptura Múltipla nas configurações 1 e 2. 76

Lista de Abreviaturas e Siglas

AAC	:	Amostragem Adaptativa por Conglomerados
AAS	:	Amostragem Aleatória Simples
AASc	:	Amostragem Aleatória Simples com Reposição
AASs	:	Amostragem Aleatória Simples sem Reposição
CV	:	Coeficiente de Variação
EB	:	Estimador de Bailey
EC	:	Estimador de Chapman
ELP	:	Estimador de Lincoln-Petersen
ER	:	Erro Relativo
ES	:	Estimador de Schnabel
ESE	:	Estimador de Schumacher-Eschmeyer
HH	:	Estimador de Hansen-Hurwitz
HHmod	:	Estimador de Hansen-Hurwitz Modificado
HT	:	Estimador de Horvitz-Thompson
HTmod	:	Estimador de Horvitz-Thompson Modificado
IC	:	Intervalo de Confiança
M2C2E	:	Método 2-Camadas e 2-Estimadores
MCRM	:	Método de Captura e Recaptura Múltipla
MCRS	:	Método de Captura e Recaptura Simples
MO	:	Método Ótimo
NB	:	Normal Bivariada

Sumário

1 Introdução **1**

1.1 Contextualização 3

1.2 Objetivos 4

1.3 Organização 4

2 Trabalhos Relacionados **6**

3 Amostragem Adaptativa por Conglomerados **9**

3.1 Conceitos Básicos da AAC 9

3.1.1 Célula, Grade e Vizinhança 9

3.1.2 Rede, Borda e Conglomerado 11

3.2 Metodologia da AAC 12

3.3 Estimadores usando na AAC 15

3.3.1 Estimador de Horvitz-Thompson 15

3.3.2 Estimador de Hansen-Hurwitz 16

3.4 Critério de Parada da AAC 17

4 Método de Captura e Recaptura **18**

4.1 Método de Captura e Recaptura Simples 18

4.1.1 Estimador de Lincoln-Petersen 19

4.1.2 Estimador de Chapman 21

4.1.3 Estimador de Bailey 21

4.2 Método de Captura e Recaptura Múltipla 22

4.2.1 Estimador de Schnabel 23

4.2.2 Estimador de Schumacher-Eschmeyer 24

4.3 Critérios de Parada para o Método de Captura e Recaptura Múltipla . . . 24

4.3.1 Usando Coeficiente de Variação 26

4.3.2 Usando Intervalo de Confiança 27

5 Framework Proposto **29**

5.1 Metodologia 29

5.1.1 Estimadores Modificados 31

5.1.2 Etapas do M2C2E com Ilustrações 33

5.2 Vantagens e Desvantagens 35

6 Estudo com Dados Sintéticos **36**

6.1 Criação dos Dados Sintéticos 37

6.2 Implementando o Critério de Parada no MCRM 37

6.3 Verificando os Métodos: M2C2E, MO e MCRM 41

6.4 Convergência do Estimador de Schnabel 48

7 Aplicação a Dados Reais **50**

7.1 Descrição do Conjunto de Dados 50

7.2 Resultados 51

7.2.1 Primeira Configuração 58

7.2.2 Segunda Configuração 58

7.2.3 Análise dos Métodos 70

7.2.4 Teste de Hipótese para Comparação dos Métodos 78

8 Considerações Finais **80**

8.1 Contribuições 82

8.2 Trabalhos Futuros 82

Referências **83**

Apêndice A – Aspectos Computacionais **87**

A.1 População Rara e Agrupada: Geração e Contagem 87

A.2 Método de Captura e Recaptura Múltipla: Estimador de Schnabel e Critério de Parada 90

A.3 Amostragem Adaptativa por Conglomerados 93

Capítulo 1

Introdução

Em diversas áreas de estudo, não sendo diferente em Computação, é comum surgir a necessidade de quantificar o número de elementos da população o qual também será chamado de total populacional ou variável de interesse. Contudo, saber o número total de elementos em uma rede pode necessitar de métodos amostrais complexos, partindo do princípio que a estrutura da população venha a ser agrupada, ou seja, com maiores concentrações em determinadas áreas, e por consequência existindo muitas áreas sem a variável de interesse. Vale lembrar que a otimização de recursos para evitar proliferação de anomalias em rede, por exemplo, dependem do tamanho e da forma como a população apresenta-se distribuída.

Nos casos em que as variáveis de interesse são apenas encontradas em pequenos números, tais como falhas raras em um software, define-se este tipo de população como rara, Thompson [48]. Alguns problemas em redes computacionais podem ser caracterizados por eventos raros e essa tendência aumenta a importância de estudar as populações raras e agrupadas, visando solucionar problemas antes de chegarem a todos os usuários ou mesmo determinar de antemão a magnitude deles. Em populações deste tipo, caso fosse utilizado o estimador mais comumente empregado para total populacional que é o estimador da amostragem aleatória simples - AAS, as estimativas obtidas apresentariam erros altos, devido ao grande número de regiões amostrais sem a característica em estudo.

Dentre os estimadores possíveis, Blower et al. [6] analisou o princípio para a utilização do estimador de Lincoln-Petersen - ELP no qual os elementos são marcados e liberados de volta à população, posteriormente, uma segunda amostra é tirada após um período de tempo e o número de elementos anteriormente já marcados é anotado a cada amostra. King et al. [30] explicam com mais detalhes esse método chamado de Captura e Recaptura, além de apresentar as suposições necessárias à aplicação do estimador de Lincoln-Petersen,

tais como necessidade de população fechada (não havendo nascimentos, mortes e migração no período de estudo); todos os elementos têm a mesma probabilidade de captura; e na ocasião da recaptura todos os elementos previamente observados podem ser identificados. Fundamentalmente, o método é constituído por uma captura e uma recaptura, as quais são extrações de uma amostra aleatória simples sem reposição para cada uma.

Contudo, caso o número de recapturados seja igual a zero, utilizar o estimador de Lincoln-Petersen - ELP não reflete o valor do parâmetro, pois em sua formulação contém o número de recapturados no denominador, o qual seria zerado e, matematicamente, não está definida a divisão por zero. Portanto, nesse cenário usar uma única recaptura não se aplica e outros estimadores foram propostos para solucionar essa questão, tais como o estimador de Chapman [17] e o estimador de Bailey [5].

Observando tal problema, Schnabel [37] contornou através do Método de Captura e Recaptura Múltipla - MCRM. Esse método é composto por uma captura inicial e $R = \{R_1, R_2, R_3, R_4, ..., R_k\}$ recapturas ao longo do tempo. Dessa forma, o MCRM apresenta seus estimadores apropriados, tais como estimador de Schnabel, estimador de Schumacher e Eschmeyer, dentre outros.

Um outro problema que surge quando pensa-se em múltiplas recapturas é quando parar o processo, ou seja, o momento em que as estimativas estão tão próximas do parâmetro do total populacional que acréscimos de recaptura trarão melhorias pouco significativas. Na literatura, muitos critérios de parada foram implementados, entretanto faltam estudos conclusivos sobre regras de parada com foco no MCRM. Contudo, nesta dissertação foi estudado e implementado o critério de parada proposto por Singham et al. [40] [41] [42] que visa os processos sequenciais.

A Amostragem Adaptativa por Conglomerados - AAC foi proposta por Thompson [48] para estimar o total populacional sobre o contexto de populações raras e agrupadas usando o estimador de Horvitz-Thompson - HT, por exemplo. O desenho amostral da AAC consiste em sobrepor uma grade a população de interesse, amostrar células dessa grade e observar todos os elementos dentro da célula selecionada, caso a célula contenha algum elemento de interesse, as células vizinhas a ela também serão incluídas na amostra até chegar à primeira célula vizinha que não contenha a variável de interesse.

Tendo em vista que a AAC leva em consideração a possibilidade de observar todos os elementos de interesse dentro da célula para esse caso deu-se o nome de Método Ótimo - MO. Contudo, pode não ser viável encontrá-los em todos os cenários como, por exemplo, para elementos extremamente difíceis de serem capturados por estarem escondidos

ou misturados a outra população com grande número de elementos sem a variável de interesse ao estudo em questão. Turk et al. [51] fizeram uma revisão sobre os principais desenvolvimentos na AAC desde sua introdução por Thompson [48] e 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 algumas situações. Essa condição C tornou-se a principal questão de pesquisa a ser desenvolvida nesta dissertação, levando a implementação do framework proposto.

Portanto, foi proposto um framework chamado de Método 2-Camadas e 2-Estimadores - M2C2E que procura solucionar o problema de não ser possível encontrar todos os elementos 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 et al. [40] [41] [42] para não exceder em número de recapturas, e posteriormente, utiliza-se o estimador de Horvitz-Thompson ou estimador de Hansen-Hurwitz usual da AAC para estimar o total populacional da grade.

Sendo assim, foram geradas populações sintéticas com o intuito de analisar o critério de parada proposto por Singham et al. [40] [41] [42] ao MCRM e o desempenho dos três métodos: M2C2E, MO e MCRM. Além disso, um caso de uso aos sistemas distribuídos foi estudado com dados reais de táxis no município do Rio de Janeiro no dia 22 de junho de 2016 conectados a um aplicativo de transporte através das redes móveis de internet com acesso fornecido pelas principais operadoras de telefonia nacional. Os dados reais foram utilizados para estudar os métodos clássicos na literatura MO e MCRM e o framework proposto.

1.1 Contextualização

As dificuldades em estimar o total populacional em redes compostas por população rara e agrupada encontram-se em diferentes áreas de concentração da computação, tais como engenharia de software, redes de computadores e sistemas distribuídos. Uma aplicação possível está em estimar o total populacional em uma região a partir do acesso a um determinado aplicativo. A Figura 1.1 representa um problema de aplicação real para determinar o total populacional que será desenvolvido, desta dissertação, a partir das coordenadas geográficas dos táxis por hora sobrepostas por uma grade.

Suponha que um aplicativo contenha três grupos envolvidos: os usuários; a empresa prestadora de serviço; e uma organização concorrente. A empresa prestadora de serviço

detém todas as informações sobre os usuários do aplicativo, inclusive o número de usuários por período de tempo. Seria possível a organização concorrente obter essa informação sem que a empresa prestadora lhe forneça quaisquer dados, levando em consideração que a distribuição dos usuários seja rara e agrupada?

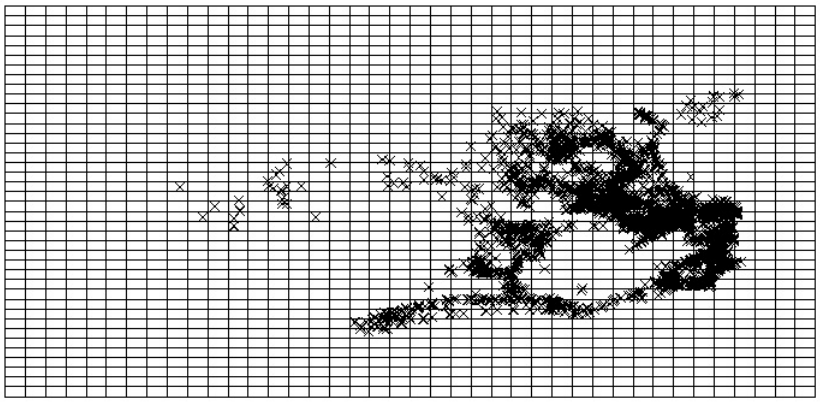


Figura 1.1: População real de táxis conectados a um aplicativo transporte através de redes móveis às 10:00hs do dia 22 de junho de 2016 no município do Rio de Janeiro representada pelas coordenadas em formato de "x" e sobreposta por uma grade com 1600 células.

1.2 Objetivos

Pretende-se por objetivos gerais, propor e analisar um método para tornar a AAC um plano alinhado nos casos onde existe a impossibilidade de encontrar todos os elementos desejados e estabelecer o critério de parada para o número de recapturas do Método de Captura e Recaptura Múltipla verificando se a estimativa resultante na parada é satisfatória e converge para o valor do parâmetro.

Os objetivos específicos são implementar o Método de Captura e Recaptura Múltipla usando populações sintéticas sobrepostas por grades de tamanhos diferentes e variando o número de células selecionadas na amostra inicial; realizar um estudo com dados sintéticos para analisar o desempenho do MCRM, do M2C2E e do MO; concluir quais são os melhores cenários para aplicar cada um deles; e, por fim, avaliar suas aplicações em um sistema distribuído constituído por dados reais.

1.3 Organização

Esta dissertação está organizada em 8 capítulos, na qual o Capítulo 1 é a introdução e no Capítulo 2 é possível encontrar a apresentação dos trabalhos relacionados com alguns

autores que também trabalharam com a Amostragem Adaptativa por Conglomerados e com o Método de Captura e Recaptura Simples e Múltipla.

A Amostragem Adaptativa por Conglomerados - AAC é explicada no Capítulo 3 através de seus conceitos básicos e metodologia. As formulações dos estimadores de Horvitz-Thompson e de Hansen-Hurwitz juntamente com suas variâncias são apresentadas. A explicação de como é construído o critério de parada, a partir da própria estrutura da AAC, é realizada no fim do capítulo.

No Capítulo 4, estão descritos o Método de Captura e Recaptura Simples e o Método de Captura e Recaptura Múltipla. Assim como os estimadores apropriados e mais utilizados na literatura. É possível observar no Capítulo 4 uma discussão sobre Critérios de Parada para o Método de Captura e Recaptura Múltipla usando coeficiente de variação e intervalo de confiança. Em seguida, o Capítulo 5 apresenta um framework proposto para correção da lacuna na AAC com metodologia, notações, estimadores modificados, etapas com ilustrações, vantagens e desvantagens.

No Capítulo 6, encontra-se o estudo com dados sintéticos. Foram geradas populações sintéticas visando avaliar a viabilidade do critério de parada proposto por Singham et al. [40] [41] [42] para o número de recapturas do MCRM. Nesse capítulo também foi verificada a AAC para 4 tipos de redes sintéticas com a variância fixada comparando com o framework proposto e os resultados são apresentados.

O desempenho do MO, do MCRM e do M2C2E com dados reais de táxis no município do Rio de Janeiro extraídos no dia 22 de junho de 2016 totalizando 110369 coordenadas geográficas distribuídas ao longo do dia foram analisadas no Capítulo 7. Por fim, o Capítulo 8 conclui-se dando um breve resumo sobre o que foi desenvolvido, comentando sobre os resultados obtidos e acrescentou um pouco sobre os trabalhos que podem ser desenvolvidos futuramente.

Capítulo 2

Trabalhos Relacionados

Na área de Computação, Mills [33] desenvolveu a primeira pesquisa usando o Método de Captura e Recaptura na engenharia de software. O objetivo do estudo era semear falsos defeitos antes de iniciar os testes principais e durante as varreduras detectar falsos defeitos e defeitos reais. Além disso, usou o estimador Lincoln-Peterson para obter uma estimativa do número total de defeitos, assumindo que ambos os defeitos tivessem as mesmas probabilidades de serem detectados.

Levando em consideração os estimadores, Tallmon et al. [46] afirmaram ter problemas para estimar o tamanho efetivo da população, porque a maioria dos estimadores isoladamente se apresentaram imprecisos ou enviesados e Thompson [49] salientou a dificuldade de dimensionar o tamanho total de populações raras e agrupadas. Na mesma linha de pensamento, Solberg et al. [44] adicionaram que embora haja uma variedade de métodos disponíveis para estimar abundância e densidade das populações, a maioria dos estudos depende de apenas um estimador e poucos estudos realizam comparações entre eles ou avaliaram criticamente a adequação de cada um.

Dentre os trabalhos que realizaram comparações entre os estimadores específicos do MCRM, Mares et al. [32] compararam as estimativas médias de Lincoln-Petersen, Schnabel e Schumacher-Eschmeyer e mostraram que todos os resultados subestimaram o tamanho real da população. Oliveira [34] realizou um estudo comparativo considerando um algoritmo computacional para gerar as amostras aleatórias de dados e o interesse principal era analisar uma amostra inicial e uma amostra secundária com tamanho de 150%, 100%, 75%, 50% e 25% em relação ao tamanho da primeira amostra e registrar o número de elementos marcados para cada uma das recapturas.

Portanto, nas pesquisas de Oliveira [34], o interesse é variar os tamanhos de captura

e de recaptura com a finalidade de analisar qual é comportamento dos estimadores de Lincoln-Petersen, Chapman e Bailey. Observou-se que, depois das iterações, as estimativas de Lincoln-Petersen foram as que obtiveram as melhores aproximações para o total populacional no caso em que a segunda amostra era maior do que a primeira amostra. Contudo, ao ser levado em consideração os erros padrão e os intervalos de confiança, as estimativas usando o estimador de Chapman apresentaram menores erros e amplitude.

Jorge et al. [29] também realizaram um estudo para verificar o comportamento dos estimadores Lincoln-Petersen, Chapman e Bailey e concluíram que o melhor estimador considerando o erro quadrático médio e o erro padrão foi o estimador de Bailey. Quando o mesmo estudo foi repetido por eles para maiores populações, os estimadores foram equivalentes.

Hall [24] afirmou que em todos os seus experimentos as estimativas de Schnabel foram as que apresentaram maior acurácia. Em populações nas quais de 25% a 50% da população é marcada são melhores estimadas pelo estimador de Schumacher e Eschmeyer, enquanto as estimativas de Schnabel são mais eficientes quando apenas uma pequena parte é marcada.

Nos estudos de Budrys et al. [10], comparações entre os estimadores de Chapman, Schnabel e Schumacher-Eschmeyer foram realizadas e notou-se que as estimativas de Schumacher-Eschmeyer são aparentemente razoáveis. Nos casos de alta taxa de recaptura, os estimadores baseado no modelo de população fechada (Chapman, Schnabel e Schumacher-Eschmeyer) forneceram resultados semelhantes com confiança relativamente estreita de intervalos e abrangendo o tamanho real de população. Nos casos de baixa taxa de recaptura, o estimador de Schnabel foi menos suscetível à flutuação ou falta parcial de recapturas nas séries de censos, mas a combinação do estimador de Chapman e o estimador de Schumacher-Eschmeyer forneceram intervalo de confiança mais estreito.

	Lincoln-Petersen	Chapman	Schnabel	Schumacher
Accettura et al. [1]	✓	✓	✓	✓
Akanda et al. [3]	✓	–	✓	–
Blower et al. [6]	✓	–	–	–
King et al. [30]	✓	✓	–	–
Oliveira [34]	–	✓	✓	–
Schnabel [37]	–	–	✓	–
Schumacher et al. [38]	–	–	–	✓

Tabela 2.1: Estimadores utilizados no Método de Captura e Recaptura Simples e Múltipla apresentados em trabalhos anteriores.

	Horvitz-Thompson	Hansen-Hurwitz
Brown et al. [8]	✓	✓
Cochran [13]	✓	✓
Hansen et al. [25]	–	✓
Horvitz et al. [26]	✓	–
Thompson et al. [48]	✓	✓

Tabela 2.2: Estimadores utilizados na Amostragem Adaptativa por Conglomerados apresentados em trabalhos anteriores.

Nas Tabelas 2.1 e 2.2, os exemplos de autores que utilizaram os estimadores mais comumente empregados na AAC, no MCRS e no MCRM em seus trabalhos são apresentados. Nas colunas, estão os estimadores e nas linhas, as referências, no interior da tabela "✓" significa que o autor apresentou tal estimador e "–" para não mencionou.

Em relação a populações raras e agrupadas, Townsend et al. [50] enfatizaram que estimativas precisas para este tipo de população, usando implementação computacional, podem exigir tempos elevados de execução. Thompson [48] implementou uma região de estudo com 400 células para estimar o número de objetos-ponto raros e agrupados os quais foram produzidos por um processo de Poisson conforme Diggle [18] com cinco núcleos através de uma distribuição Uniforme na região e uma distribuição de Poisson com média igual a 40 números de objetos-ponto dispersos em relação aos núcleos com uma distribuição Gaussiana com desvio padrão de $\phi = 0,02$.

Accettura et al. [1] questionaram a relação entre a eficácia do Método de Captura e Recaptura para dimensionar redes e sua complexidade computacional, eles também afirmam que a otimização de serviço ou sistema requer estimativas precisas do número de itens-chave envolvidos tais como nós, usuários, arquivos, pacotes e fluxos. Contudo, Accettura et al. [1] não se aplicaram a resolver o problema de estimar populações raras e agrupadas, assim como Peng et al. [36] e outros autores os quais focaram em grandes populações.

A principal problemática desta dissertação refere-se a impossibilidade ou dificuldade em determinar o número total de elementos dentro da célula na AAC. Esta questão foi trazida por Turk et al. [51], os quais comentaram ser difícil ou impossível de determinar em algumas situações. Outros autores que também utilizaram a AAC, tais como Gattone et al. [22], Brown et al. [9] e Singh et al. [39], não procuraram solucionar essa questão e desconsideraram essa limitação em suas pesquisas.

Capítulo 3

Amostragem Adaptativa por Conglomerados

A Amostragem Adaptativa por Conglomerados - AAC foi proposta por Thompson [48] e foi considerada como um planejamento amostral eficiente para populações raras e agrupadas. Este tipo de população encontra-se em pequenos números e concentrada em grupos esparsos e dispersos em uma região, por definição. A ideia principal da AAC é sobrepor uma grade a região cujos elementos de interesse estão contidos e utilizar a própria estrutura da população para construir o plano amostral.

3.1 Conceitos Básicos da AAC

Na literatura sobre AAC, observa-se a introdução de alguns conceitos próprios tais como: célula, grade, vizinhança, conglomerado, rede e borda. Nas subseções a seguir, serão apresentadas as definições para cada um deles.

3.1.1 Célula, Grade e Vizinhança

Considere $[a, b]$ um intervalo fechado em $\mathbb{R}$, isto é, $[a, b] = \{t \in \mathbb{R} \,|\, a \leq t \leq b\}$. Seja o subconjunto $\mathcal{P} = \{t_0, t_1, \cdots, t_{v-1}, t_v\}$ de $[a, b]$ de uma partição, de ordem $v+1$ do intervalo $[a, b]$ se satisfaz a seguinte condição: $t_0 = a < t_1 < \cdots < t_{v-1} < b = t_v$. Note que $v + 1$ é o número de elementos da particão $\mathcal{P}$. Cada retângulo $R_{ij} = [t_i, t_{i+1}]$ x $[t_j, t_{j+1}]$ com vértices nos pontos $(t_i, t_j), (t_{i+1}, t_j), (t_i, t_{j+1})$ e (t_{i+1}, t_{j+1}) são chamados de células em $\mathbb{R}^2$ para $i, j = 0, 1, ..., v - 1$. A Figura 3.1 ilustra uma célula em $\mathbb{R}^2$.

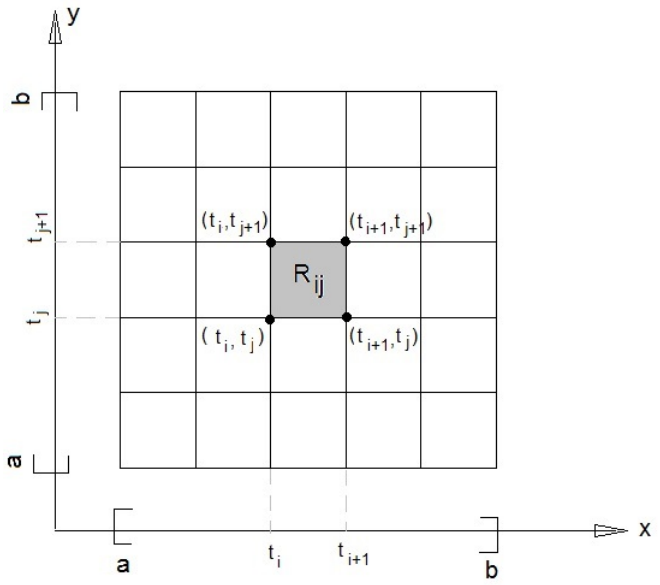


Figura 3.1: Ilustração de célula no plano cartesiano.

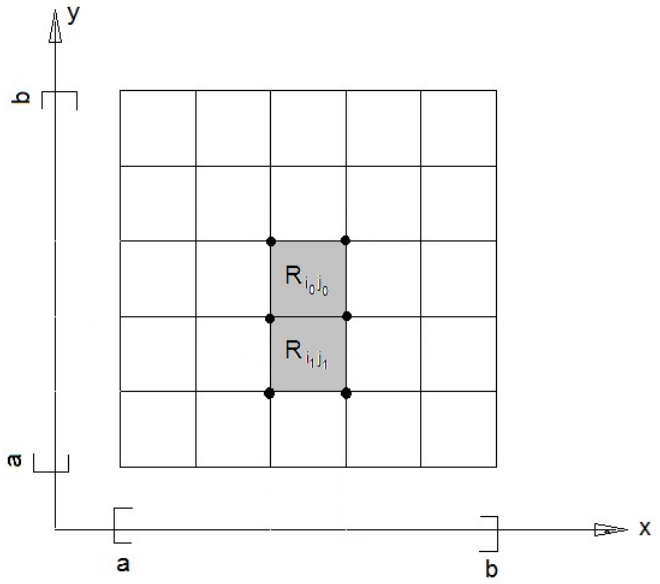


Figura 3.2: Ilustração de células vizinhas no plano cartesiano.

Já uma grade é o conjunto de todas as células com a mesma área em $\mathbb{R}^2$ para uma fixada partição P de $[a, b]$. Tem-se também que duas células R_{i_0,j_0} e R_{i_1,j_1} são vizinhas quando possuem um lado em comum de acordo com a Figura 3.2, ou seja, os segmentos de reta:

$$\{((1-\lambda)t_i + \lambda t_{i+1}, t_j) \in \mathbb{R}^2/\lambda \in [0,1]\},\ \{((1-\lambda)t_i + \lambda t_{i+1}, t_{j+1}) \in \mathbb{R}^2/\lambda \in [0,1]\},$$
$$\{(t_i, (1-\lambda)t_j + \lambda t_{j+1}) \in \mathbb{R}^2/\lambda \in [0,1]\} \text{ e } \{(t_{i+1}, (1-\lambda)t_j + \lambda t_{j+1}) \in \mathbb{R}^2/\lambda \in [0,1]\}$$

onde λ é uma variável no intervalo $[0, 1]$ com a finalidade de definir um segmento de reta.

3.1.2 Rede, Borda e Conglomerado

Com os elementos de interesse espalhados pela grade, em locais fixos, de forma rara e agrupada, define-se rede pelo conjunto formado por células nas quais os elementos de interesse podem ser encontrados. As bordas são células que não contêm elementos de interesse, mas são vizinhas de células que satisfazem a condição. O conglomerado é a união das células de rede e das células de borda.

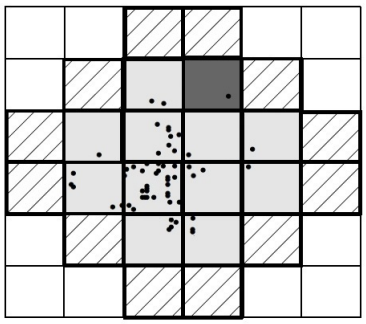


Figura 3.3: Ilustração de um conglomerado da AAC com 24 células.

Os conceitos básicos definidos são visualizados na Figura 3.3 que é um conglomerado com 24 células. Observe que os quadrados em cinza claro e o único quadrado em cinza escuro são células de rede, logo constituem a rede, em particular o quadrado em cinza escuro representa a primeira célula observada na rede. Os quadrados hachurados formam as células de borda. Finalmente, o conglomerado é a união de todos estes quadrados, exceto os quadrados em branco.

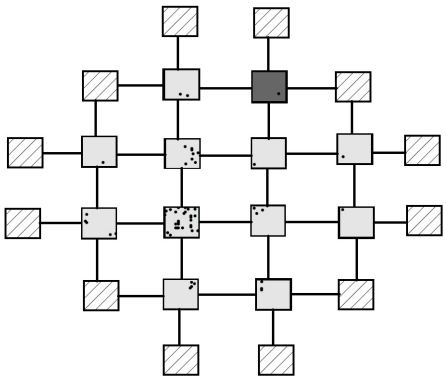


Figura 3.4: Ilustração de vizinhança na AAC através da utilização de arestas.

A Figura 3.4 é o mesmo conglomerado da Figura 3.3, sendo que as células foram

afastadas e arestas que ligam as células foram adicionadas para representar o conceito de vizinhança. Note que apenas as células em cinza e a célula em cinza escuro emitem arestas para os 4 lados (direita, esquerda, em cima e em baixo), porque elas contêm elementos de interesse e as células de borda uma vez que não contêm, elas não adicionam nenhuma outra célula e apenas recebem arestas das células que formam a rede.

3.2 Metodologia da AAC

Seja uma determinada população rara e agrupada disposta em uma região. Conforme a AAC é necessário sobrepor uma grade sobre a região de estudo e selecionar amostras de células desta grade para conhecer o total populacional. A ideia principal da AAC é utilizar a estrutura da rede para construir o plano amostral. Na AAC, quando um elemento de interesse é encontrado, observa-se as localizações vizinhas. Uma vez que a vizinhança revela com maior probabilidade outras concentrações de elementos com a característica de interesse dessa população. Na Figura 3.5 é possível observar uma grade com $N = 400$ células no $\mathbb{R}^2$ e os pontos representam a população de interesse.

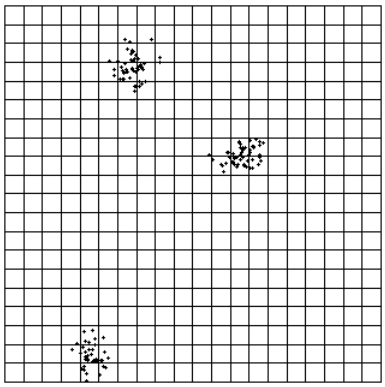
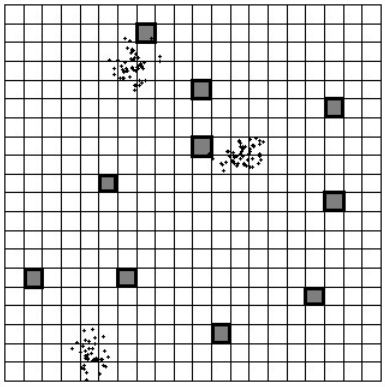


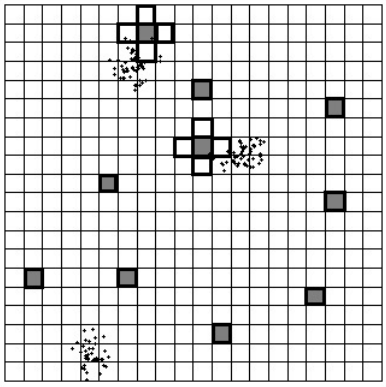
Figura 3.5: Exemplo de uma população rara e agrupada sobreposta por uma grade com 400 células.

A AAC inicia-se com a extração de um subconjunto da grade, ou seja, uma amostra ao acaso de z_1 células com reposição (AASc) ou sem reposição (AASs) na qual todas as células 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 extrair todos os elementos dela e das células vizinhas, até chegar às células que não contenham nenhum elemento da população de interesse, em outras palavras, até obter a borda. Tal fato, indica que o tamanho amostral, ou seja,

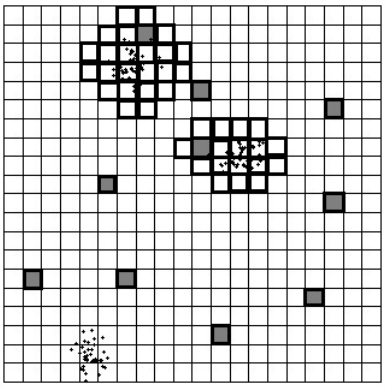
o número de células observadas até a parada do método é uma variável aleatória. Isto ocorre, porque dentro das células vizinhas podem haver elementos de interesse e, portanto, a seleção das vizinhas delas continua o processo ou, por outro lado, podem não haver elementos e seleção é finalizada.



(a) Etapa 1



(b) Etapa 2



(c) Etapa 3

Figura 3.6: Ilustração da metodologia da AAC em uma população rara e agrupada sobreposta por uma grade 400 células.

Na Figura 3.6, encontra-se uma ilustração da metodologia da AAC para a população da Figura 3.5, assumindo que os elementos de interesse foram sobrepostos por uma grade, os passos posteriores são apresentados nas etapas a seguir:

1. Extrair uma amostra aleatória simples com ou sem reposição de z_1 células da grade. Essa etapa está representada na Figura 3.6(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 as quais compartilham lados com a célula i serão adicionadas a amostra. Note que na Figura 3.6(a) duas células satisfizeram a condição C, então apenas as células vizinhas a estas foram observadas, como ilustrado na Figura 3.6(b);

3. O processo de observar células vizinhas continuará até chegar às células que não satisfazem a condição de interesse, conforme a Figura 3.6(c).

Seja n a variável que representa o conjunto formado por todas células pertencentes a amostra. Uma célula i pode ser incluída na amostra de três maneiras diferentes: se ela for selecionada na amostra inicial z_1; se ela for qualquer célula da rede; ou 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 a 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, mas seja selecionada na amostra inicial das z_1 células, 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 $a_i = 0$. Na Figura 3.3 tem-se que $m_i = 12$ e $a_i = 0$, sendo que a célula i seja qualquer célula que pertença a rede em cinza claro. Vale ressaltar que se i é célula de borda, não se pode garantir que i também não seja borda de uma outra rede que não foi observada.

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 de seleção é dada por:

$$p_i = \frac{m_i + a_i}{N}, \tag{3.1}$$

em que N é o número total de células construídas na grade em $\mathbb{R}^2$.

Outra probabilidade importante é a probabilidade de inclusão da célula i na amostra através da AASs, onde n é número total de células observadas ao final do processo, é dada por:

$$\pi_i = 1 - P(\{i \text{ não estar incluída entre as células } n\}) = 1 - \frac{\binom{N - m_i - a_i}{n}}{\binom{N}{n}}.$$

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 a Equação 3.1. Contudo, a probabilidade de inclusão da célula i dentro da AASc é dada pela seguinte expressão:

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

3.3 Estimadores usando na AAC

A Amostragem Adaptativa por Conglomerados - AAC leva em consideração para desenvolver suas técnicas 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. Sendo assim, a AAC utiliza os estimadores para seleção de unidades amostrais com probabilidades desiguais, tais como o estimador de Horvitz-Thompson na Seção 3.3.1 e o estimador de Hansen-Hurwitz na Seção 3.3.2.

3.3.1 Estimador de Horvitz-Thompson

O estimador Horvitz-Thompson - HT é considerado um estimador não tendencioso do total populacional para população com probabilidades desiguais de seleção e as células z_1 selecionadas por uma AAS sem reposição. Seja π_i a probabilidade da célula i estar na amostra, π_j a probabilidade da célula j estar na amostra e π_{ij} a probabilidade da célula i e da célula j estarem simultaneamente na amostra. Segundo Horvitz et al. [26], a expressão do estimador de Horvitz-Thompson do total populacional é dada por:

$$\widehat{\tau}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}, \tag{3.2}$$

na qual $s = \{1, 2, ..., n\}$ é o conjunto por todas as células visitadas ao final do processo da AAC e y_i é o número total de elementos de interesse na célula i.

Mas o estimador de Horvitz-Thompson $\widehat{\tau}_{HT}$ também pode ser reescrito da seguinte maneira:

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

onde I_i é uma variável indicadora de inclusão da célula i na amostra. Ou seja, $I_i = 1$, se i pertence a s e $I_i = 0$, caso contrário.

A variância do estimador de Horvitz-Thompson $Var(\widehat{\tau}_{HT})$ é dada por:

$$Var(\widehat{\tau}_{HT}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j. \tag{3.3}$$

O estimador não tendencioso para a variância de Horvitz-Thompson usando a amostra de células de tamanho n tem a seguinte expressão:

$$\widehat{Var}(\widehat{\tau}_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j.$$

Segundo Horvitz et al. [26], a amostragem com probabilidade desigual de seleção pode reduzir consideravelmente a variância dos estimadores quando comparados com planos que assumem probabilidades iguais de seleção.

3.3.2 Estimador de Hansen-Hurwitz

Uma alternativa aos casos em que a amostra inicial z_1 seja extraída por uma AAS com reposição foi estudada por Hansen et al. [25], o estimador de Hansen-Hurwitz - HH. Esse estimador é não tendencioso para o total da população e é possível ver sua expressão a seguir:

$$\widehat{\tau}_{HH} = \frac{1}{n} \sum_{i \in s} \frac{y_i}{p_i}, \tag{3.4}$$

sendo $s = \{1, 2, ..., n\}$ o conjunto das células visitadas, y_i é o número total de elementos de interesse na célula i e p_i a probabilidade de seleção da i-ésima célula da população, para $i = 1, \ldots, N$.

E sua variância é dada por:

$$Var(\widehat{\tau}_{HH}) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - \tau\right)^2. \tag{3.5}$$

Uma vez que τ é um parâmetro populacional (número total de pontos na grade) e portanto desconhecido, $Var(\widehat{\tau}_{HH})$ é também um parâmetro populacional que precisa ser estimado. O estimador não tendencioso dessa variância tem a seguinte expressão:

$$\widehat{Var}(\widehat{\tau}_{HH}) = \frac{1}{n(n-1)}\sum_{i\in s} \left(\frac{y_i}{p_i} - \widehat{\tau}_{HH}\right)^2.$$

3.4 Critério de Parada da AAC

No plano amostral adaptativo, como descrito por Thompson [48], inicia-se por uma Amostragem Aleatória Simples - AAS sem reposição ou com reposição de z_1 células. No momento em que um elemento da população é encontrado dentro de qualquer uma das células z_1 sorteadas inicialmente, a AAC orienta seguir observando as células vizinhas a elas, até chegar às células que não contenham nenhum elemento de interesse (são as chamadas células de borda). Portanto, a AAC utiliza sua própria estrutura para finalizar o processo, pois o critério de parada está na chegada às células de borda.

Capítulo 4

Método de Captura e Recaptura

Uma solução aos problemas de contagem é usar o Método de Captura e Recaptura no qual os elementos da população são capturados, marcados e liberados a população de origem para futuras recapturas como explicado por Otis et al. [35]. Dentro do Método de Captura e Recaptura, vale destacar que o termo capturado é usado para expressar a primeira captura de cada elemento e o termo recapturado é usado para denominar os elementos capturados duas ou mais vezes.

Dunn et al. [20] afirmaram que o Método de Captura e Recaptura, além de prover estimativas relevantes, tem a vantagem de consumir menos tempo e ser mais barato. Tal fato justifica-se porque não existe a necessidade de contar todos os elementos com a característica desejada dentro de uma região.

Nas Seções 4.1 e 4.2, serão apresentados tanto o Método da Captura e Recaptura Simples como o Método da Captura e Recaptura Múltipla. Apesar de estarem divididos em seções diferentes, ambos os métodos existem as seguintes condições fundamentais: coletas amostrais precisam ser aleatórias e independentes entre elas; todos os elementos precisam ter probabilidades iguais de serem capturados; e entre as recapturas, a identificação aplicada ao elemento de pesquisa não poderá desaparecer ao longo de todo o estudo.

4.1 Método de Captura e Recaptura Simples

O Método de Captura e Recaptura Simples - MCRS é usado para estimar o total populacional. A estrutura desse método está baseada na extração de apenas duas amostras aleatórias independentes e representativas da população. Suponha que haja interesse em estimar o tamanho total N de uma população que não apresenta mudança ao longo do

tempo, ou seja, nenhum elemento entra ou sai através de nascimento, morte, imigração ou emigração.

Um número de elementos da população n_1 são capturados no primeiro dia, marcados de alguma forma com a finalidade de serem identificados na próxima etapa e liberados. Na etapa seguinte, espera-se algum tempo podendo ser horas, dias ou conforme interesse da pesquisa para os elementos marcados na etapa inicial e os não marcados se misturem. Posteriormente a um intervalo de tempo finito, mais elementos são capturados e serão denotados por n_2. Seja m_1 o número de elementos marcados antes na primeira amostra, portanto $m_1 = 0$. Vale notar que na segunda amostra de n_2 elementos podem haver m_2 já marcados e eles serão chamados de recapturados. Tem-se que $n_2 - m_2$ serão os capturados pela primeira vez.

Antes de introduzir os estimadores utilizados no MCRS, dependendo do estimador, os fatores que causam estimadores viesados, isto é, quando o valor esperado do estimador é diferente do parâmetro $E[\widehat{N}] \neq N$ foram discutidos por Coeli et al. [14] e apresentados a seguir:

- Se as amostras não forem independentes, $\widehat{N}$ será viesado.
- Se a probabilidade de captura na segunda amostra for maior do que na primeira, $\widehat{N}$ irá subestimar N.
- Se a probabilidade de captura na segunda amostra será menor do que na primeira, $\widehat{N}$ irá superestimar N.

4.1.1 Estimador de Lincoln-Petersen

Esse estimador foi introduzido por Lincoln [31] e composto por duas coletas: a captura que é uma amostra aleatória sem reposição; e uma única reamostragem da população correspondendo a recaptura com tamanho igual ou diferente da dimensão da primeira amostra.

Antes de aplicar este estimador, salientaram Briand et al. [7] a necessidade de observar certas suposições que podem diferir para diferentes populações, por exemplo, as inspeções de biologia podem ser diferentes das inspeções de software, mas as suposições básicas para utilizar o estimador Lincoln-Petersen são as mesmas, a seguir:

- A população é fechada;

- Todos os elementos têm a mesma probabilidade de captura; e
- Na ocasião da recaptura, todos os elementos previamente observados podem ser identificados.

O termo "população fechada"é utilizado, pois é um dos pressupostos dos estimadores apresentados. Esse pressuposto é usado para facilitar as implementações, simulações e comparações. Por exemplo, Teo [47] tinha o objetivo de comparar algoritmos e para isso fixou o tamanho de sua população em 10, 20, 30, 50 e 100 elementos.

Seja $t = 1$ o instante da primeira amostra. Define-se n_1 como sendo o número de elementos pertencentes à amostra, no instante $t = 1$. Agora, seja $t = 2$ o instante da segunda amostra. Define-se n_2 como o número de elementos da amostra no instante $t = 2$, ou seja, a quantidade de elementos da segunda amostra; e defina m_2 como o número de elementos encontrados na interseção das amostras n_1 e n_2, isto é, o número de elementos encontrados em ambas as amostras n_1 e n_2. Por construção, $m_1 = 0$, visto que não existem elementos marcados antes da primeira amostra ser coletada. Se $m_2 \neq 0$, tem-se a expressão dada pela Equação 4.1 que é o estimador do total populacional de Lincoln-Petersen.

$$\widehat{N}_{ELP} = \frac{n_1 n_2}{m_2}. \tag{4.1}$$

A Figura 4.1 apresenta a ilustração de uma população com $N = 100$, uma captura $n_1 = 12$ foi realizada e devolvida à população, posteriormente, a distribuição espacial dos elementos muda do instante $t = 1$ para o instante $t = 2$, tal fato ocorreu para que a população com $N = 100$ misturem-se e então uma nova captura foi retirada $n_2 = 8$, uma vez que nenhum elemento foi marcado antes da primeira amostra, tem-se que $m_1 = 0$ e observou-se 1 elemento em comum as duas amostras, portanto $m_2 = 1$. Caso a Equação 4.1 fosse aplicada, a estimativa seria de $\widehat{N}_{ELP} = 96$ elementos.

No caso em que $m_2 = 0$, logo $\widehat{N}_{ELP}$ é infinito, o estimador de Lincoln-Petersen não está bem definido, pois tem esperança e variância infinitas. Por isso, surgiu a necessidade do estimador de Chapman e do estimador de Bailey os quais propuseram modificações no estimador de Lincoln-Petersen, com a finalidade de obter estimadores com média e variância finitas.

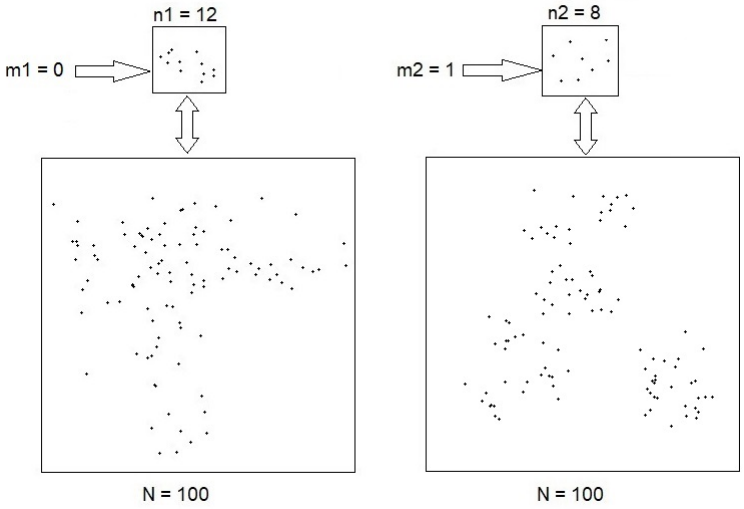


Figura 4.1: Ilustração da estrutura do estimador de Lincoln-Petersen.

4.1.2 Estimador de Chapman

O estimador de Chapman - EC, o qual foi proposto por Chapman [17], é considerado não tendencioso para valores elevados de n_1 e n_2. Oliveira [34] demonstrou que $E[\widehat{N}] = N$ e que a estimativa da variância é sempre finita. Para suprir as deficiências do estimador de Lincoln-Petersen no caso de $m_2 = 0$, tem-se o estimador de Chapman com a notação semelhante a do estimador de Lincoln-Petersen, dada pela Equação 4.1.2, a seguir:

$$\widehat{N}_{EC} = \frac{(n_1+1)(n_2+1)}{(m_2+1)} - 1.$$

A variância estimada é dada por:

$$\widehat{var}(\widehat{N}_{EC}) = \widehat{var}\left(\frac{(n_1+1)(n_2+1)}{(m_2+1)} - 1\right) = \widehat{var}\left(\frac{(n_1+1)(n_2+1)}{(m_2+1)}\right).$$

A segunda igualdade é válida por propriedade de variância na qual tem-se que a variância de constante é igual a zero.

4.1.3 Estimador de Bailey

Com o objetivo de eliminar o viés do estimador de Lincoln-Petersen para o total populacional, Bailey [5] propôs o estimador de Bailey - EB dado por:

$$\widehat{N}_{EB} = \frac{n_1(n_2+1)}{m_2+1}.$$

Contudo, este estimador traz a estimativa com viés relativo o qual pode ser sério mesmo se o tamanho da amostra é bastante grande, conforme apresentado por Bailey [5].

4.2 Método de Captura e Recaptura Múltipla

O Método de Captura e Recaptura Múltipla - MCRM foi proposto por Schnabel [37], visando a extração de um número maior que dois de amostras independentes, visto que a utilização de apenas uma captura e uma recaptura pode não ser suficiente nos casos em que número de recapturados seja igual a zero como visto na Seção 4.1. A ilustração do MCRM é apresentada na Figura 4.2.

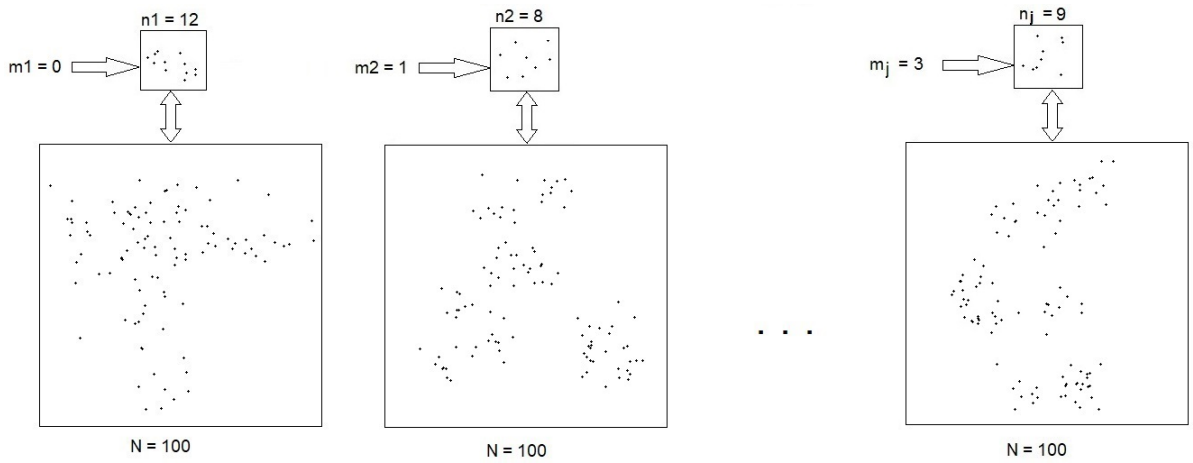


Figura 4.2: Ilustração do Método de Captura e Recaptura Múltipla.

Vale a pena enfatizar que o Método de Captura e Recaptura Múltipla visa obter estimativas mais precisas para realizar inferências sobre a população. Por outro lado, múltiplas recapturas podem ser uma tarefa bastante difícil por depender da capacidade de acesso à população, aumentar os custos de pesquisa e o tempo computacional.

Notação	Descrição
N	número total de elementos da população
k	número máximo de recapturas
j	número de amostras coletadas, onde $j = \{1, 2, ..., k\}$
n_j	número de elementos a cada amostra j, $j = \{1, 2, ..., k\}$
m_j	número de elementos recapturados na amostra j de tamanho n_j

Tabela 4.1: Notações específicas para Método Captura e Recaptura Múltipla.

Esse método é composto por um vetor C_1 que representa a captura de elementos e R recapturas as quais são denotadas pelo conjunto de vetores $R = \{R_1, R_2, R_3, ..., R_k\}$, os vetores $M = \{M_1, M_2, ..., M_k\}$ para contabilizar os novos elementos que aparecem a cada recaptura e as outras notações específicas são dadas pela Tabela 4.1. Além disso, o Método de Captura e Recaptura Múltipla - MCRM conta com estimadores próprios, tais como estimador de Schnabel e estimador de Schumacher e Eschmeyer.

4.2.1 Estimador de Schnabel

O estimador de Schnabel - ES foi desenvolvido por Schnabel [37] 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 pela Equação 4.2.

$$\widehat{N}_{schn} = \frac{\sum_{j=2}^{k} n_j M_j}{\sum_{j=2}^{k} m_j}. \tag{4.2}$$

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

A variância do estimador de Schnabel é apresentada como $1/\widehat{N}_{schn}$:

$$Var(1/\widehat{N}_{schn}) = \frac{\sum_{j=2}^{k} m_j}{(\sum_{j=2}^{k} n_j M_j)^2}.$$

A Tabela 4.2 contém a variável j, onde $j = 1$ representa a captura inicial e para $j = \{2, 3, 4, ..., k\}$ tem-se as recapturas, a variável M_j representa o número total de elementos distintos capturados com pelo menos uma marca durante todo o processo, as variáveis $\widehat{n}_1$, $\widehat{n}_2$, ..., $\widehat{n}_{k-1}$ são as estimativas de Schnabel a cada recaptura e as demais variáveis n_j e m_j são o número de elementos na j-ésima amostra e o número de elementos encontrados em ambas as amostras, respectivamente. Uma vez que não existem elementos marcados na primeira amostra, tem-se que $m_1 = 0$, $M_1 = 0$ e $M_2 = n_1$.

Trata-se de uma generalização do estimador de Lincoln-Petersen, mas no caso em que existe a possibilidade de extrair mais do que duas amostras de diferentes tamanhos ou não para n_j. A estrutura desse estimador leva em consideração uma amostra inicial de n_1 elementos extraídos e marcados, em seguida, devolvidos à população de origem;

j	n_j	m_j	M_j	$\widehat{N}_{schn}$
1	n_1	m_1	M_1	...
2	n_2	m_2	$M_2 = M_1 + n_1 - m_1$	$\widehat{n}_1 = \frac{n_2 M_2}{m_2}$
3	n_3	m_3	$M_3 = M_2 + n_2 - m_2$	$\widehat{n}_2 = \frac{n_2 M_2 + n_3 M_3}{m_2 + m_3}$
:	:	:	:	:
$j+1$	n_{j+1}	m_{j+1}	$M_{j+1} = M_j + n_j - m_j$	$\widehat{n}_j = \frac{n_2 M_2 + n_3 M_3 + ... + n_{j+1} M_{j+1}}{m_2 + m_3 + ... + m_{j+1}}$
:	:	:	:	:
k	n_k	m_k	$M_k = M_{k-1} + n_{k-1} - m_{k-1}$	$\widehat{n}_{k-1} = \frac{n_2 M_2 + n_3 M_3 + ... + n_k M_k}{m_2 + m_3 + ... + m_k}$

Tabela 4.2: Estruturação das variáveis referente ao estimador de Schnabel.

posteriormente, recolhe-se uma segunda amostra de tamanho n_2, anota-se o número de elementos já marcados em n_1 na variável m_2, e marcam-se novamente todos os elementos; o processo repete-se um determinado número k de vezes.

4.2.2 Estimador de Schumacher-Eschmeyer

O estimador de Schumacher-Eschmeyer - ESE foi proposto por Schumacher et al. [38]. A Equação 4.3 fornece a expressão do estimador de Schumacher-Eschmeyer, a diferença em relação a Equação 4.2 está na multiplicação de M_j no numerador e no denominador.

$$\widehat{N}_{schEsch} = \frac{\sum_{j=2}^{k} n_j M_j^2}{\sum_{j=2}^{k} m_j M_j}. \tag{4.3}$$

Assim como o estimador de Schnabel, o estimador de Schumacher-Eschmeyer não está definido para todos os valores de j, pois se o denominador $\sum_{j=2}^{k} m_j M_j$ for igual a zero, o estimador de Schumacher-Eschmeyer não está matematicamente definido.

4.3 Critérios de Parada para o Método de Captura e Recaptura Múltipla

Os critérios de parada têm sido usados para fornecer auxílio à tomada de decisão. A escolha incorreta do momento de parada pode adicionar viés aos resultados da pesquisa como discutido por Dalal et al. [16], 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 a 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 a insatisfação dos clientes; e se o teste continuar até o limite máximo permitido, existe o custo do esforço de teste elevado".

Zielinski el al. [54] reafirmaram as hipóteses apresentadas por Dalal et al. [16] em seu trabalho sobre os algoritmos de otimização cujo objetivo principal era a convergência e o objetivo secundário era usar o mínimo esforço computacional. Estudos recentes sobre os métodos de parada precoce foram propostos por Streeter [45], Golovin et al. [23], Domhan et al. [19] e Yao et al. [53] usando modelos paramétricos e não-paramétricos, visando refutar a primeira hipótese apontada por Dalal et al. [16].

Conforme Costanza et al. [15], a regra de parada mais comumente empregada nos anos 70 era o teste F para determinar a distância adequada entre cada resultado e a parada ocorre logo antes do primeiro resultado não significativo. No artigo de Chao et al. [12], foi apresentado o problema de definir critério de parada para teste de software podendo parar a qualquer momento $t > 0$ (caso contínuo) ou parar quando um erro é encontrado (caso discreto) e estimar o número de erros, ou seja, falhas no software, levando em consideração as seguintes variáveis: (i) custo por unidade de tempo; (ii) custo da ocorrência de um erro com o usuário; (iii) coeficiente de variação das taxas de falha; (iv) tempo total esperado de uso do software; e (v) tempo total do usuário até a próxima revisão.

É possível encontrar muitos critérios de parada disponíveis, mas não foi possível evidenciar estudos conclusivos sobre regras de parada com foco no Método de Captura e Recaptura Múltipla. Briand et al. [7] questionaram sobre qual seria o melhor momento para que as suas inspeções de software pararem ou se deveriam continuar até atingir um nível adequado de qualidade, mas não chegaram a desenvolver uma solução. Assim como o artigo de Smith et al. [43] que apontou a dificuldade em determinar em qual ocasião que as recapturas devem cessar, mas ainda que plotando função de risco e ganhos para a estimativa, não concluiu sobre critério de parada.

El et al. [21], durante as suas reinspeções, afirmaram que o momento de parar as inspeções é um problema e que os estimadores do Método de Captura e Recaptura nem sempre são precisos para decidir se devem parar ou voltar a inspecionar e adicionaram que as organizações precisam definir seus limites de eficácia para suas inspeções. Um caso particular apresentado foi aplicado as inspeções de software, lembrando que se elas excederem o limite superior ou interior tornam propício para decidir em parar as inspeções. Hwang et al. [27] trabalharam com a fixação de cinco tempos de parada, ou seja, $t =$

$\{1; 1,25; 1,5; 2; 4\}$ e apenas afirmaram que se o tempo de parada for relativamente curto o estimador se comporta de maneira instável, por outro lado, quando o tempo de parada é aumentado e existem mais dados disponíveis, o estimador proposto supera o anterior em relação ao viés.

No artigo de Accettura et al. [1], o critério de parada foi determinado *a priori*, pois o número de recapturas foi decidido de antemão e com tamanhos de amostra fixados em 1. Entretanto, se o critério de parada proposto por Accettura et al. [1] fosse utilizado com o objetivo de interromper o MCRM quando um número predeterminado de elementos novos fosse capturado ou quando um número predeterminado de recapturas fosse alcançado, forneceria um critério de parada subjetivo.

4.3.1 Usando Coeficiente de Variação

O coeficiente de variação - CV é uma medida normalizada de dispersão de uma distribuição de probabilidade definida como a razão entre o desvio padrão σ e a média μ. Essa medida é amplamente utilizada para comparar conjuntos de dados com diferentes unidades. Tem-se a sua expressão na Equação 4.4.

$$CV = \frac{\sigma}{\mu}. \tag{4.4}$$

Entretanto, o coeficiente de variação pode ser obtido através de dados amostrais $\widehat{\gamma}$ onde $\widehat{\gamma} \in [0, +\infty[$ é definido como:

$$\widehat{\gamma} = \frac{S}{\bar{X}}, \tag{4.5}$$

onde S é o desvio padrão amostral e $\bar{X}$ é a média obtida a partir dos dados amostrais.

Yang et al. [52] definiram o $CV \leq 0,5$ como critério de parada ótimo, mas essa escolha pode levar a um viés no procedimento, visto que para algumas populações esse valor determinístico não contribuirá a uma correta tomada de decisão. Vale mencionar que o critério de parada ótimo é aquele que possui o melhor custo benefício entre tempo de pesquisa e estimativa mais próxima do valor verdadeiro do parâmetro.

Castagliola et al. [11] avaliaram o CV como uma abordagem bem-sucedida dentro de Controle Estatístico do Processo, mesmo quando a média do processo e o desvio padrão não são constantes. Dalal et al. [16] também apresentaram o uso do coeficiente de variação na Equação 4.5 com o propósito de parar um experimento.

4.3.2 Usando Intervalo de Confiança

Uma abordagem promissora é usar a meia largura do Intervalo de Confiança - IC como critério de parada, conforme proposto por Singham et al. [40] [41] [42]. Seja $X = \{X_1, X_2, ..., X_k\}$ uma amostra de tamanho k do total populacional. Seja $\bar{X}$ uma estimativa para 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:

$$\left[\bar{X} - t_{\eta,k-1}\sqrt{\frac{S_k^2}{k}}, \bar{X} + t_{\eta,k-1}\sqrt{\frac{S_k^2}{k}}\right]. \tag{4.6}$$

Define-se $MK_{\eta,k}$ como a meia largura do IC:

$$MK_{\eta,k} = t_{\eta,k-1}\sqrt{\frac{S_k^2}{k}}. \tag{4.7}$$

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 por Singham et al. [40] [41] [42] afirmou que é necessário escolher um valor de precisão desejado, parar quando a meia largura do IC, ou seja, a variável $MK_{\eta,k}$ for menor ou igual a δ e define-se k^* como a iteração na qual ocorre a parada de um processo sequencial.

$$k^* = \arg\min_{k>0} MK_{\eta,k} \leq \delta, \tag{4.8}$$

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

Usando a Inequação 4.8, 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. Seja $\widehat{n}^*$ a estimativa na k^*-ésima recaptura. A seguir, são descritas as etapas para implementação do critério de parada com a garantia de confiança η baseada na Inequação 4.8 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, a variância e a meia largura das estimativas obtidas até k.

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.

A adequação da proposta de Singham et al. [40] [41] [42] ao contexto dessa dissertação ao MCRM foi possível devido ao valor de δ poder ser desenvolvido a partir da primeira estimativa válida dos estimadores usuais, tal como o estimador de Schnabel.

$$\delta = \widehat{n}_1 \varepsilon, \tag{4.9}$$

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

A Equação 4.9 possibilita utilizar a proposta de Singham et al. [40] [41] [42] 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). Tal fato, viabiliza um critério de parada ao MCRM menos subjetivo em comparação a proposta de Yang et al. [52] em que existe apenas uma variável predeterminada.

Capítulo 5

Framework Proposto

A partir da observação de uma premissa que se tornou uma lacuna 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 - 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 dificuldade de captura, por exemplo, por estarem escondidos ou misturados a outra população com grande número de elementos sem a variável de interesse em questão.

5.1 Metodologia

Esse framework proposto, M2C2E, como o próprio nome do método aponta, é o processo que contém duas camadas e dois estimadores. Inicia-se de forma semelhante a AAC, sobrepondo uma grade a população 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 $\widehat{y}_i$ dentro da célula i selecionada.

Na Figura 5.1, é possível observar que 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 ou o estimador de Schumacher-Eschmeyer.

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

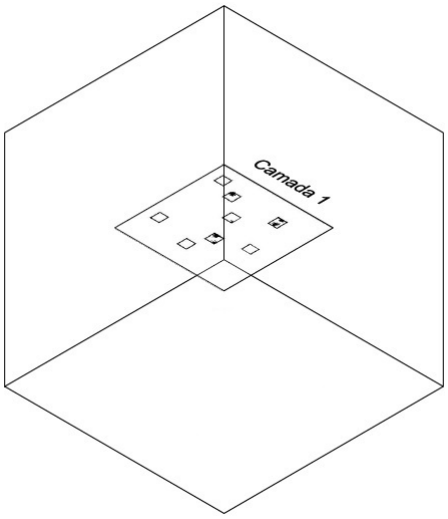


Figura 5.1: Ilustração do Método 2-Camadas e 2-Estimadores na camada 1.

proposto por Singham et al. [40] [41] [42]. 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 5.1.1.

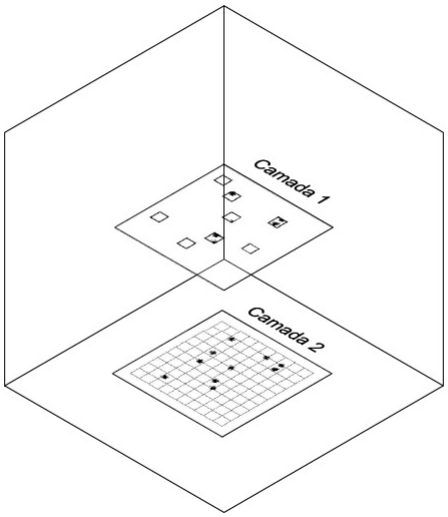


Figura 5.2: Ilustração do Método 2-Camadas e 2-Estimadores na camada 1 e na camada 2.

Na Figura 5.2, as camadas 1 e 2 são apresentadas simultaneamente com a ilustração de seus conteúdos. Na camada 2, ocorre o final do método resultando no dimensionamento do total populacional na grade. A Figura 5.2 contém a imagem da camada 1 que é um

recorte de 8 células da camada 2 e a imagem da camada 2 que representa uma rede sintética sobreposta por uma grade 10x10.

Notação	Descrição
τ	total populacional
N	tamanho na grade, ou seja, número de células
n	número total de células selecionadas
z_1	tamanho inicial da amostra de células
n_j	número de elementos a cada amostra j, $j = \{1, 2, ..., k\}$

Tabela 5.1: Notações para o M2C2E.

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, enquanto 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. A Tabela 5.1 contém as notações a serem utilizadas no M2C2E.

5.1.1 Estimadores Modificados

Seja $\widehat{N}^*_{schn}$ o estimador de Schnabel até a última recaptura a qual ocorre na k^*-ésima recaptura apresentado na Equação 5.1, onde a variável k^* é obtida pelo critério de parada proposto Singham et al. [40] [41] [42] representado pela Inequação 4.8 e pela Equação 4.9. Seja $\widehat{n}^*_{schn_i}$ a estimativa de Schnabel na k^*-ésima recaptura para a célula i obtida através das informações trazidas pela amostra ao estimador de Schnabel na última recaptura.

$$\widehat{N}^*_{schn} = \frac{\sum_{j=2}^{k^*} n_j M_j}{\sum_{j=2}^{k^*} m_j}. \tag{5.1}$$

No estimador de Horvitz-Thompson modificado do total populacional $\widehat{\tau}_{HT_mod}$, tem-se que o parâmetro y_i, que significa o número total de elementos na célula i, é substituído por $\widehat{n}^*_{schn_i}$, é dado por:

$$\widehat{\tau}_{HT_mod} = \sum_{i=1}^{N} \frac{\widehat{n}^*_{schn_i}}{\pi_i}, \tag{5.2}$$

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

Proposição 1: Se $\pi_i > 0$, onde $i = (1, 2, ..., N)$, então $\widehat{\tau}_{HT_mod} = \sum_{i=1}^{N} \dfrac{\widehat{n}^*_{schn_i}}{\pi_i}$ é o estimador de τ e sua variância é dada por:

$$Var(\widehat{\tau}_{HT_mod}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} \widehat{n}^{*2}_{schn_i} + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} \widehat{n}^*_{schn_i} \widehat{n}^*_{schn_j}.$$

Prova: Seja I_i a variável indicadora de inclusão da célula i na amostra que recebe o valor 1, se a i-ésima unidade é amostrada e recebe 0, caso contrário. Dessa forma, I_i segue uma distribuição Bernoulli, com probabilidade π_i. Tem-se que: $E(I_i) = \pi_i$ e $V(I_i) = \pi(1 - \pi)$.

Seja o produto $I_i I_j$ igual a 1, se a i-ésima e a j-ésima unidades aparecerem simultaneamente na amostra. Considerando $Cov(I_i I_j)$ como a definição,

$$Cov(I_i I_j) = E(I_i I_j) - E(I_i)E(I_j) = \pi_{ij} - \pi_i \pi_j.$$

Sendo $\widehat{n}^*_{schn_i}$ a estimativa de Schnabel no critério de parada para cada célula i, portanto $\widehat{n}^*_{schn_i}$ é um valor fixo e I_i como variável aleatória com distribuição Bernoulli:

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

$\widehat{\tau}_{HT_mod}$ será um estimador não tendencioso para τ, se $\widehat{n}^*_{schn_i}$ convergir para y_i e $E(I_i) = \pi_i$.

$$\begin{aligned} V(\widehat{\tau}_{HT_mod}) &= \sum_{i=1}^{N} \left(\frac{\widehat{n}^*_{schn_i}}{\pi_i} \right)^2 V(I_i) + 2 \sum_{i=1}^{N} \sum_{j>1}^{N} \frac{\widehat{n}^*_{schn_i}}{\pi_i} \frac{\widehat{n}^*_{schn_j}}{\pi_j} Cov(I_i I_j) = \\ &= \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} \widehat{n}^{*2}_{schn_j} + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} \widehat{n}^*_{schn_i} \widehat{n}^*_{schn_j}, \end{aligned}$$

como prova da Proposição 1.

O estimador de Hansen-Hurwitz modificado para o total da população $\widehat{\tau}_{HH_mod}$ é possível ver sua expressão na Equação 5.3 a seguir:

$$\widehat{\tau}_{HH_mod} = \frac{1}{n} \sum_{i=1}^{N} \frac{\widehat{n}^*_{schn_i}}{p_i}, \tag{5.3}$$

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

Proposição 2: Se a variável n representa o número total de células visitadas na grade e $pi > 0$, onde $i = 1, 2, ..., N$, então $\widehat{\tau}_{HH_mod} = \frac{1}{n}\sum_{i=1}^{N}\frac{\widehat{n}^*_{schn_i}}{p_i}$ é o estimador de τ e sua variância é dada por:

$$Var(\widehat{\tau}_{HH_mod}) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{\widehat{n}^*_{schn_i}}{p_i} - \widehat{\tau}_{HH_mod}\right)^2.$$

Prova: Seja W_i o número de vezes que a i-ésima célula da grade aparece no estimador. Tem-se que W_i segue uma distribuição Binomial com $E(W_i) = np_i$ e $V(W_i) = np_i(1-p_i)$.

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

$\widehat{\tau}_{HH_mod}$ será um estimador não tendencioso para τ, se $\widehat{n}^*_{schn_i}$ convergir para y_i e $E(W_i) = np_i$.

Sendo $V(\widehat{\tau}_{HH_mod}) = E(\widehat{\tau}^2_{HH_mod}) - E(\widehat{\tau}_{HH_mod})^2$.

$$\begin{aligned} V(\widehat{\tau}_{HH_mod}) &= V\left(\frac{1}{n}\sum_{i=1}^{N}\frac{\widehat{n}^*_{schn_i}}{p_i}\right) = \left[\frac{1}{n}\sum_{i=1}^{N}\frac{\widehat{n}^{*2}_{schn_i}p_i}{p_i^2} - \left(\frac{1}{n}\sum_{i=1}^{N}\frac{\widehat{n}^*_{schn_i}}{p_i}\right)^2\right] = \\ &= \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{\widehat{n}^*_{schn_i}}{p_i} - \widehat{\tau}_{HH_mod}\right)^2, \end{aligned}$$

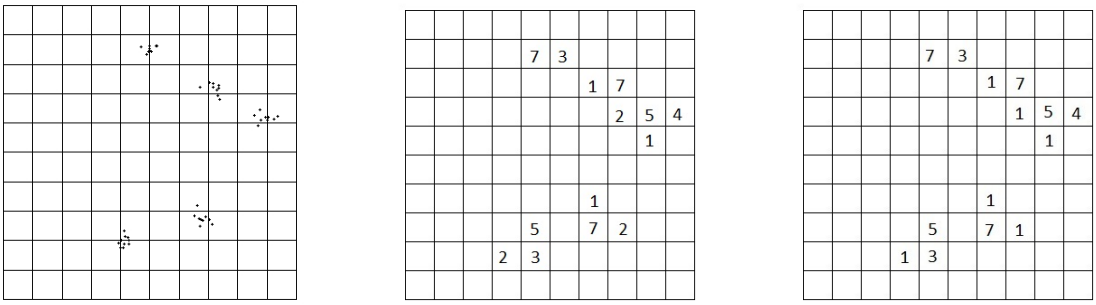
como prova da Proposição 2.

5.1.2 Etapas do M2C2E com Ilustrações

Considere uma população sobreposta por uma grade 10x10 (100 células) conforme a Figura 5.3. As estimativas de Schnabel na parada com $\varepsilon = 5\%$ na Figura 5.3 (c) foram obtidas a partir do algoritmo em Apêndice A2 - Estimador de Schnabel e Critério de Parada.

Partindo do princípio que os elementos de interesse foram sobrepostos por uma grade e cada célula tem o seu valor de k^* de forma independente, aplicam-se para a implementação do M2C2E as etapas a seguir:

1. Selecionar z_1 células da grade a partir de uma AASs ou AASc (Figura 5.4).



(a) Representação com pontos

(b) Representação com os totais reais

(c) Representação com as estimativas de Schnabel na parada com $\varepsilon = 5\%$

Figura 5.3: População sintética sobreposta por uma grade com $N = 100$ células e $\tau = 50$ elementos de interesse.

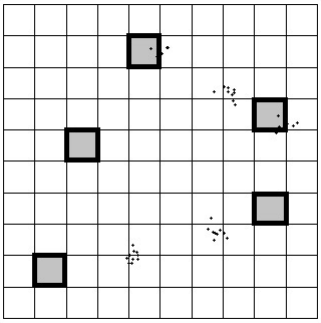


Figura 5.4: População sintética sobreposta por uma grade com 100 células com seleção de 5 células iniciais.

2. Se houver elementos de interesse nas células selecionadas na etapa 1, as células que compartilham lados com elas são observadas (Figura 5.5), aplicar nelas o estimador Schnabel até chegar à k^*-ésima recaptura na Equação 5.1 e obter as estimativas $\widehat{n}^*_{schn_i}$ (Figura 5.3 (c)).

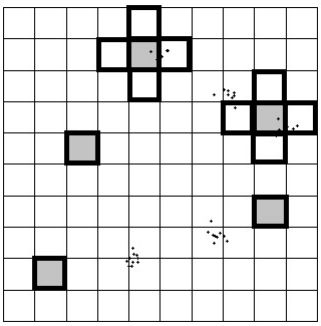


Figura 5.5: População sintética sobreposta por uma grade com 100 células com seleção de 5 células iniciais e com as vizinhas imediatas.

3. Adicionar as células vizinhas das vizinhas (Figura 5.6), novamente aplicar o estimador de Schnabel até chegar à k^*-ésima recaptura na Equação 5.1 e obter as estimativas $\widehat{n}^*_{schn}$ para todas elas (Figura 5.3 (c)) até chegar em células vizinhas que não contêm elementos de interesse.

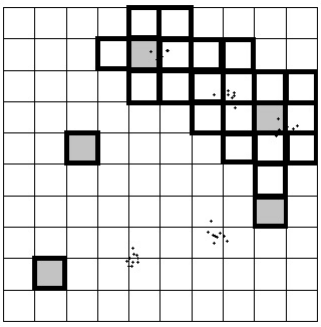


Figura 5.6: População sintética sobreposta por uma grade com 100 células com seleção de 5 células iniciais e com todas as vizinhas.

4. Inserir $\widehat{n}^*_{schn_i}$ ao invés de y_i ao estimador de Horvitz-Thompson na Equação 3.2 ou ao estimador de Hansen-Hurwitz na Equação 3.4 e retornar a estimativa para o total populacional na grade usando $\widehat{\tau}_{HT_mod}$ ou $\widehat{\tau}_{HH_mod}$ de acordo com as Equações 5.2 e 5.3.

5.2 Vantagens e Desvantagens

A vantagem do M2C2E é não precisar pressupor que todos os elementos dentro das células são conhecidos e possivelmente capturados. Dentre as desvantagens, a introdução de mais uma estimação dentro do processo da AAC, o número inicial de células a serem utilizadas e o número máximo de elementos a cada recaptura ou o tempo de coleta usado a cada recaptura continuam sendo predeterminados, pois o número de elementos a serem coletados a cada recaptura, por exemplo, influencia no número total de recapturas na camada 1, ou seja, quanto menor esse número, maior é a quantidade de recapturas necessárias.

Capítulo 6

Estudo com Dados Sintéticos

Neste capítulo, tem-se por objetivo gerar populações sintéticas, analisar o critério de parada para o número de recapturas, implementar e avaliar o Método de Captura e Recaptura Múltipla, o Método 2-Camadas e 2-Estimadores e o Método Ótimo. Oliveira [34] afirmou que quando se passa para a aplicação do Método de Captura e Recaptura Múltipla, o algoritmo necessário é mais complexo e mais difícil de implementar, devido ao número de amostras. Contudo, todas as implementações realizadas a seguir consideraram múltiplas amostras aleatórias e foi possível aplicar o Método de Captura e Recaptura Múltipla com até 50 recapturas.

As populações sintéticas são geradas com totais populacionais iguais a $N = \{100, 1000, 10000\}$ para observar o comportamento do critério de parada proposto por Singham et al. [40] [41] [42] ao MCRM e outras duas com $\tau = \{1000, 2000\}$ sobrepostas por grades de tamanho 10x10 (10 linhas e 10 colunas) e 20x20 (20 linhas e 20 colunas) com a finalidade de testar os métodos de estimação apresentados. As grades são formadas por células e o número de células é obtido multiplicando o número de linhas pelo número de colunas, ou seja, uma grade 10x10 tem 100 células, já uma grade 20x20 tem 400 células. Vale lembrar que no MCRM, por construção, é uma grade 1x1 e a notação para o total populacional é N, entretanto quando trata-se da AAC a notação usual é τ e para o M2C2E usa-se τ.

Todas as implementações foram realizadas no software RStudio Cloud[1] baseado em nuvem. Dois estudos com dados sintéticos foram feitos, sendo o primeiro para definir um critério de parada com foco no MCRM e, posteriormente, para comparar o M2C2E com o modelo onde é possível localizar todos os elementos dentro da célula chamado de Método Ótimo - MO.

[1] https://rstudio.cloud

6.1 Criação dos Dados Sintéticos

De modo semelhante a geração populacional em Thompson [48], as populações sintéticas utilizando o processo de Poisson apresentado por Diggle et al. [18], que consiste na geração de pontos numa região, foram criadas com a finalidade de implementar o critério de parada e avaliar o M2C2E, o MO e o MCRM. As coordenadas dos pontos iniciais são geradas a partir de uma distribuição Uniforme definida na região. Estes pontos são conhecidos como "pais"ou núcleo da rede em análise.

As coordenadas dos "filhos" (X_f, Y_f) seguem uma distribuição Normal Bivariada - NB com um vetor de médias igual as coordenadas dos "pais"(X_p, Y_p) e matriz de variância cuja variância é fixada na diagonal conforme a Expressão 6.1. Quanto maior a variância ϕ, maior é a distância dos "filhos"com relação aos pais, isto é equivalente a populações menos agrupadas.

$$\begin{bmatrix} X_f \\ Y_f \end{bmatrix} \sim NB\left(\begin{bmatrix} X_p \\ Y_p \end{bmatrix}, \phi \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right). \tag{6.1}$$

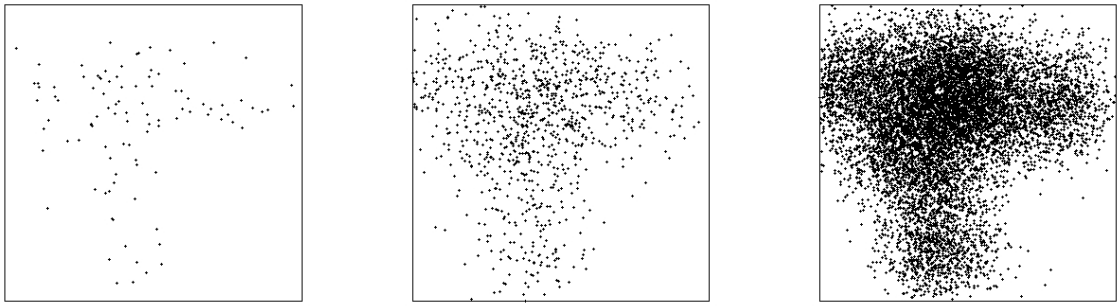
O algoritmo para realizar essa geração encontra-se no Apêndice A.1 - Geração da população rara e agrupada. Visando identificar os elementos durante a implementação, cada par ordenado das coordenadas foi atribuído a uma identificação chamada de ID do elemento, como pode ser visto no Apêndice A.1 - Para identificar os elementos.

6.2 Implementando o Critério de Parada no MCRM

Com a finalidade de analisar o critério de parada proposto por Singham et al. [40] [41] [42], foram implementadas três populações sintéticas $N = 100$, $N = 1000$ e $N = 10000$. Na Figura 6.1, todos os pontos são equiprováveis e o processo de escolha dos pontos em uma captura é aleatório com tamanho fixo ou variável. Suponha que as Figuras 6.1(a), 6.1(b) e 6.1(c) representem elementos dentro de uma célula, esse fato ajuda a visualizar os possíveis comportamentos do M2C2E na camada 1 ao variar o número total de elementos. Ainda que uma célula com 10000 elementos seja um exemplo extremo e, portanto, pouco provável para populações raras e agrupadas, é importante avaliar o critério de parada com diferentes tamanhos populacionais.

As Figuras 6.2, 6.3 e 6.4 contêm as estimativas de Schnabel a cada recaptura geradas pelo algoritmo no Apêndice A.2 - Estimador de Schnabel referente as populações sintéticas

nas Figuras 6.1(a), 6.1(b) e 6.1(c), respectivamente. De modo geral, nas Figuras 6.2, 6.3 e 6.4, a linha na horizontal indica o valor do parâmetro e a linha na vertical aponta à recaptura referente ao critério de parada que utiliza a Inequação 4.8 para δ com erro de 5% em relação ao valor da primeira estimativa válida do estimador de Schnabel com algoritmo no Apêndice A.2 - Critério de Parada.



(a) $N = 100$ (b) $N = 1000$ (c) $N = 10000$

Figura 6.1: Populações sintéticas com 100, 1000 e 10000 elementos.

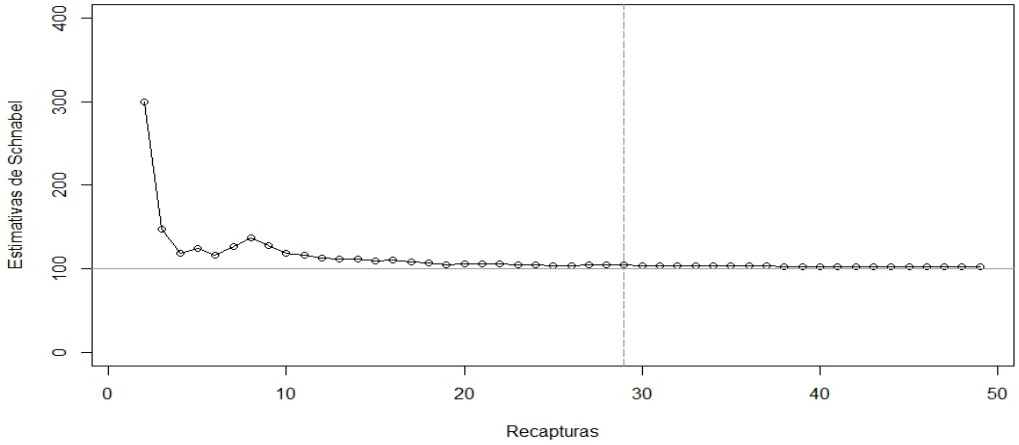


Figura 6.2: Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 10$ referente a população sintética com 100 elementos.

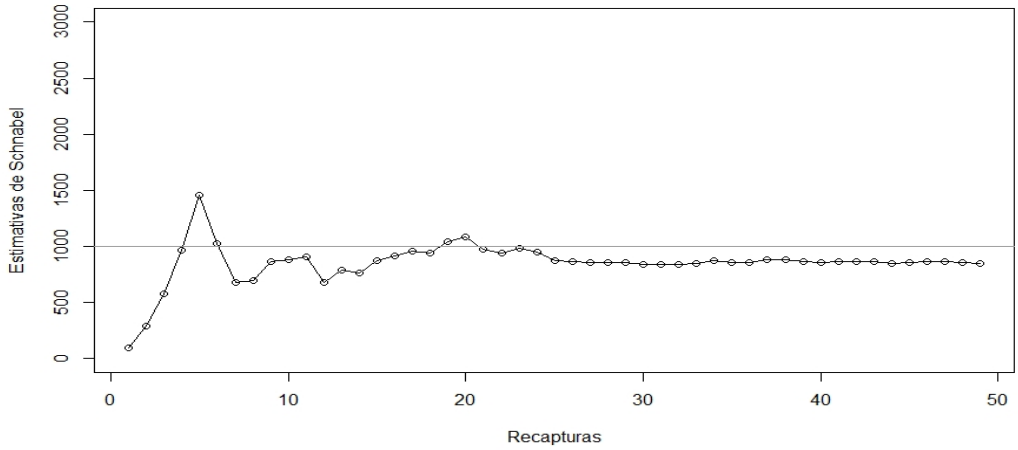


Figura 6.3: Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 10$ referente a população sintética com 1000 elementos.

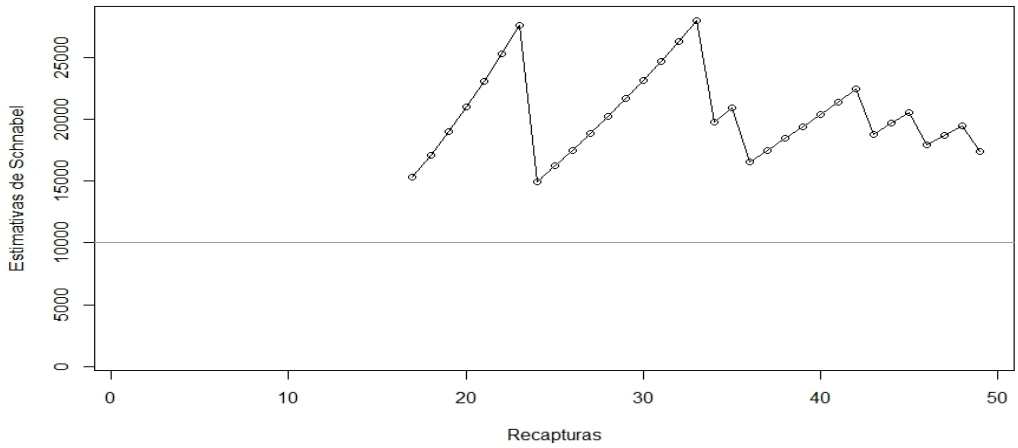


Figura 6.4: Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 10$ referente a população sintética com 10000 elementos.

Note que na Figura 6.2, a parada ocorreu na 29ª recaptura e retornou a estimativa de Schnabel na parada $\widehat{n}^*_{schn} = 104$, onde $N = 100$. Por outro lado, nas Figuras 6.3 e 6.4, a ausência da linha na vertical pode ser percebida. Essa ausência, no caso da Figura 6.3, aponta para a necessidade de mais recapturas, ou seja, a parada ocorrerá após a 50ª recaptura.

Entretanto, na Figura 6.4, uma vez que a população é muito densa, encontrar elementos em comum entre as recapturas é mais difícil, fazendo com que o número de recapturados m_j seja igual a zero. Note que as primeiras 16 estimativas não aparecem na Figura 6.4, pois $\widehat{N}_{schn}$ não está, matematicamente, bem definido e após a 16ª recaptura

resulta em estimações ruins, esse problema é referente ao estimador de Schnabel. Uma solução encontrada para esse caso é aumentar o número de elementos nas recapturas de $n_1 = n_2 = ... = 10$ para $n_1 = n_2 = ... = 100$, por exemplo. O resultado com essa solução é observado na Figura 6.5 onde a linha pontilhada na vertical representa a parada na 44ª recaptura e a estimativa resultante nela foi de $\widehat{n}^*_{schn} = 9979$ elementos usando δ com erro 10% em relação ao valor da primeira estimativa de Schnabel.

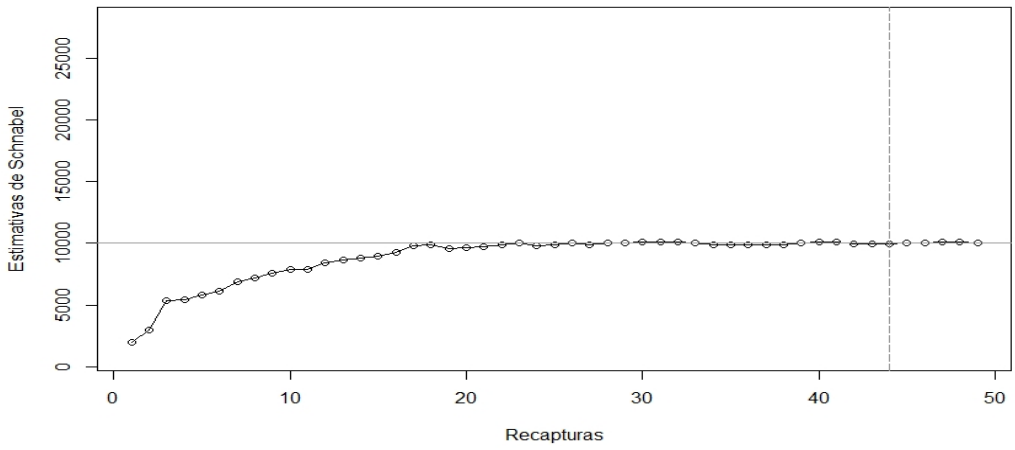
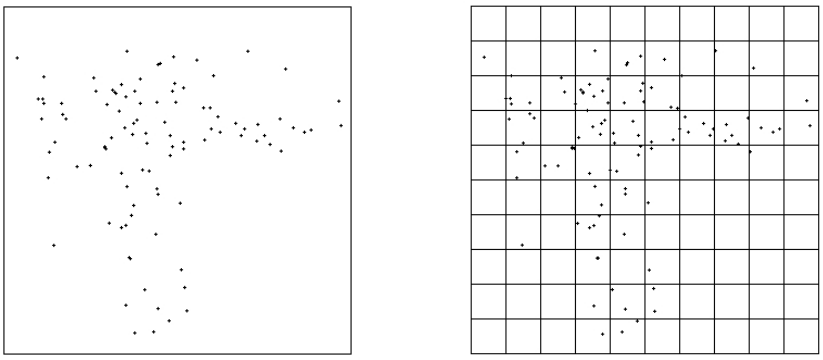


Figura 6.5: Estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 100$ referente a população sintética com 10000 elementos.

As limitações no critério de parada proposto por Singham et al. [40] [41] [42] quando aplicado ao MCRM são:

- Nos casos em que o número de elementos na rede é extremamente pequeno ($N < 100$, por exemplo), pela possibilidade de não encontrar elementos de interesse n_j.
- Em células muito densas ($N \geq 10000$), número de recapturados m_j igual a zero e, sendo assim, não obter estimativas usando o estimador de Schnabel, por exemplo.

Vale lembrar que a captura sempre ocorrerá na existência de pelo menos um elemento no MCRM. Nos casos onde o número de elementos na grade seja pequeno, a Figura 6.6 representa uma possível solução que seria uma transição para o método de populações raras tal como para a AAC. No cenário em que a população é densa, a Figura 6.5 na qual houve um aumento no número de elementos a cada recaptura $n_1, n_2, ...n_j$ de 10 para 100 apresentou uma correção plausível a utilização do critério de parada proposto por Singham et al. [40] [41] [42].



(a) MCRM com $N = 100$ (b) AAC com 100 células

Figura 6.6: Exemplo de refinamento de grade para populações extremamente pequenas saindo do MCRM grade 1x1 para a AAC grade 10x10.

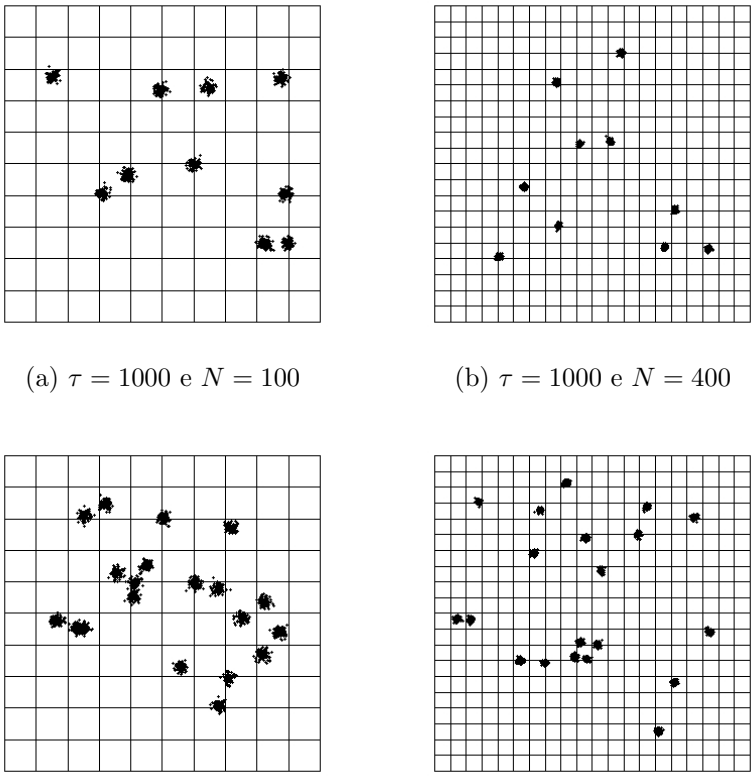
6.3 Verificando os Métodos: M2C2E, MO e MCRM

Na verificação para saber se o M2C2E pode ser uma alternativa ao AAC, quatro exemplos de populações raras e agrupadas dispostas em uma região, vistas na Figura 6.7, foram implementados seguindo a metodologia apresentada na Seção 6.1. Em particular, na Figura 6.7, a variância foi fixada em $\phi = 0,1$ a qual representa o grau de agrupamento entre os pontos.

A implementação foi realizada com 10 mil replicações para cada valor de z_1 inicial, totalizando 400 mil replicações, sendo 200 mil para o M2C2E e 200 mil para o MO, em 2 diferentes refinamentos de grade $N = 100$ e $N = 400$ e 2 diferentes números de elementos na rede $\tau = 1000$ e $\tau = 2000$. O algoritmo para implementar a camada 1 do M2C2E encontra-se no Apêndice A.2 - Método de Captura e Recaptura Múltipla, e para a camada 2 do M2C2E e o MO, utilizou-se o algoritmo em Affonso [2] apresentado no Apêndice A.3 - Amostragem Adaptativa por Conglomerados.

A Tabela 6.1 apresenta estimações para cada tamanho de amostra inicial $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$, onde N é o número de células da grade, a média das estimativas de Horvitz-Thompson foi representada por $E(\widehat{\tau}_{HT})$, o erro relativo referente a cada estimativa média $ER = \dfrac{|\tau - E(\widehat{\tau}_{HT})|}{\tau}$ representado em termo percentual, a variância amostral foi representada por $Var(\widehat{\tau}_{HT})$, o $IC_{95\%}$ é dado por 6.2, $\alpha = 0,05$ e $z_{\alpha/2} = 1,96$.

$$\left[E(\widehat{\tau}_{HT}) - z_{\alpha/2}\sqrt{\frac{Var(\widehat{\tau}_{HT})}{k}}; E(\widehat{\tau}_{HT}) + z_{\alpha/2}\sqrt{\frac{Var(\widehat{\tau}_{HT})}{k}}\right]. \tag{6.2}$$



(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 6.7: Populações sintéticas utilizadas para a comparação entre M2C2E e MO.

O valor da eficiência dos métodos são obtidos da seguinte maneira:

$$ef_{MO}^{M2C2E} = \frac{Var(\widehat{\tau}_{HT_mod})}{Var(\widehat{\tau}_{HT})}. \tag{6.3}$$

É possível observar na Tabela 6.1 os casos em que $ef_{MO}^{M2C2E} < 1$ indicando serem as estimativas do M2C2E mais eficiência que as estimativas do MO, caso contrário, as estimativas do MO são mais eficientes em relação a variabilidade do que as estimativas do M2C2E. Embora, o caso em que $\tau = 1000$ e $N = 400$ aponta para a interpretação de que o M2C2E seja mais eficiente que MO, as estimativas médias do total populacional, para todos os casos de z_1, estão fora do IC, uma vez que o valor do parâmetro é 1000, ou seja, essas estimativas médias são tendenciosas.

Ainda que nos demais cenários a ef_{MO}^{M2C2E} seja maior que 1, indicando que o MO é melhor, mas será apenas nos casos em que todos os elementos dentro da célula são encontrados. Note que o resultado do IC na Tabela 6.1, no caso em que $\tau = 2000$ e $N = 100$, contém o valor do parâmetro $\tau = 2000$ para todos os valores de z_1, no caso $\tau = 2000$ e $N = 400$ para $z_1 = 1\%N$ contém $\tau = 2000$ e quando $\tau = 1000$ e $N = 100$ para

$z_1 = 1\%N$ e $z_1 = 5\%N$ também contém o valor do parâmetro que é $\tau = 1000$. Os boxplots das variáveis ER e ef_{MO}^{M2C2E} nas Figuras 6.8 e 6.9 indicam que o M2C2E ajusta-se melhor em grade com menor número de células e maiores valores de τ.

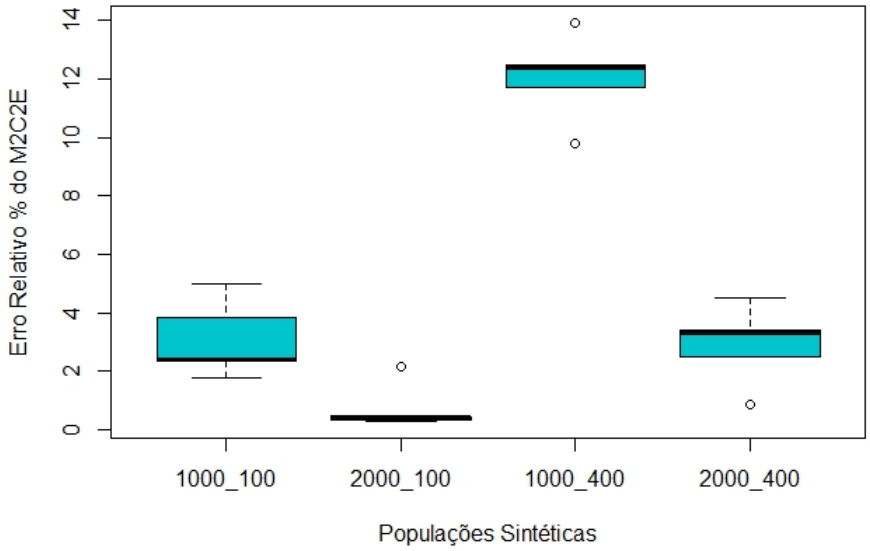


Figura 6.8: Boxplots com os erros relativos das estimativas médias do M2C2E usando o estimador de Horvitz-Thompson modificado das populações sintéticas $\tau = 1000$ e $N = 100$, $\tau = 1000$ e $N = 400, \tau = 2000$ e $N = 100$, $\tau = 2000$ e $N = 400$.

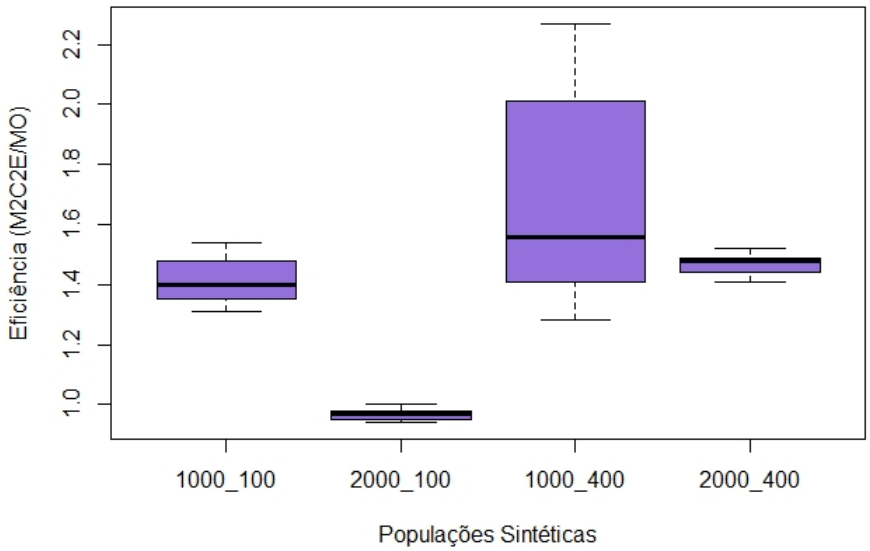
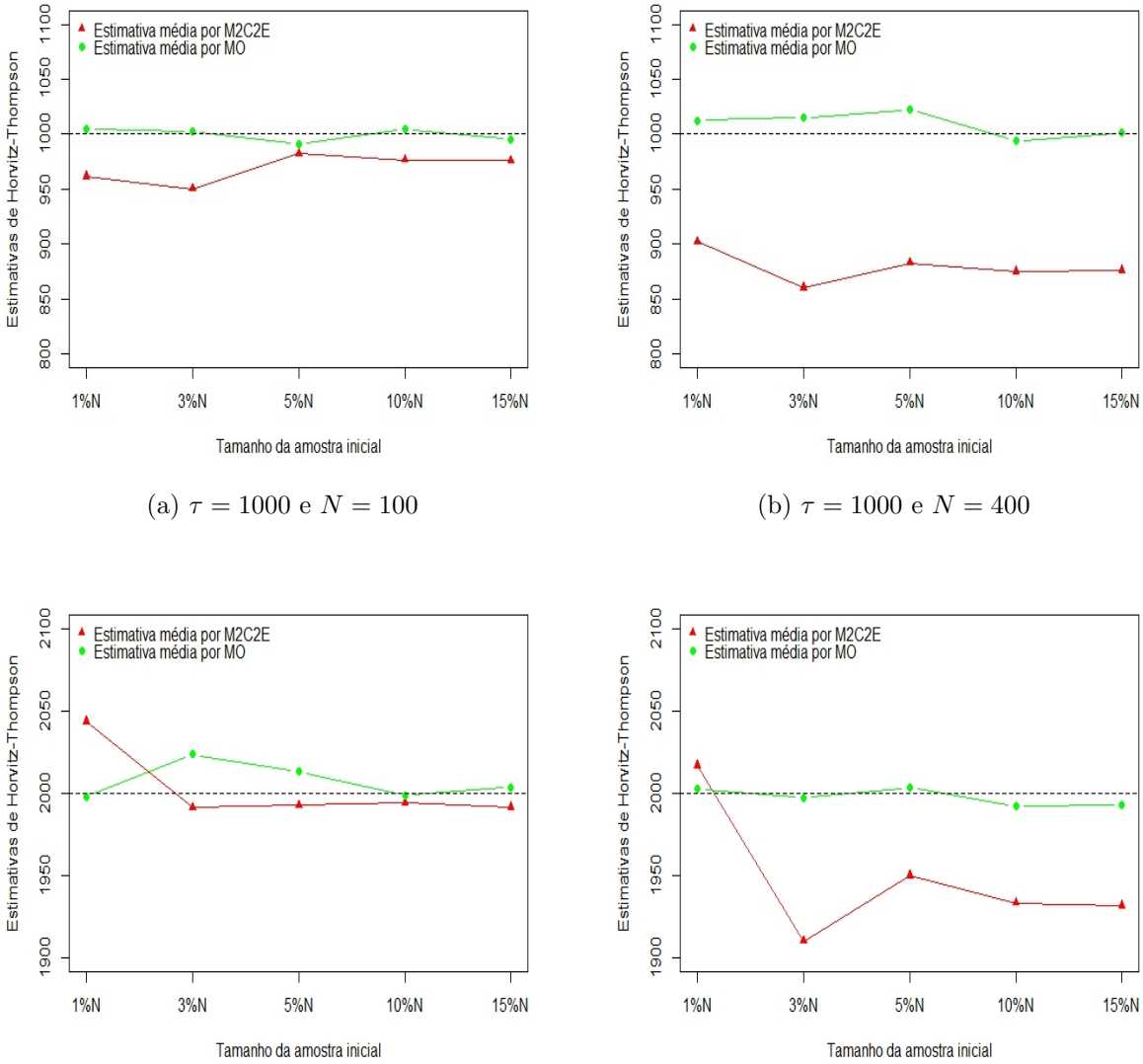


Figura 6.9: Boxplots com as eficiências entre o M2C2E e o MO das populações sintéticas $\tau = 1000$ e $N = 100$, $\tau = 1000$ e $N = 400, \tau = 2000$ e $N = 100$, $\tau = 2000$ e $N = 400$.

A Figura 6.10 contém as estimativas médias do MO (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 tamanho inicial de z_1 referentes as populações sintéticas apresentadas na Figura 6.7. Nos casos em que $\tau = 1000$ e $N = 100$ na Figura 6.10(a) e $\tau = 2000$ e $N = 100$ na Figura 6.10(c), as estimativas médias estão visualmente mais próximas da linha pontilhada a qual representa o total populacional verdadeiro.



(a) $\tau = 1000$ e $N = 100$

(b) $\tau = 1000$ e $N = 400$

(c) $\tau = 2000$ e $N = 100$

(d) $\tau = 2000$ e $N = 400$

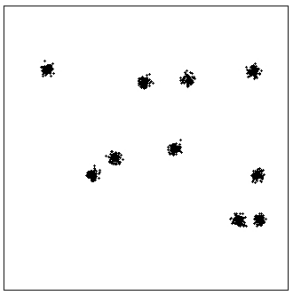
Figura 6.10: Estimativas médias referente ao número total das populações sintéticas $\tau = 1000$ e $N = 100$, $\tau = 1000$ e $N = 400$, $\tau = 2000$ e $N = 100$, $\tau = 2000$ e $N = 400$ usando o Método Ótimo - MO e o Método 2-Camadas e 2-Estimadores - M2C2E.

População	z_1	M2C2E				MO			ef_{MO}^{M2C2E}
		$E(\widehat{\tau}_{HT_mod})$	ER	$Var(\widehat{\tau}_{HT_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HT})$	ER	$Var(\widehat{\tau}_{HT})$	
$\tau = 1000$ e $N = 100$	1%N	961,80	3,82%	4917504,24	[917,76 ; 1005,04]	1005,14	0,51%	3631218,76	1,35
	3%N	950,21	4,98%	1520255,24	[926,17 ; 974,25]	1002,66	0,26%	1157849,89	1,31
	5%N	982,27	1,77%	930557,50	[963,46 ; 1001,08]	991,22	0,88%	662075,63	1,40
	10%N	976,67	2,33%	408895,04	[964,20 ; 989,14]	1004,25	0,43%	276826,99	1,48
	15%N	976,13	2,39%	244810,99	[966,48 ; 985,78]	995,46	0,45%	158738,76	1,54
$\tau = 1000$ e $N = 400$	1%N	902,21	9,78%	5492473,51	[856,51 ; 947,91]	1012,52	1,25%	5490762,94	1,00
	3%N	860,59	13,94%	1741530,52	[834,86 ; 886,32]	1014,95	1,49%	1850634,05	0,94
	5%N	882,81	11,72%	1028563,66	[863,03 ; 902,59]	1022,61	2,26%	1077371,48	0,95
	10%N	875,01	12,50%	459279,29	[861,79 ; 888,23]	993,77	0,62%	469889,13	0,98
	15%N	876,12	12,39%	285934,89	[865,69 ; 886,55]	1001,42	0,14%	292959,03	0,97
$\tau = 2000$ e $N = 100$	1%N	2043,66	2,18%	11861599,27	[1975,84 ; 2110,16]	1997,93	0,10%	9266988,73	1,28
	3%N	1991,21	0,44%	3429983,80	[1955,09 ; 2027,32]	2023,54	1,16%	2433200,05	1,41
	5%N	1992,72	0,36%	1732835,22	[1967,05 ; 2018,39]	2013,23	0,66%	1108046,48	1,56
	10%N	1994,32	0,28%	638359,02	[1978,74 ; 2009,90]	1998,82	0,06%	316497,84	2,01
	15%N	1991,60	0,42%	346537,08	[1980,12 ; 2003,08]	2003,67	0,18%	152744,83	2,27
$\tau = 2000$ e $N = 400$	1%N	2016,97	0,85%	14777143,57	[1942,01 ; 2091,93]	2002,65	0,13%	10071325,16	1,48
	3%N	1910,19	4,49%	4495092,98	[1868,84 ; 1951,34]	1997,46	0,12%	3199013,88	1,41
	5%N	1950,01	2,50%	2731743,12	[1917,78 ; 1982,24]	2003,27	0,16%	1895579,28	1,44
	10%N	1933,49	3,33%	1225881,48	[1911,89 ; 1955,08]	1992,03	0,40%	824359,85	1,49
	15%N	1931,94	3,40%	731497,29	[1915,26 ; 1948,62]	1993,23	0,34%	482427,97	1,52

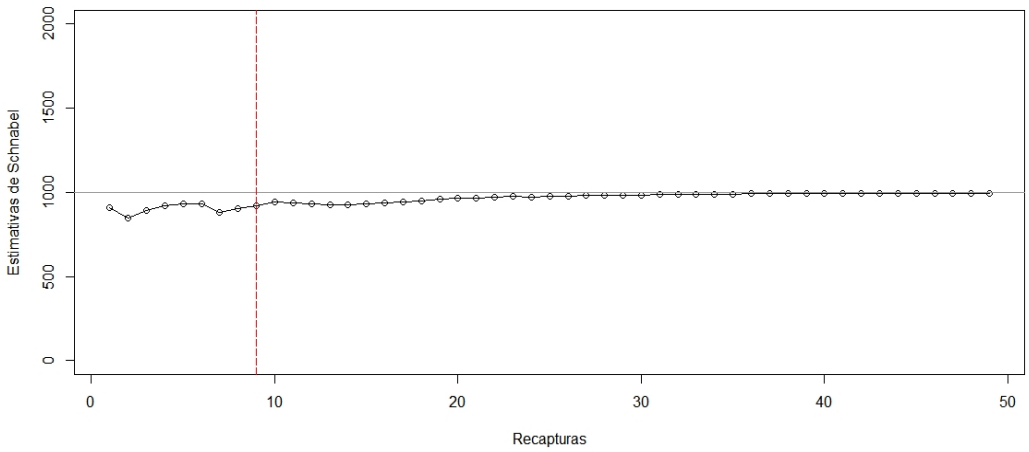
Tabela 6.1: Resultado da implementação considerando o valor fixo $\phi = 0,1$ com $n_1 = n_2 = ... = n_j = 10$ e variando os valores de τ e N.

O teste de hipótese T^2 Hotelling é um teste paramétrico multivariado para testar a igualdade entre os vetores de estimativas médias $\vec{\mu}$. Conforme Johnson [28], o único pré-requisito para realizar esse teste é a normalidade dos dados o qual pode ser verificar pelo Teorema do Limite Central, pois as médias ou os totais de amostras grandes e aleatórias são aproximadamente normais ou pelo teste de normalidade Kolmogorov-Smirnov os quais são válidos a este cenário, mas ainda que o pré-requisito não fosse válido, Arnold [4] apresentou a não normalidade dentro do teste T^2 Hotelling.

Após realizar um teste de hipótese T^2 Hotelling para verificar a igualdade entre os vetores de estimativas médias de Horvitz-Thompson com as seguintes hipóteses: H0: $\vec{\mu}_{M2C2E_{HT_mod}} = \vec{\mu}_{MO_{HT}}$ versus H1: $\vec{\mu}_{M2C2E_{HT_mod}} \neq \vec{\mu}_{MO_{HT}}$, a única população 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 do MO e do M2C2E para essa população não podem ser consideradas diferentes.

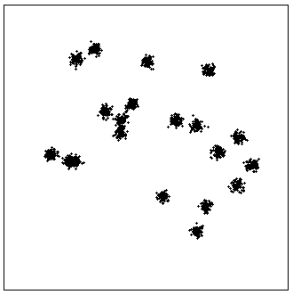


(a) $N = 1000$

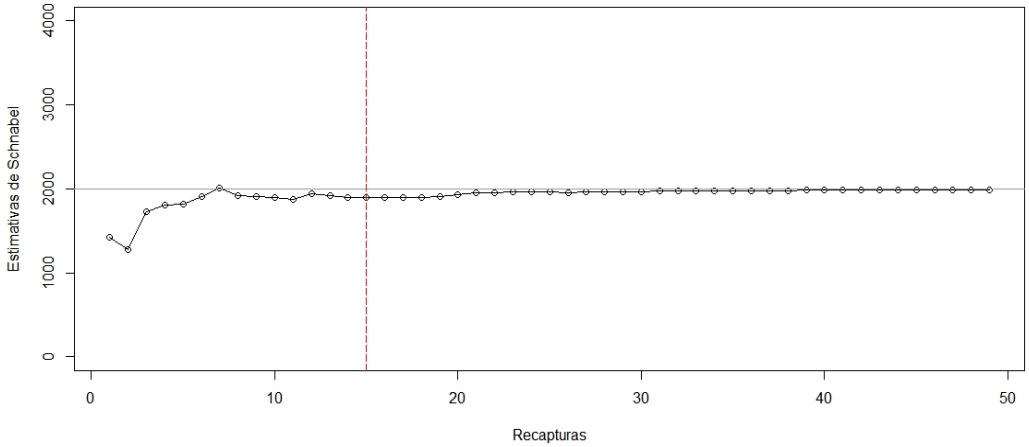


(b) Implementação referente a população com 1000 elementos

Figura 6.11: População sintética com 1000 elementos e estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 100$.



(a) $N = 2000$



(b) Implementação referente a população com 2000 elementos

Figura 6.12: População sintética com 2000 elementos e estimativas de Schnabel a cada recaptura para o número total de elementos com $n_1 = n_2 = ... = 100$.

As Figuras 6.11 e 6.12 têm o propósito de analisar o MCRM usando o critério de parada com erros de 2% e 5%, respectivamente. Na Figura 6.11, a parada ocorreu na 9ª recaptura, conforme representado pela linha pontilhada vermelha na vertical, retornando a estimativa de Schnabel $\widehat{n}^*_{schn} = 921$ elementos sendo que $N = 1000$, variância amostral na 9ª recaptura referente as estimativas de Schnabel anteriores igual a 888,17 e $ER = 7,90\%$ que é maior do que os erros relativos do MO ($\tau = 1000$ e $N = 100$ e com $\tau = 1000$ e $N = 400$) e do M2C2E ($\tau = 1000$ e $N = 100$). Na Figura 6.12, tem-se que a 15ª recaptura seria a última com estimativa de Schnabel $\widehat{n}^*_{schn} = 1901$ elementos, onde $N = 2000$, variância amostral na 15ª recaptura referente as estimativas de Schnabel anteriores igual a 27495,08 e o erro relativo $ER = 4,95\%$ indica que o MCRM implementado separadamente contém um erro relativo maior que todos os cenários apresentados na Tabela 6.1 para $\tau = 2000$ no MO e no M2C2E, indicando que o MCRM apresentou a estimativa mais tendenciosa no contexto na Figura 6.12 que na Figura 6.11.

6.4 Convergência do Estimador de Schnabel

Com a finalidade de observar a convergência das estimativas de Schnabel a cada recaptura, 50 replicações para 50 recapturas foram realizadas, tem-se uma matriz de captura representada por $C_{50,100}$ e 49 matrizes de recaptura $R_{50,100}$, sendo 100 o número de elementos selecionados em 50 replicações, cujos conteúdos são os IDs dos escolhidos. Através dessas matrizes, é possível produzir os vetores com as estimativas de Schnabel $\vec{n}_{50,1}$, conforme apresentado a seguir:

$$\overset{C_{50,100}}{\begin{bmatrix} 1017 & 1860 & \cdots & 1569 \\ 557 & 1685 & \cdots & 1218 \\ \vdots & \vdots & \ddots & \vdots \\ 1531 & 510 & \cdots & 741 \end{bmatrix}} \overset{R^1_{50,100}}{\begin{bmatrix} 1498 & 191 & \cdots & 961 \\ 1490 & 13510 & \cdots & 64 \\ \vdots & \vdots & \ddots & \vdots \\ 1267 & 1645 & \cdots & 771 \end{bmatrix}} \Longrightarrow \overset{\vec{n}^1_{50,1}}{\begin{bmatrix} 1428,57 \\ 2000,00 \\ \vdots \\ 2000,00 \end{bmatrix}}$$

$$\vdots$$

$$\overset{C_{50,100}}{\begin{bmatrix} 1017 & 1860 & \cdots & 1569 \\ 557 & 1685 & \cdots & 1218 \\ \vdots & \vdots & \ddots & \vdots \\ 1531 & 510 & \cdots & 741 \end{bmatrix}} \overset{R^1_{50,100}}{\begin{bmatrix} 1498 & 191 & \cdots & 961 \\ 1490 & 1351 & \cdots & 64 \\ \vdots & \vdots & \ddots & \vdots \\ 1267 & 1645 & \cdots & 771 \end{bmatrix}} \cdots \overset{R^{49}_{50,100}}{\begin{bmatrix} 93 & 8 & \cdots & 49 \\ 86 & 7 & \cdots & 11 \\ \vdots & \vdots & \ddots & \vdots \\ 40 & 20 & \cdots & 5 \end{bmatrix}} \Longrightarrow \overset{\vec{n}^{49}_{50,1}}{\begin{bmatrix} 1988,08 \\ 2011,01 \\ \vdots \\ 2015,66 \end{bmatrix}}$$

Nas Figuras 6.13 e 6.14, os boxplots são resultados dos vetores com as estimativas de Schnabel $\vec{n}_{50,1} = \{\vec{n}^1_{50,1}, \vec{n}^2_{50,1}, ..., \vec{n}^{49}_{50,1}\}$ e apresentam a convergência das estimativas referente as populações nas Figuras 6.11(a) e 6.12(a). Note que a medida que o número de recapturas aumenta a distribuição das estimativas de Schnabel converge para o valor de parâmetro de $\tau = 1000$ e $\tau = 2000$, respectivamente.

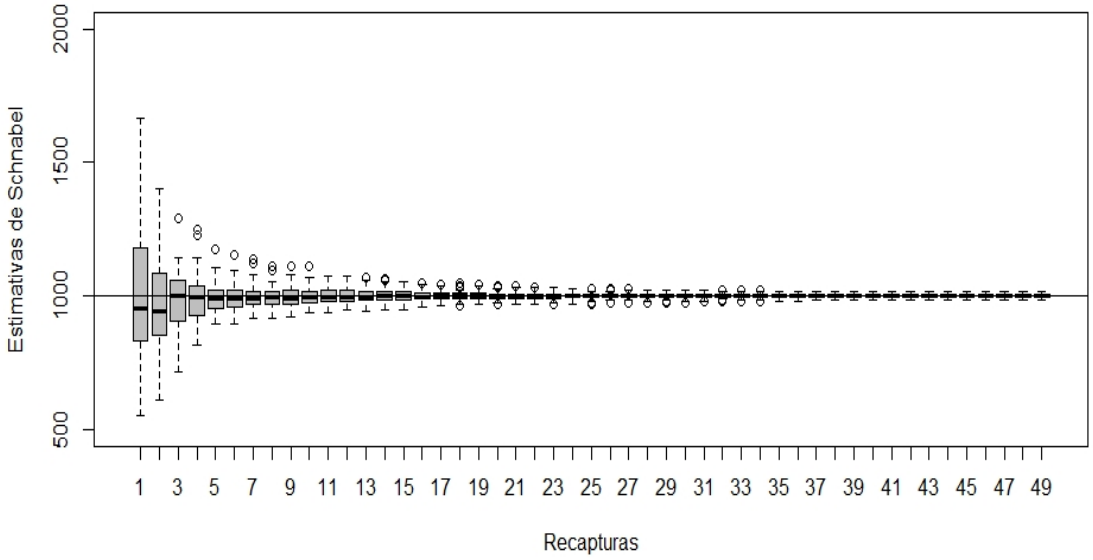


Figura 6.13: Boxplots com 50 replicações para cada número de recaptura dos dados sintéticos de 1000 elementos com $n_1 = n_2 = ... = 100$.

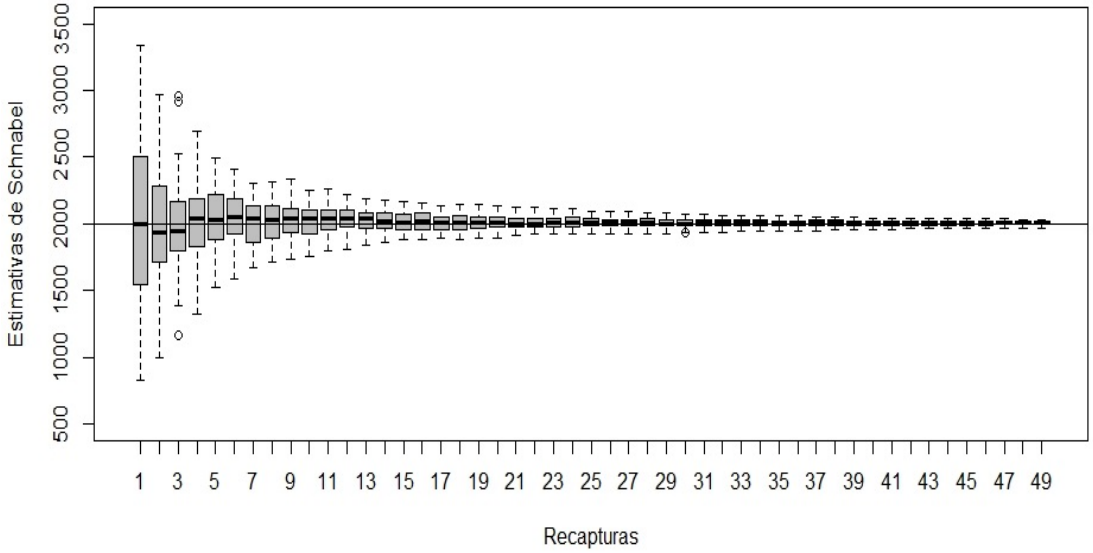


Figura 6.14: Boxplots com 50 replicações para cada número de recaptura dos dados sintéticos de 2000 elementos com $n_1 = n_2 = ... = 100$.

Capítulo 7

Aplicação a Dados Reais

Neste capítulo, o objetivo é analisar os métodos MO, M2C2E e o MCRM em um conjunto de dados reais. Em particular, nesse estudo com dados reais, utilizou-se os dados de um aplicativo de celular para táxis que fornecia, dentre outras informações, a localização geográfica dos táxis conveniados em circulação no município do Rio de Janeiro no dia 22 de junho de 2016.

7.1 Descrição do Conjunto de Dados

No município do Rio de Janeiro, existem concentrações de táxis em região com maior renda e em locais como aeroportos, shoppings e rodoviárias, em outras palavras, a presença de táxis varia em relação a localização do município. A concorrência com carros particulares que fazem serviço de transporte por aplicativo é um exemplo de fator que causa a redução do número de táxi em operação. Portanto, esses dados são exemplos de população rara e agrupada conectada por uma rede móvel a um aplicativo. 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.

O banco de dados a ser utilizado é constituído pelas coordenadas geográficas (latitude e longitude), data, hora e operadora de telefonia móvel do motorista desses táxis ao longo de um dia. Nesse caso, o objetivo será estimar o número total de táxis por hora no aplicativo, essa unidade de tempo foi escolhida por ser a menor unidade de tempo apresentada pelos 24 conjuntos de dados sendo um para cada hora, uma vez que a escolha de unidades maiores como por período do dia ou por dia, poderia haver duplicidade na contagem, já que o mesmo táxi pode sair do aplicativo e entrar novamente no mesmo dia, por exemplo. Com a finalidade de atingir o objetivo, as coordenadas serão sobrepostas a

região de interesse que é município do Rio de Janeiro.

A Figura 7.1 apresenta o contorno do município do Rio de Janeiro com área de 1200[1] km^2 em vermelho sobreposto por uma grade de 16x16 (256 células) que representa a região de estudo para os dados reais, exceto pelas ilhas.

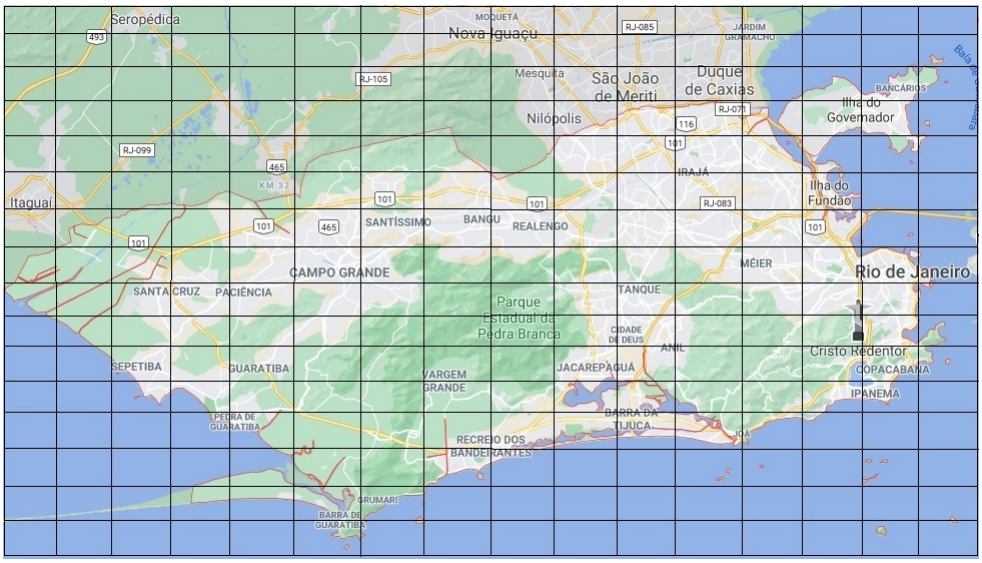


Figura 7.1: Contorno do município do Rio de Janeiro sobreposto por uma grade com 256 células.

7.2 Resultados

Inicialmente, o número total de dados no dia 22 de junho de 2016 era 578297 observações. Após a limpeza do banco de dados, devido às duplicações restaram 110369 coordenadas relevantes. Observou-se que esses dados não estão divididos igualmente ao longo do dia, ou seja, existem horários em que a circulação dos táxis conveniados ao aplicativo é maior, conforme a Figura 7.2 na qual o total de táxis por hora é apresentado. Respeitando os pressupostos dos estimadores a serem utilizados, a rede de táxis foi considerada como uma população fechada no horário de estudo (01:00hs, 02:00hs, ... , 24:00hs). O *software* RStudio Cloud foi utilizado para programar os algoritmos, consolidar os resultados e confeccionar os gráficos.

[1]https://www.ibge.gov.br/cidades-e-estados/rj/rio-de-janeiro.html

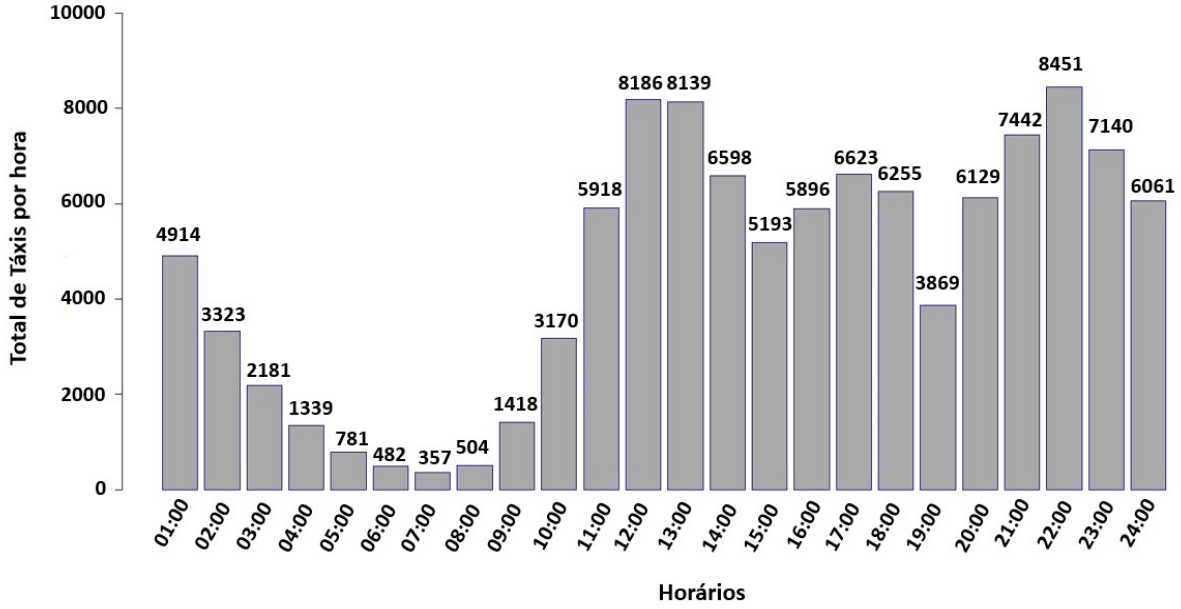


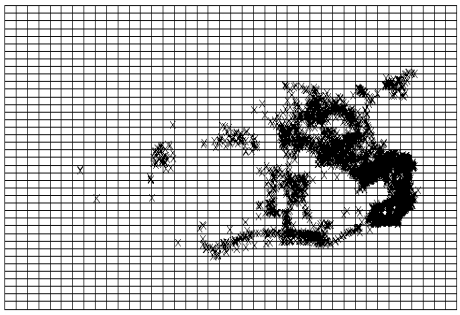
Figura 7.2: Gráfico de barras do número total de táxis no dia 22 de junho de 2016 no município do Rio de Janeiro por hora.

A fim de verificar se o M2C2E seria de fato uma alternativa ao MO, repetiu-se o estudo realizado no Capítulo 6 utilizando os dados reais de táxis, variando $z_1 \in \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$, realizou-se 100 replicações de amostras adaptativas para cada tamanho inicial de z_1 e para cada horário. Vale destacar que em situações reais, as desvantagens do MCRM grade 1x1 (1 célula) em populações raras e agrupadas estão na possibilidade de não encontrar nenhum elemento da rede em áreas consideradas grandes e não encontrar recapturados ao longo do estudo.

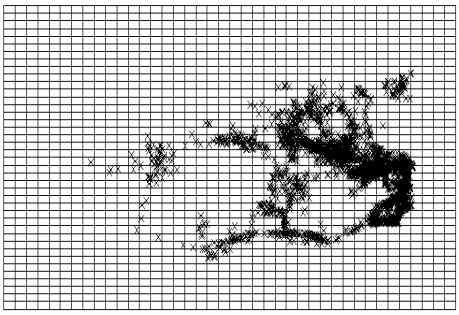
As Figuras 7.3, 7.4, 7.5 e 7.6 correspondem as distribuições espaciais por hora de táxis no dia 22 de junho de 2016 ao longo de um dia no município do Rio de Janeiro sobreposta por uma grade 40x40, ou seja, 1600 células. Os "x"representam a posição (latitude e longitude) dos táxis no respectivo horário. Note que a concentração de táxi varia no decorrer do dia, por exemplo, a Figura 7.4 (a) é o horário de 07:00hs no qual a concentração de táxis é menor com 357 elementos conectados ao aplicativo, e, por outro lado, a Figura 7.6 (d) referente ao horário de 22:00hs é o de maior concentração com 8451 elementos conectados ao aplicativo. Com o objetivo de facilitar a análise visual, as grades foram divididas em 4 blocos que representam o período do dia (madrugada, manhã, tarde e noite) e as Figuras 7.3, 7.4, 7.5 e 7.6 referem-se ao período madrugada, manhã, tarde e noite, respectivamente.

Nas Subseções 7.2.1 e 7.2.2, foram criadas duas configurações: uma utilizando, o estimador de Schnabel (camada 1) e o estimador de Horvitz-Thompson modificado (camada 2); a outra, com o estimador de Schnabel (camada 1) e o estimador de Hansen-Hurwitz modificado (camada 2). Uma vez que o MCRM por construção é uma grade 1x1, não existe a variação de z_1 que é o número de células selecionadas iniciais.

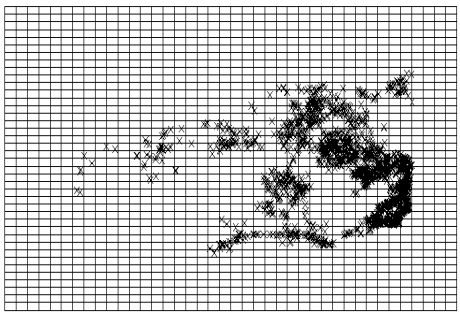
A Tabela 7.6 contém os resultados da estimação dos táxis usando MCRM com $n_1 = n_2 = ... = n_j = 100$, os horários, o valor verdadeiro de táxis no horário representado pela variável N, a média das estimativas de Schnabel $E(\widehat{N}_{schn})$ a partir de 50 replicações para cada número de recaptura, o erro relativo referente a estimativa média em termos percentuais ER, a variância $Var(\widehat{N}_{schn})$ e o intervalo de confiança de 95% para a estimativa média $E(\widehat{N}_{schn})$ usando Expressão 6.2 e as eficiências ef_{MCRM}^{M2C2E} que foram calculadas em relação aos valores de variância do M2C2E usando o estimador de Horvitz-Thompson modificado na camada 2.



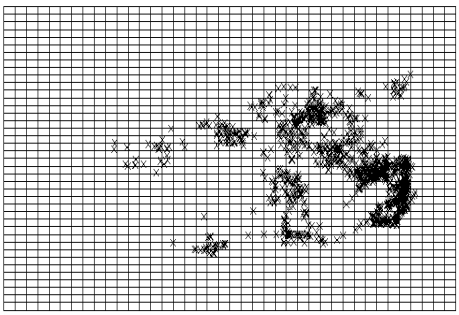
(a) 01:00hs



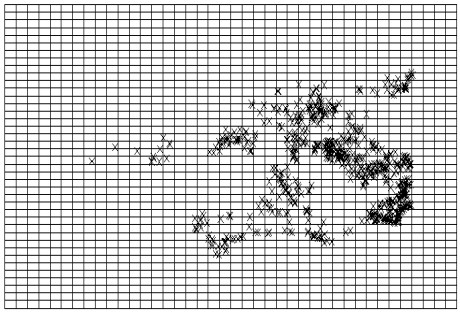
(b) 02:00hs



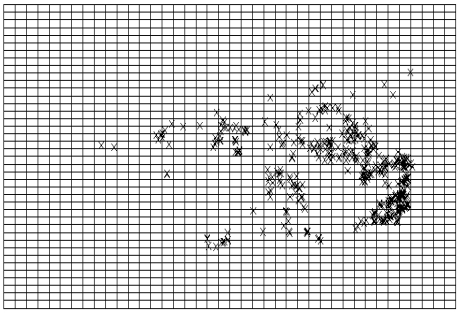
(c) 03:00hs



(d) 04:00hs

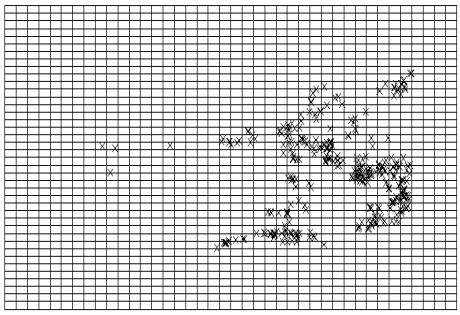


(e) 05:00hs

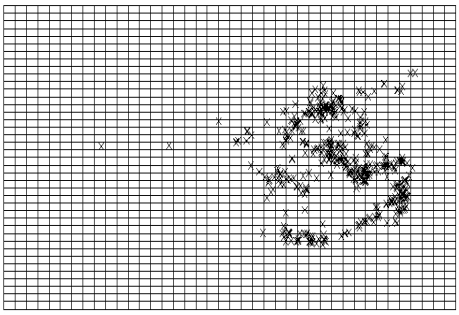


(f) 06:00hs

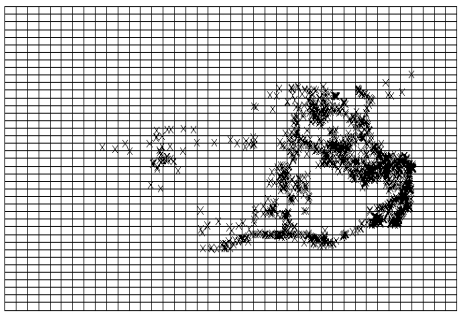
Figura 7.3: Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da madrugada entre 01:00hs - 06:00hs no município do Rio de Janeiro sobreposta por grade 40x40.



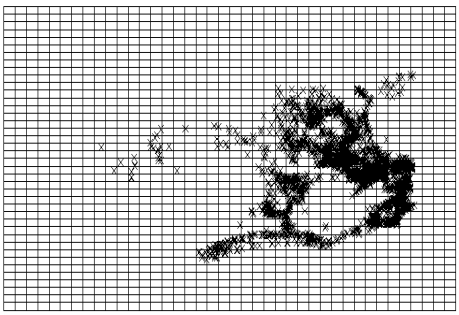
(a) 07:00hs



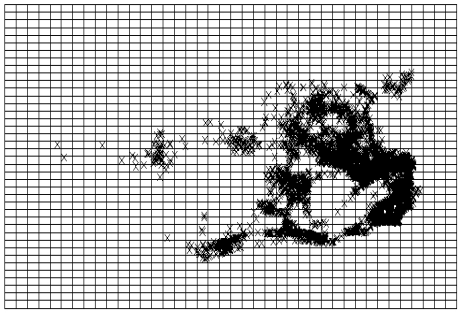
(b) 08:00hs



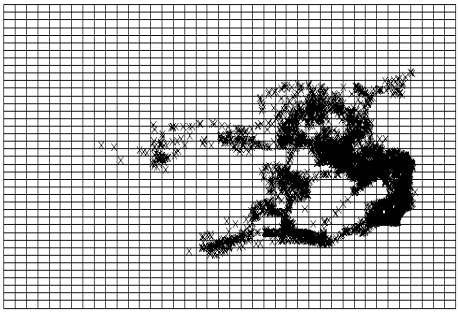
(c) 09:00hs



(d) 10:00hs

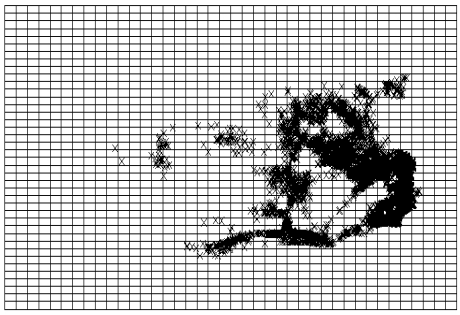


(e) 11:00hs

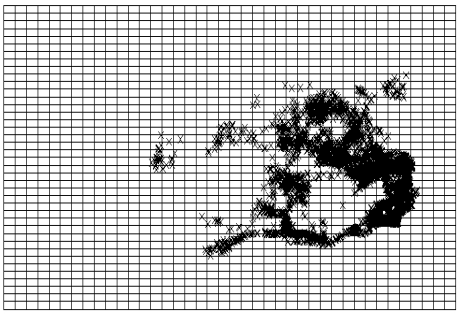


(f) 12:00hs

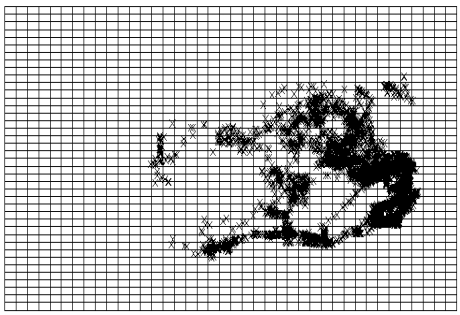
Figura 7.4: Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da manhã entre 07:00hs - 12:00hs no município do Rio de Janeiro sobreposta por grade 40x40.



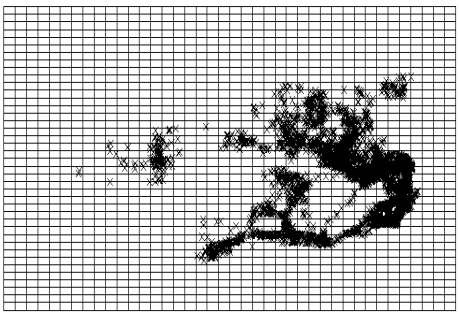
(a) 13:00hs



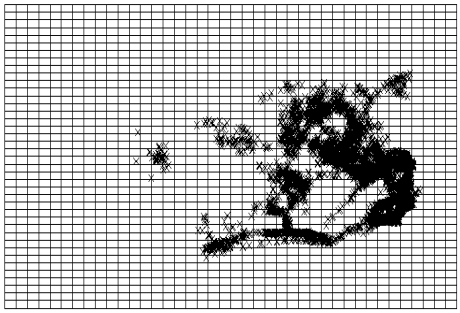
(b) 14:00hs



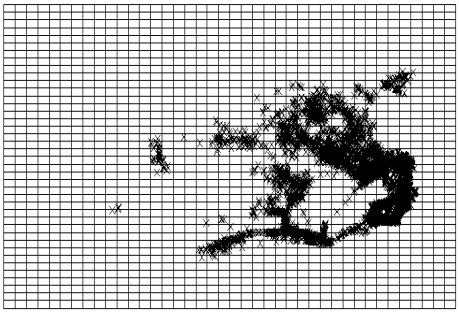
(c) 15:00hs



(d) 16:00hs

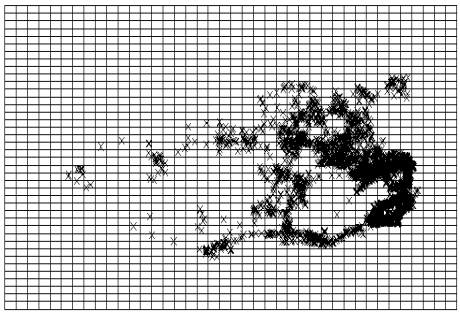


(e) 17:00hs

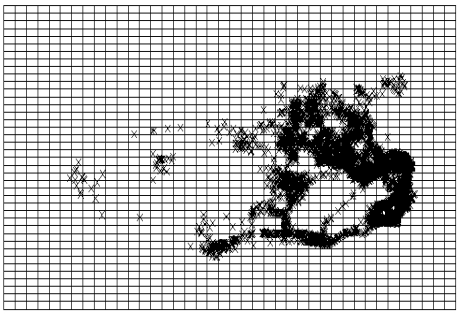


(f) 18:00hs

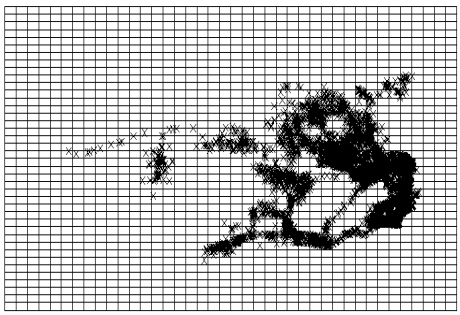
Figura 7.5: Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da tarde entre 13:00hs - 18:00hs no município do Rio de Janeiro sobreposta por grade 40x40.



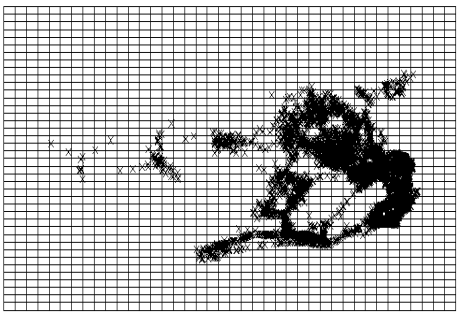
(a) 19:00hs



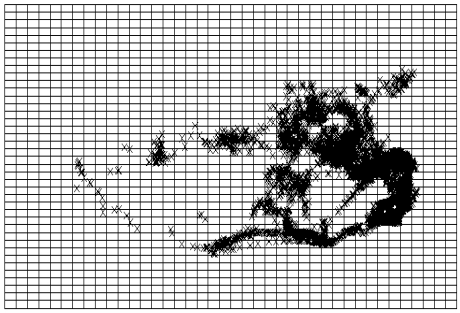
(b) 20:00hs



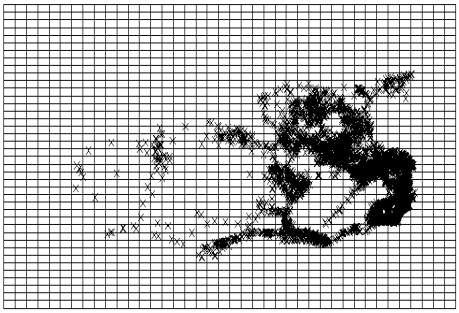
(c) 21:00hs



(d) 22:00hs



(e) 23:00hs



(f) 24:00hs

Figura 7.6: Distribuição espacial por hora da população de táxis no dia 22 de junho de 2016 no período da noite entre 19:00hs - 24:00hs no município do Rio de Janeiro sobreposta por grade 40x40.

7.2.1 Primeira Configuração

A primeira configuração visa implementar o estimador de Schnabel na camada 1 e o estimador de Horvitz- Thompson modificado na camada 2, variando $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$ e o critério de parada na camada 1, conforme a Inequação 4.8 que foi utilizada para δ com erro de 5% em relação a primeira estimativa. As especificações de cada método empregado são as seguintes: MCRM com o estimador de Schnabel $n_1 = n_2 = ... = n_j = 100$; MO com o estimador de Horvitz-Thompson; e M2C2E - camada 1 com estimador de Schnabel $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos e camada 2 com o estimador de Horvitz-Thompson modificado.

As Tabelas 7.1, 7.2, 7.3, 7.4 e 7.5 contêm os horários, o total populacional τ, o tamanho da amostra inicial z_1, para o caso do M2C2E: a estimativa média $E(\widehat{\tau}_{HT_mod})$, o erro relativo referente a estimativa média em termos percentuais ER, a variância $Var(\widehat{\tau}_{HT_mod})$ e o intervalo de confiança de 95% para a estimativa média $E(\widehat{\tau}_{HT_mod})$ usando a Expressão 6.2; e para o MO: a estimativa média $E(\widehat{\tau}_{HT})$, o erro relativo referente a estimativa média em termos percentuais ER, a variância $Var(\widehat{\tau}_{HT})$. Por fim, a eficiência ef_{MO}^{M2C2E} que é a razão entre as variâncias na Equação 6.3.

7.2.2 Segunda Configuração

A segunda configuração em relação a primeira muda o estimador utilizado na camada 2. Portanto, tem-se o estimador de Schnabel na camada 1 e o estimador de Hansen-Hurwitz modificado variando $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$ na camada 2. Para cada um dos três métodos, tem-se as seguintes configurações: MCRM - estimador de Schnabel $n_1 = n_2 = ... = n_j = 100$; MO - estimador de Hansen-Hurwitz; e M2C2E - Camada 1 - estimador de Schnabel $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos e Camada 2 - estimador de Hansen-Hurwitz modificado.

As Tabelas 7.7, 7.8, 7.9, 7.10 e 7.11 contêm os horários, o total populacional τ, o tamanho da amostra inicial z_1, para o M2C2E: a estimativa média $E(\widehat{\tau}_{HH_mod})$, o erro relativo ER referente a estimativa média em termos percentuais, a variância $Var(\widehat{\tau}_{HH_mod})$ e o intervalo de confiança de 95% para a estimativa média $E(\widehat{\tau}_{HH_mod})$; e para o MO, a estimativa média $E(\widehat{\tau}_{HH})$, o erro relativo ER referente a estimativa média em termos percentuais, a variância $Var(\widehat{\tau}_{HH})$. Usando $Var(\widehat{\tau}_{HH_mod})$ e $Var(\widehat{\tau}_{HH})$ ao invés de $Var(\widehat{\tau}_{HT_mod})$ e $Var(\widehat{\tau}_{HT})$ na Equação 6.3, tem-se por fim a eficiência ef_{MO}^{M2C2E}.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HT_mod})$	ER	$Var(\widehat{\tau}_{HT_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HT})$	ER	$Var(\widehat{\tau}_{HT})$	
01 : 00	4914	1%N	4724,41	3,86%	3879923,93	[4685,80 ; 4763,02]	4880,82	0,67%	3108376,67	1,24
		3%N	4912,43	0,03%	9187,59	[4910,55 ; 4914,31]	4911,01	0,06%	7791,12	1,18
		5%N	4900,42	0,27%	4861,93	[4899,05 ; 4901,79]	4905,12	0,18%	5300,33	0,92
		10%N	4891,55	0,45%	2274,38	[4890,62 ; 4892,48]	4919,63	0,11%	2271,45	1,00
		15%N	4901,56	0,25%	1043,99	[4900,93 ; 4902,19]	4909,56	0,09%	870,08	1,20
02 : 00	3323	1%N	3708,44	10,39%	746863,74	[3691,50 ; 3725,38]	3374,08	1,54%	1953982,42	0,38
		3%N	3310,21	0,38%	234581,76	[3300,71 ; 3319,70]	3350,67	0,83%	32169,04	7,29
		5%N	3338,61	0,47%	17117,25	[3336,04 ; 3341,17]	3350,01	0,81%	19653,68	0,87
		10%N	3318,41	0,14%	5832,38	[3303,44 ; 3333,37]	3325,08	0,06%	6014,93	0,97
		15%N	3323,40	0,01%	2231,46	[3322,47 ;3324,33]	3323,19	0,00%	2471,77	0,90
03 : 00	2181	1%N	2218,06	1,69%	817002,82	[2200,34 ; 2235,77]	2157,93	1,06%	862331,28	0,95
		3%N	2204,07	1,06%	16671,84	[2178,76 ; 2229,38]	2189,56	0,39%	11988,91	1,39
		5%N	2197,07	0,74%	7191,71	[2195,41 ; 2198,73]	2174,99	0,27%	6447,03	1,11
		10%N	2189,26	0,38%	3036,77	[2178,46 ; 2200,06]	2183,64	0,12%	2633,86	1,15
		15%N	2187,39	0,29%	1076,38	[2186,74 ; 2188,03]	2173,94	0,32%	1429,12	0,75
04 : 00	1339	1%N	1259,39	5,94%	736895,37	[1242,56 ; 1276,22]	1280,93	4,33%	792020,90	0,93
		3%N	1320,61	1,37%	85621,79	[1314,87 ; 1326,34]	1346,38	0,55%	81185,20	1,05
		5%N	1315,06	1,79%	20039,47	[1312,28 ; 1317,83]	1348,13	0,68%	16731,38	1,20
		10%N	1304,47	2,58%	5303,38	[1290,19 ; 1318,74]	1341,02	0,15%	4685,81	1,13
		15%N	1308,66	2,26%	1717,69	[1300,54 ; 1316,78]	1330,62	0,62%	2763,63	1,29
05 : 00	781	1%N	756,95	3,08%	187143,01	[748,47 ; 765,42]	757,72	2,98%	161260,84	1,16
		3%N	784,27	0,42%	23751,63	[781,24 ; 787,29]	766,22	1,89%	29445,44	0,80
		5%N	760,15	2,67%	9051,70	[758,28 ; 762,01]	786,97	0,76%	12152,80	0,75
		10%N	785,21	0,53%	3476,42	[784,05 ; 786,36]	779,49	0,19%	4467,67	0,78
		15%N	780,59	0,05%	1506,10	[779,83 ; 781,35]	786,41	0,69%	2297,08	0,65

Tabela 7.1: Configuração 1 - Resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HT_mod})$	ER	$Var(\widehat{\tau}_{HT_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HT})$	ER	$Var(\widehat{\tau}_{HT})$	
06 : 00	482	1%N	431,59	10,45%	127215,31	[424,60 ; 438,58]	519,28	7,73%	130825,12	0,98
		3%N	494,75	2,65%	35259,22	[491,07 ; 498,43]	499,28	3,58%	36226,36	0,98
		5%N	485,73	0,77%	16255,33	[483,23 ; 488,23]	490,63	1,79%	15056,40	1,08
		10%N	464,74	3,58%	5365,52	[463,30 ; 466,17]	482,03	0,00%	3953,10	1,36
		15%N	473,05	1,86%	2543,09	[472,06 ; 474,04]	485,29	0,68%	1934,81	1,31
07 : 00	357	1%N	342,45	4,07%	66031,08	[337,41 ; 347,48]	343,32	3,83%	67979,92	0,97
		3%N	322,71	9,60%	13521,07	[320,43 ; 324,99]	342,43	4,08%	19060,53	0,71
		5%N	331,50	7,14%	5242,19	[330,08 ; 332,91]	371,22	3,98%	7792,46	0,67
		10%N	320,61	10,19%	2821,01	[319,57 ; 321,65]	359,52	0,71%	3465,85	0,81
		15%N	321,71	9,88%	1395,51	[320,97 ; 322,44]	353,69	0,93%	1344,03	1,03
08 : 00	504	1%N	548,01	8,73%	119734,86	[541,23 ; 554,79]	491,68	2,44%	121424,94	0,98
		3%N	509,46	1,08%	21123,42	[506,61 ; 512,31]	506,36	0,47%	20798,65	1,01
		5%N	501,27	0,54%	8525,91	[499,46 ; 503,08]	515,49	2,28%	5834,26	1,46
		10%N	506,30	0,45%	2202,99	[505,38 ; 507,22]	501,24	0,55%	1717,39	1,28
		15%N	504,60	0,12%	932,88	[504,00 ; 505,20]	504,37	0,07%	1133,61	0,82
09 : 00	1418	1%N	1340,45	5,47%	645461,55	[1324,70 ; 1356,19]	1467,73	3,51%	551721,73	1,17
		3%N	1392,14	1,82%	53544,05	[1387,60 ; 1396,67]	1403,26	1,04%	51769,92	1,03
		5%N	1424,32	0,44%	12315,64	[1422,14 ; 1426,49]	1411,52	0,45%	9423,41	1,31
		10%N	1413,06	0,35%	3779,94	[1411,85 ; 1414,26]	1410,69	0,52%	2545,31	1,48
		15%N	1412,73	0,37%	985,16	[1412,11 ; 1413,34]	1420,70	0,19%	828,50	1,19
10 : 00	3170	1%N	3135,69	1,08%	1563265,05	[3111,18 ; 3160,19]	3225,00	1,73%	1295392,91	1,21
		3%N	3196,92	0,85%	1574,92	[3196,14 ; 3197,70]	3171,77	0,05%	1968,34	0,80
		5%N	3187,65	0,55%	1052,12	[3187,01 ; 3188,28]	3177,85	0,25%	1036,67	1,01
		10%N	3188,60	0,58%	509,32	[3188,16 ; 3189,04]	3172,77	0,09%	446,02	1,14
		15%N	3188,02	0,57%	189,15	[3187,75 ; 3188,29]	3169,88	0,00%	183,97	1,03

Tabela 7.2: Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HT_mod})$	ER	$Var(\widehat{\tau}_{HT_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HT})$	ER	$Var(\widehat{\tau}_{HT})$	
11 : 00	5918	1%N	5888,62	0,49%	3057122,01	[5854,35 ; 5922,89]	5959,46	0,70%	2710245,03	1,13
		3%N	5904,63	0,22%	4083,91	[5903,38 ; 5905,88]	5926,36	0,14%	4715,93	0,86
		5%N	5909,54	0,14%	2274,21	[5908,60 ; 5910,47]	5915,35	0,04%	2078,29	1,09
		10%N	5902,67	0,25%	850,93	[5902,09 ; 5903,24]	5918,55	0,00%	776,39	1,09
		15%N	5903,71	0,24%	416,17	[5903,31 ; 5904,11]	5918,01	0,00%	355,65	1,17
12 : 00	8186	1%N	8425,05	2,92%	3769645,78	[8386,99 ; 8463,10]	8217,63	0,38%	4922731,43	0,77
		3%N	8186,39	0,00%	6321,69	[8184,83 ; 8187,95]	8197,77	0,14%	8116,19	0,78
		5%N	8199,50	0,16%	3681,92	[8198,31 ; 8200,69]	8187,01	0,01%	3601,59	1,02
		10%N	8194,38	0,10%	1247,72	[8193,69 ; 8195,07]	8184,46	0,02%	1014,63	1,23
		15%N	8194,33	0,10%	557,92	[8193,86 ; 8194,79]	8188,41	0,03%	443,71	1,25
13 : 00	8139	1%N	8335,78	2,42%	5224450,45	[8290,98 ; 8380,58]	8240,56	1,24%	5976606,86	0,87
		3%N	8158,32	0,24%	4068,57	[8157,07 ; 8159,57]	8140,75	0,02%	3355,16	1,21
		5%N	8145,95	0,08%	1956,18	[8145,08 ; 8146,81]	8135,56	0,04%	1863,69	1,05
		10%N	8143,17	0,05%	708,27	[8142,65 ; 8143,69]	8138,96	0,00%	751,45	0,94
		15%N	8149,02	0,12%	281,43	[8148,69 ; 8149,35]	8139,78	0,00%	408,04	0,69
14 : 00	6598	1%N	6614,77	0,25%	4385167,05	[6573,72 ; 6655,81]	6317,16	4,26%	5990397,69	0,73
		3%N	6605,23	0,11%	3746,56	[6604,03 ; 6606,43]	6602,96	0,08%	3170,93	1,18
		5%N	6597,70	0,00%	1814,70	[6596,86 ; 6598,53]	6596,03	0,03%	1648,43	1,10
		10%N	6592,70	0,08%	747,90	[6592,16 ; 6593,23]	6597,37	0,00%	702,62	1,06
		15%N	6596,89	0,02%	376,84	[6596,51 ; 6597,27]	6601,14	0,05%	381,98	0,99
15 : 00	5193	1%N	4885,54	5,92%	4161518,04	[4845,56 ; 4925,52]	4971,05	4,27%	3964670,91	1,05
		3%N	5202,15	0,17%	10802,14	[5200,11 ; 5204,19]	5179,65	0,26%	7625,92	1,42
		5%N	5194,22	0,02%	5850,72	[5192,72 ; 5195,72]	5191,92	0,02%	6201,13	0,94
		10%N	5196,56	0,07%	2363,30	[5195,61 ; 5197,51]	5192,58	0,00%	2549,36	0,93
		15%N	5201,55	0,16%	1285,34	[5200,85 ; 5202,25]	5195,02	0,04%	1419,26	0,91

Tabela 7.3: Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HT_mod})$	ER	$Var(\widehat{\tau}_{HT_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HT})$	ER	$Var(\widehat{\tau}_{HT})$	
16 : 00	5896	1%N	5884,42	0,19%	3687570,33	[5846,78 ; 5922,06]	5998,00	1,73%	3149106,05	1,17
		3%N	5867,31	0,48%	18613,08	[5864,63 ; 5869,98]	5876,96	0,32%	15972,22	1,16
		5%N	5894,04	0,03%	12182,68	[5891,88 ; 5896,20]	5894,89	0,02%	9265,22	1,31
		10%N	5873,97	0,37%	2616,54	[5872,97 ; 5874,97]	5900,08	0,07%	3720,29	0,70
		15%N	5885,68	0,18%	1291,93	[5884,97 ; 5886,38]	5899,21	0,05%	1199,98	1,07
17 : 00	6623	1%N	6669,80	0,71%	3406488,19	[6633,62 ; 6705,98]	6722,67	1,50%	2935498,41	1,16
		3%N	6634,60	0,18%	5626,51	[6633,13 ; 6636,07]	6637,38	0,22%	5820,17	0,97
		5%N	6627,81	0,07%	2639,24	[6626,80 ; 6628,82]	6625,35	0,03%	3134,99	0,84
		10%N	6627,97	0,07%	1060,51	[6627,49 ; 6628,45]	6625,83	0,04%	1020,62	1,04
		15%N	6635,08	0,18%	607,45	[6634,59 ; 6635,56]	6620,79	0,03%	673,28	0,90
18 : 00	6255	1%N	6478,13	3,57%	3068939,43	[6443,79 ; 6512,47]	6204,33	0,81%	5029368,43	0,61
		3%N	6230,35	0,39%	23607,37	[6227,34 ; 6233,36]	6252,04	0,05%	26724,70	0,88
		5%N	6210,55	0,71%	8721,96	[6208,72 ; 6212,38]	6262,28	0,12%	15535,92	0,56
		10%N	6216,95	0,61%	4280,55	[6215,67 ; 6218,23]	6266,85	0,19%	5231,94	0,82
		15%N	6222,45	0,52%	1607,44	[6221,66 ; 6223,23]	6266,35	0,18%	2047,20	0,78
19 : 00	3869	1%N	3791,88	1,99%	2214457,14	[3762,71 ; 3821,04]	3986,68	3,04%	1363113,07	1,62
		3%N	3885,44	0,42%	15215,35	[3883,02 ; 3887,85]	3855,49	0,34%	11057,96	1,38
		5%N	3870,41	0,04%	6443,88	[3868,84 ; 3871,98]	3872,89	0,18%	6938,30	0,93
		10%N	3870,04	0,02%	3098,33	[3868,95 ; 3871,13]	3873,35	0,11%	2846,50	1,08
		15%N	3879,97	0,28%	1523,97	[3879,20 ; 3880,74]	3870,64	0,04%	1716,72	0,89
20 : 00	6129	1%N	6277,69	2,42%	3012670,62	[6243,67 ; 6311,71]	6274,21	2,37%	2935773,43	1,03
		3%N	6117,39	0,19%	7201,52	[6115,73 ; 6119,05]	6138,45	0,15%	5374,22	1,34
		5%N	6115,35	0,22%	3948,09	[6114,12 ; 6116,58]	6121,71	0,12%	3408,87	1,16
		10%N	6105,58	0,38%	1239,16	[6104,89 ; 6106,27]	6122,52	0,11%	1042,23	1,19
		15%N	6113,40	0,25%	759,52	[6112,86 ; 6113,94]	6127,16	0,03%	726,35	1,05

Tabela 7.4: Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HT_mod})$	ER	$Var(\widehat{\tau}_{HT_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HT})$	ER	$Var(\widehat{\tau}_{HT})$	
21 : 00	7442	1%N	7170,89	3,64%	6887385,35	[7119,45 ; 7222,33]	7199,57	3,26%	6935555,37	0,99
		3%N	7423,93	0,24%	6391,71	[7422,36 ; 7425,49]	7451,09	0,12%	7894,44	0,81
		5%N	7443,04	0,01%	3614,11	[7441,86 ; 7444,22]	7437,09	0,06%	3554,17	1,01
		10%N	7432,79	0,12%	1506,08	[7432,03 ; 7433,55]	7440,14	0,02%	1438,17	1,05
		15%N	7434,75	0,09%	555,38	[7434,29 ; 7435,21]	7442,52	0,00%	621,76	0,89
22 : 00	8451	1%N	8435,79	0,18%	7064425,21	[8383,69 ; 8487,88]	8153,17	3,52%	9011605,89	0,78
		3%N	8495,81	0,53%	30368,65	[8492,39 ; 8499,22]	8451,84	0,00%	32953,86	0,92
		5%N	8437,59	0,16%	16435,91	[8435,08 ; 8440,10]	8448,56	0,03%	18332,80	0,89
		10%N	8424,48	0,31%	5623,14	[8423,01 ; 8425,95]	8450,10	0,01%	6701,48	0,83
		15%N	8440,43	0,13%	2341,06	[8439,48 ; 8441,38]	8453,37	0,03%	2340,34	1,00
23 : 00	7140	1%N	7289,20	2,09%	4701441,79	[7246,70 ; 7331,69]	7284,92	2,03%	3924385,43	1,20
		3%N	7262,24	1,71%	12472,00	[7260,05 ; 7264,43]	7145,31	0,07%	14134,65	0,88
		5%N	7251,23	1,56%	5809,90	[7249,73 ; 7252,72]	7144,21	0,06%	6517,85	1,09
		10%N	7266,66	1,77%	2572,20	[7265,67 ; 7267,65]	7139,41	0,00%	3021,22	0,85
		15%N	7258,13	1,65%	1376,52	[7257,40 ; 7258,85]	7141,12	0,01%	1464,15	0,94
24 : 00	6061	1%N	6334,23	4,51%	3214735,07	[6299,09 ; 6369,37]	6002,77	0,96%	4409369,74	0,73
		3%N	6121,63	1,00%	34377,68	[6117,99 ; 6125,26]	6036,83	0,39%	24229,60	1,42
		5%N	6116,10	0,91%	20203,70	[6113,31 ; 6118,88]	6052,20	0,14%	19639,99	1,03
		10%N	6122,67	1,02%	6883,52	[6121,04 ; 6124,29]	6066,19	0,08%	7383,25	0,93
		15%N	6129,62	1,13%	2735,19	[6128,59 ; 6130,64]	6068,15	0,12%	3375,85	0,81

Tabela 7.5: Configuração 1 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	N	$E(\widehat{N}_{schn})$	ER	$Var(\widehat{N}_{schn})$	$IC_{95\%}$	ef_{MCRM}^{M2C2E}
01 : 00	4914	4731,52	3,71%	309794,00	[4577,24 ; 4885,79]	{12,52; 0,03; 0,01; 0,00; 0,00}
02 : 00	3323	3154,55	5,06%	108627,10	[3063,19 ; 3245,91]	{6,87; 2,16; 0,15; 0,05; 0,02}
03 : 00	2181	2255,96	3,44%	60530,08	[2187,76 ; 2324,16]	{13,49; 0,27; 0,12; 0,05; 0,02}
04 : 00	1339	1212,17	9,47%	2531,28	[1198,22 ; 1226,11]	{291,12; 33,82; 7,92; 2,09; 0,67}
05 : 00	781	787,25	0,80%	722,36	[779,80 ; 794,69]	{259,07; 32,88; 12,53; 4,81; 2,08}
06 : 00	482	447,85	7,08%	341,87	[442,72 ; 452,98]	{372,11; 103,13; 47,54; 15,69; 7,44}
07 : 00	357	379,91	6,42%	225,54	[375,75 ; 384,07]	{292,76; 59,95; 23,24; 12,51; 6,19}
08 : 00	504	493,09	2,16%	318,90	[488,14 ; 498,04]	{375,46; 66,23; 26,73; 6,91; 2,93}
09 : 00	1418	1228,61	13,35%	9090,40	[1202,18 ; 1255,04]	{71,00; 5,89; 1,35; 0,41; 0,11}
10 : 00	3170	2852,22	10,02%	108644,30	[2837,42 ; 2867,02]	{14,39; 0,01; 0,00; 0,00; 0,00}
11 : 00	5918	5671,69	4,16%	231506,80	[5538,32 ; 5805,06]	{13,21; 0,02; 0,00; 0,00; 0,00}
12 : 00	8186	7144,48	12,72%	911029,10	[6879,91 ; 7409,05]	{4,14; 0,00; 0,00; 0,00; 0,00}
13 : 00	8139	7324,57	10,01%	2584773,00	[6878,93 ; 7770,21]	{2,02; 0,00; 0,00; 0,00; 0,00}
14 : 00	6598	5846,51	11,39%	772077,80	[5602,95 ; 6090,07]	{5,68; 0,00; 0,00; 0,00; 0,00}
15 : 00	5193	5235,04	0,81%	3684272,00	[5602,95 ; 6090,07]	{1,13; 0,00; 0,00; 0,00; 0,00}
16 : 00	5896	6089,92	3,29%	110932,90	[5997,60 ; 6182,24]	{33,24; 0,17; 0,11; 0,02; 0,01}
17 : 00	6623	5895,66	17,34%	829725,40	[5643,17 ; 6148,15]	{4,10; 0,00; 0,00; 0,00; 0,00}
18 : 00	6255	5474,49	12,48%	762827,10	[5232,40 ; 5716,58]	{4,02; 0,03; 0,01; 0,00; 0,00}
19 : 00	3869	3606,64	6,78%	90177,72	[3523,40 ; 3689,88]	{24,56; 0,17; 0,07; 0,03; 0,02}
20 : 00	6129	5686,55	7,22%	390310,90	[5513,38 ; 5859,72]	{7,72; 0,01; 0,01; 0,00; 0,00}
21 : 00	7442	8405,63	12,95%	8582680,00	[7593,58 ; 9217,68]	{0,80; 0,00; 0,00; 0,00; 0,00}
22 : 00	8451	9399,58	11,22%	2512848,00	[8960,18 ; 9838,97]	{2,81; 0,01; 0,00; 0,00; 0,00}
23 : 00	7140	7023,07	1,64%	1802909,00	[6650,88 ; 7395,25]	{2,60; 0,00; 0,00; 0,00; 0,00}
24 : 00	6061	5177,24	12,38%	184049,00	[5058,32 ; 5296,15]	{17,47; 0,18; 0,11; 0,04; 0,01}

Tabela 7.6: Resultados do MCRM com $n_1 = n_2 = ... = n_j = 100$ para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HH_mod})$	ER	$Var(\widehat{\tau}_{HH_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HH})$	ER	$Var(\widehat{\tau}_{HH})$	
01 : 00	4914	1%N	4265,59	13,19%	7270014,35	[4212,74 ; 4318,44]	4858,80	1,12%	9149847,22	0,79
		3%N	4928,96	0,30%	3051395,06	[4894,72 ; 4963,19]	5252,38	6,89%	3337941,11	0,91
		5%N	5035,46	2,47%	1885001,57	[5008,55 ; 5062,37]	4713,41	4,08%	1811600,82	1,04
		10%N	4815,61	2,00%	940235,38	[4796,60 ; 4834,62]	4835,18	1,60%	747537,59	1,26
		15%N	4972,44	1,19%	518007,15	[4958,33 ; 4986,55]	4784,49	2,63%	591756,15	0,87
02 : 00	3323	1%N	3821,74	15,01%	4224769,71	[3418,87 ; 4224,60]	3209,94	3,40%	3817107,94	1,11
		3%N	3449,57	3,81%	1910721,55	[3422,48 ; 3476,66]	3396,37	2,21%	1755728,51	1,09
		5%N	3281,95	1,23%	861569,93	[3263,76 ; 3300,14]	3393,05	2,11%	860548,65	1,00
		10%N	3318,07	0,15%	458342,92	[3304,80 ; 3331,34]	3359,10	1,09%	490938,40	0,93
		15%N	3321,25	0,05%	332018,33	[3309,96 ; 3332,54]	3203,31	3,60%	267088,49	1,24
03 : 00	2181	1%N	2073,75	4,92%	1907505,28	[2046,68 ; 2100,82]	2038,19	6,55%	1715950,47	1,11
		3%N	2061,78	5,47%	701373,34	[2045,36 ; 2078,19]	2269,33	4,05%	663766,54	1,06
		5%N	2180,50	0,02%	390157,40	[2168,26 ; 2192,74]	2022,96	7,25%	347735,62	1,12
		10%N	2188,57	0,35%	202762,84	[2179,74 ; 2197,39]	2226,64	2,09%	201537,44	1,00
		15%N	2219,92	1,78%	125313,43	[2212,98 ; 2226,86]	2218,13	1,70%	147841,36	0,85
04 : 00	1339	1%N	1367,33	2,12%	1304342,98	[1344,94 ; 1389,71]	1309,60	2,19%	752221,21	1,73
		3%N	1329,19	0,73%	428626,42	[1316,36 ; 1342,02]	1352,02	0,97%	360812,24	1,19
		5%N	1340,97	0,14%	191487,54	[1332,39 ; 1349,54]	1397,51	4,37%	209759,79	0,91
		10%N	1337,52	0,11%	111305,18	[1330,98 ; 1344,06]	1377,97	2,91%	87247,26	1,27
		15%N	1308,17	2,30%	75820,02	[1302,77 ; 1313,57]	1335,48	0,26%	68186,09	1,11
05 : 00	781	1%N	703,72	9,89%	293427,14	[693,10 ; 714,34]	770,95	1,28%	302151,04	0,97
		3%N	804,20	2,97%	84488,91	[798,50 ; 809,89]	776,44	0,58%	87000,41	0,97
		5%N	752,42	3,66%	51215,44	[747,98 ; 756,86]	776,09	0,63%	59261,51	0,86
		10%N	780,36	0,08%	21625,96	[777,48 ; 783,24]	761,81	2,45%	30561,29	0,71
		15%N	771,92	1,16%	11645,88	[769,80 ; 774,03]	784,73	0,47%	16715,17	0,70

Tabela 7.7: Configuração 2 - Resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HH_mod})$	ER	$Var(\widehat{\tau}_{HH_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HH})$	ER	$Var(\widehat{\tau}_{HH})$	
06 : 00	482	1%N	456,96	5,19%	174165,20	[448,78 ; 465,14]	481,68	0,06%	127479,38	1,37
		3%N	471,74	2,13%	48512,13	[467,42 ; 476,06]	500,45	3,83%	56979,43	0,85
		5%N	476,81	1,08%	32245,26	[473,29 ; 480,33]	498,59	3,44%	33850,09	0,95
		10%N	478,09	0,81%	13736,25	[475,79 ; 480,39]	476,69	1,10%	13169,02	1,04
		15%N	489,79	1,62%	10888,61	[487,75 ; 491,83]	485,69	0,76%	7465,57	1,46
07 : 00	357	1%N	335,41	6,05%	77940,75	[329,94 ; 340,88]	331,67	7,09%	75841,04	1,03
		3%N	300,09	15,94%	19958,78	[297,32 ; 302,86]	346,55	2,93%	26664,37	0,75
		5%N	341,93	4,22%	12106,22	[339,77 ; 344,08]	366,73	2,72%	15023,52	0,80
		10%N	312,88	12,36%	7625,79	[311,17 ; 314,59]	364,42	2,08%	10122,78	0,75
		15%N	327,42	8,28%	5067,08	[326,02 ; 328,81]	355,47	0,43%	4874,20	1,04
08 : 00	504	1%N	510,24	1,24%	141255,11	[502,87 ; 517,61]	479,06	4,95%	155824,58	0,91
		3%N	518,58	2,89%	55249,02	[513,97 ; 523,19]	513,25	1,84%	49502,35	1,11
		5%N	534,36	6,02%	29666,51	[530,98 ; 537,74]	535,04	6,16%	26655.42	1,11
		10%N	517,68	2,71%	15996,78	[515,20 ; 520,16]	501,43	0,51%	16653,05	0,96
		15%N	493,27	2,13%	9591,75	[491,35 ; 495,19]	510,24	1,24%	9864,30	0,97
09 : 00	1418	1%N	1341,49	5,39%	1070237,29	[1321,21 ; 1361,77]	1434,82	1,18%	1105827,64	0,97
		3%N	1409,00	0,63%	347372,47	[1397,45 ; 1420,55]	1448,98	2,18%	416190,77	0,83
		5%N	1471,94	3,80%	232669,32	[1462,48 ; 1481,39]	1393,64	1,72%	151779,71	1,53
		10%N	1476,23	4,11%	110837,26	[1469,70 ; 1482,76]	1409,28	0,61%	84347,42	1,31
		15%N	1384,48	2,36%	58224,83	[1379,75 ; 1389,21]	1417,99	0,00%	65422,99	0,89
10 : 00	3170	1%N	3078,16	2,89%	4986466,07	[3034,39 ; 3121,93]	3127,11	1,35%	3480278,63	1,43
		3%N	3245,87	2,39%	1410061,05	[3222,59 ; 3269,14]	3074,42	3,01%	1418836,71	0,99
		5%N	3170,40	0,01%	769597,05	[3153,21 ; 3187,59]	2969,66	6,32%	713580,92	1,08
		10%N	3256,47	2,73%	407162,70	[3243,96 ; 3268,98]	3269,19	3,13%	374393,99	1,08
		15%N	3200,89	0,97%	279391,22	[3190,53 ; 3211,25]	3123,14	1,48%	321919,40	0,87

Tabela 7.8: Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HH_mod})$	ER	$Var(\widehat{\tau}_{HH_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HH})$	ER	$Var(\widehat{\tau}_{HH})$	
11 : 00	5918	1%N	6171,38	4,28%	13998700,55	[6098,05 ; 6244,71]	6538,21	10,48%	12130245,56	1,15
		3%N	5757,49	2,71%	3752982,42	[5719,52 ; 5795,46]	6166,24	4,19%	3217716,28	1,17
		5%N	5814,27	1,75%	3040421,62	[5780,09 ; 5848,45]	5782,07	2,29%	2570841,90	1,18
		10%N	6028,14	1,86%	1158762,23	[6010,71 ; 6045,56]	5925,33	0,12%	1086963,47	1,07
		15%N	6001,41	1,41%	790218,65	[5983,99 ; 6018,83]	6022,72	1,77%	723851,22	1,09
12 : 00	8186	1%N	8390,11	2,49%	20344630,19	[8301,70 ; 8478,52]	8062,90	1,50%	20958881,47	0,97
		3%N	7617,57	6,94%	8480513,64	[7529,16 ; 7705,98]	7818,42	4,49%	5694209,33	1,49
		5%N	8352,73	2,04%	4707017,81	[8310,21 ; 8395,25]	8307,04	1,48%	5886102,02	0,80
		10%N	8342,45	1,91%	2397788,42	[8312,09 ; 8372,80]	8433,82	3,02%	2223438,81	1,08
		15%N	8199,71	0,17%	1490451,70	[8175,78 ; 8223,64]	8303,01	1,43%	1238719,49	1,20
13 : 00	8139	1%N	7932,49	2,54%	19738179,30	[7845,41 ; 8019,57]	7971,02	2,06%	25044734,73	0,79
		3%N	8019,70	1,47%	8179360,62	[7963,65 ; 8075,76]	7953,21	2,28%	8043085,53	1,02
		5%N	7841,64	3,65%	5686006,30	[7794,90 ; 7888,38]	8172,59	0,41%	4663538,82	1,22
		10%N	8068,29	0,86%	2422943,96	[8037,78 ; 8098,79]	8206,47	0,83%	2250942,43	1,08
		15%N	8476,31	4,14%	1512061,26	[8452,21 ; 8500,41]	8120,29	0,23%	1504127,35	1,00
14 : 00	6598	1%N	6639,99	0,63%	16843232,38	[6559,55 ; 6720,43]	6039,53	8,46%	13986345,27	1,20
		3%N	6214,47	5,81%	6060990,82	[6166,22 ; 6262,72]	6412,57	2,81%	4941956,04	1,22
		5%N	6147,68	6,82%	3599305,82	[6112,08 ; 6183,28]	6523,87	1,12%	3263375,79	1,10
		10%N	6699,56	1,54%	1519236,10	[6675,40 ; 6723,72]	6707,51	1,65%	1212097,64	1,25
		15%N	6631,34	0,51%	911635,73	[6612,63 ; 6650,05]	6577,38	0,31%	1150708,13	0,79
15 : 00	5193	1%N	4933,08	5,00%	9845111,92	[4871,58 ; 4994,58]	4578,84	11,82%	10618423,58	0,93
		3%N	5435,37	4,67%	3995395,61	[5396,19 ; 5474,55]	5172,36	0,40%	3566456,52	1,12
		5%N	5165,68	0,53%	2074454,88	[5137,45 ; 5193,91]	5284,19	1,75%	2306913,86	0,90
		10%N	5129,45	1,22%	913156,80	[5110,72 ; 5148,18]	5361,81	3,25%	1059022,86	0,86
		15%N	5260,06	1,29%	544451,52	[5245,60 ; 5274,52]	5187,45	0,11%	624016,67	0,87

Tabela 7.9: Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HH_mod})$	ER	$Var(\widehat{\tau}_{HH_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HH})$	ER	$Var(\widehat{\tau}_{HH})$	
16 : 00	5896	1%N	5562,07	5,66%	11249067,26	[5496,33 ; 5627,81]	5959,23	1,07%	12253488,99	0,92
		3%N	6182,26	4,86%	3993636,22	[6143,09 ; 6221,43]	5849,27	0,79%	5020522,21	0,79
		5%N	5804,46	1,55%	2081868,11	[5776,18 ; 5832,74]	5881,13	0,25%	2802220,59	0,74
		10%N	5869,22	0,45%	1351645,01	[5846,43 ; 5892,01]	5952,87	0,96%	1019478,69	1,32
		15%N	5771,39	2,11%	717351,62	[5754,79 ; 5787,99]	5961,81	1,12%	879517,22	0,81
17 : 00	6623	1%N	6240,89	5,77%	12685942,26	[6171,08 ; 6310,70]	6742,82	1,81%	17999090,14	0,70
		3%N	6569,46	0,81%	3436017,33	[6533,13 ; 6605,79]	6834,10	3,19%	4869760,76	0,71
		5%N	6839,41	3,27%	2728368,71	[6807,03 ; 6871,78]	6748,32	1,89%	3580552,62	0,76
		10%N	6414,05	3,15%	1233446,87	[6392,28 ; 6435,82]	6604,55	0,28%	1115488,97	1,10
		15%N	6493,15	1,96%	699154,11	[6476,76 ; 6509,54]	6745,33	1,85%	845087,58	0,83
18 : 00	6255	1%N	6038,69	3,46%	11107935,36	[5973,37 ; 6104,01]	6525,85	4,33%	16894565,13	0,66
		3%N	6419,15	2,62%	5836458,50	[6371,80 ; 6466,50]	6034,10	3,53%	4774262,12	1,22
		5%N	6567,10	4,99%	3332307,51	[6531,32 ; 6602,88]	6381,44	2,02%	3437560,88	0,97
		10%N	6158,96	1,54%	1655430,93	[6133,74 ; 6184,18]	6099,73	2,48%	1456678,91	1,14
		15%N	6122,10	2,12%	1034701,05	[6102,16 ; 6142,04]	6363,28	1,73%	1150665,03	0,90
19 : 00	3869	1%N	3497,13	9,61%	5279315,16	[3452,09 ; 3542,16]	3894,96	0,67%	5523542,94	0,96
		3%N	3787,43	2,11%	2230445,45	[3758,16 ; 3816,70]	4065,14	5,07%	1769057,60	1,26
		5%N	3918,30	1,27%	1398657,55	[3895,12 ; 3941,48]	3854,74	0,37%	1337552,82	1,05
		10%N	4007,92	3,59%	639501,53	[3992,25 ; 4023,59]	3856,25	0,33%	551228,78	1,16
		15%N	3796,43	1,87%	387473,44	[3784,23 ; 3808,63]	3848,20	0,54%	413483,26	0,94
20 : 00	6129	1%N	6055,08	1,21%	13185279,70	[5983,91 ; 6126,25]	5911,96	3,54%	12912377,20	1,02
		3%N	6502,22	6,08%	5219469,57	[6457,44 ; 6546,99]	6225,65	1,57%	4923068,00	1,06
		5%N	6053,32	1,23%	2762046,45	[6020,75 ; 6085,89]	6293,68	2,69%	3433524,86	0,80
		10%N	6306,64	2,89%	1740010,93	[6280,78 ; 6332,49]	6195,47	1,08%	1321088,59	1,31
		15%N	6113,79	0,25%	938441,19	[6094,80 ; 6132,78]	6164,03	0,57%	969921,20	0,97

Tabela 7.10: Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

Horários	τ	M2C2E					MO			ef_{MO}^{M2C2E}
		z_1	$E(\widehat{\tau}_{HH_mod})$	ER	$Var(\widehat{\tau}_{HH_mod})$	$IC_{95\%}$	$E(\widehat{\tau}_{HH})$	ER	$Var(\widehat{\tau}_{HH})$	
21 : 00	7442	1%N	7758,97	4,26%	22625834,45	[7665,74 ; 7852,20]	6442,53	13,43%	16771747,68	1,35
		3%N	7083,69	4,81%	8721098,77	[7025,81 ; 7141,57]	7896,52	6,11%	8183421,77	1,06
		5%N	7634,03	2,58%	4061631,29	[7594,53 ; 7673,53]	7510,60	0,92%	3903195,82	1,04
		10%N	7390,03	0,70%	2299479,38	[7360,31 ; 7419,75]	7401,38	0,55%	1782969,57	1,29
		15%N	7456,08	0,19%	1131193,48	[7435,23 ; 7476,92]	7433,49	0,11%	1338938,49	0,84
22 : 00	8451	1%N	8395,84	0,65%	23480420,33	[8300,86 ; 8490,81]	8293,56	1,86%	30219250,85	0,77
		3%N	8421,50	0,35%	7721641,99	[8367,03 ; 8475,96]	8470,70	0,23%	7715937,03	1,00
		5%N	8353,55	1,15%	5497907,08	[8307,59 ; 8399,51]	8203,75	2,92%	5664337,44	0,97
		10%N	8479,86	0,34%	2522609,98	[8448,73 ; 8510,99]	8483,55	0,38%	2130889,56	1,18
		15%N	8495,11	0,52%	1543339,76	[8455,51 ; 8504,21]	8410,15	0,48%	1530278,79	1,00
23 : 00	7140	1%N	7784,25	9,02%	22681182,49	[7690,90 ; 7877,59]	6945,87	2,72%	17472456,85	1,30
		3%N	7717,80	8,09%	5835888,25	[7670,45 ; 7765,15]	7043,72	1,35%	6314408,45	0,92
		5%N	7307,69	2,35%	3787083,80	[7269,55 ; 7345,83]	7486,32	4,85%	4559535,51	0,83
		10%N	7143,81	0,05%	1704464,60	[7118,22 ; 7169,39]	6983,92	2,18%	1412697,68	1,20
		15%N	7297,96	2,21%	1074465,49	[7277,64 ; 7318,27]	7168,95	0,40%	994553,99	1,08
24 : 00	6061	1%N	6600,98	8,91%	14933031,58	[6525,24 ; 6676,72]	5850,18	3,48%	17394576,49	0,85
		3%N	6308,91	4,09%	4032546,13	[6269,55 ; 6348,27]	5974,79	1,42%	3954076,70	1,02
		5%N	6211,87	2,49%	3301912,29	[6176,25 ; 6247,48]	6014,86	0,76%	4274221,73	0,77
		10%N	6231,44	2,80%	981390,33	[6212,02 ; 6250,86]	6066,47	0,09%	1304424,65	0,75
		15%N	6205,32	2,38%	781754,96	[6187,99 ; 6222,65]	6043,10	0,29%	844741,82	0,92

Tabela 7.11: Configuração 2 - Continuação dos resultados para o total populacional de táxi no município do Rio de Janeiro no dia 22 de junho de 2016 para $n_1, n_2, ..., n_j$ capturando aleatoriamente de 1 a 10 elementos.

7.2.3 Análise dos Métodos

Nos resultados, referente a primeira configuração, para o total populacional nas Tabelas 7.1, 7.2, 7.3, 7.4 e 7.5, é possível observar que os maiores valores de variância e os erros relativos mais elevados estão em $z_1 = 1\%N$, ou seja, o tamanho inicial de células selecionadas implica na acurácia das estimativas. Observando a variável $IC_{95\%}$ no M2C2E em apenas 15 dos 120 intervalos de confiança, o parâmetro estava incluso. De forma geral, as eficiências apontam para um equilíbrio entre o M2C2E e o MO, exceto no horário de 02:00hs para $z_1 = 3\%N$ onde o MO tem uma variância 7,29 vezes menor do que o M2C2E.

A Tabela 7.6 tem um valor médio de erro relativo de 7,74% para o MCRC. Note que os valores médios para o erro relativo do MO e do M2C2E na primeira configuração são 0,78% e 1,46%, respectivamente, e na configuração 2, para MO é 2,35% e para o M2C2E é 3,09%, isso significa que o MCRM separadamente apresenta essa desvantagem em comparação ao MO e ao M2C2E. Continuando a observar os resultados do MCRM, em apenas 3 dos 24 intervalos de confiança $IC_{95\%}$, o valor do parâmetro estava contido.

Uma observação relevante a ser considerada na Tabela 7.6 é a eficiência ef_{MCRM}^{M2C2E}, nos horários em que o total populacional é menor, por exemplo, 06:00hs ou 07:00hs, a variância do M2C2E é superior que a do MCRM. Entretanto, esta superioridade pode estar apontando para um fator a ser considerado, pois, computacionalmente, todos os elementos estão a disposição e podem ser capturados no MCRM, caso o estudo fosse realizado em campo, por exemplo, capturar elementos em 1 m^2 não é igual em 10000 km^2, ou seja, a área da região não está sendo levada em consideração na metodologia do MCRM.

A segunda configuração, para o total populacional nas Tabelas 7.7, 7.8, 7.9, 7.10 e 7.11, é possível observar que as maiores variâncias estão em $z_1 = 1\%N$, mas os erros relativos mais altos não encontram-se necessariamente em $z_1 = 1\%N$. A variável $IC_{95\%}$ no M2C2E em apenas 20 dos 120 intervalos de confiança, o parâmetro estava incluso. De forma geral, as eficiências apontam para um equilíbrio entre o M2C2E e o MO.

As Figuras 7.7, 7.8, 7.9, 7.10, 7.11 correspondem as estimativas médias de 100 replicações do MO e do M2C2E para cada horário na primeira configuração mencionada na Subseção 7.2.1. As Figuras 7.12, 7.13, 7.14, 7.15, 7.16 correspondem as estimativas médias de 100 replicações do MO e do M2C2E para cada horário na segunda configuração mencionada na Subseção 7.2.2. Em ambas as configurações, à medida que o valor de z_1 aumenta as estimativas médias do M2C2E e do MO se aproximam do valor verdadeiro.

Contudo, na configuração 1, a partir do $z_1 = 3\%N$, observa-se essa aproximação na Figura 7.8, por outro lado, na configuração 2, tal fato ocorre a partir do $z_1 = 10\%N$ na Figura 7.15. Vale salientar que as estimativas médias referente a 50 replicações do MCRM não variam devido a configuração escolhida ser a primeira ou a segunda, uma vez que o MCRM não apresenta refinamento de grade e utilizou o estimador de Schnabel independente da configuração, por isso suas estimativas não apresentam alterações causadas pelo valor de z_1, visto que esse método é constituído por uma grade 1x1.

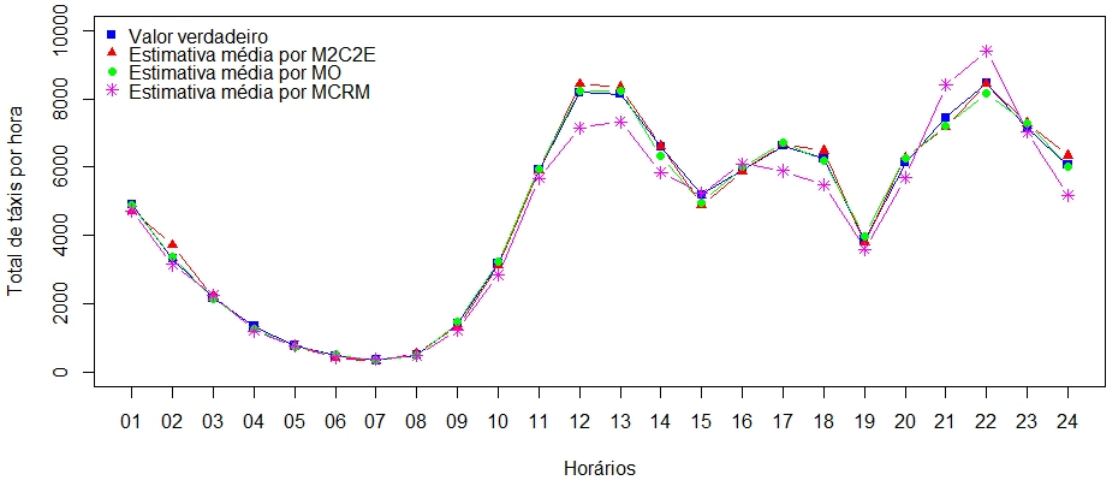


Figura 7.7: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 1\%N$.

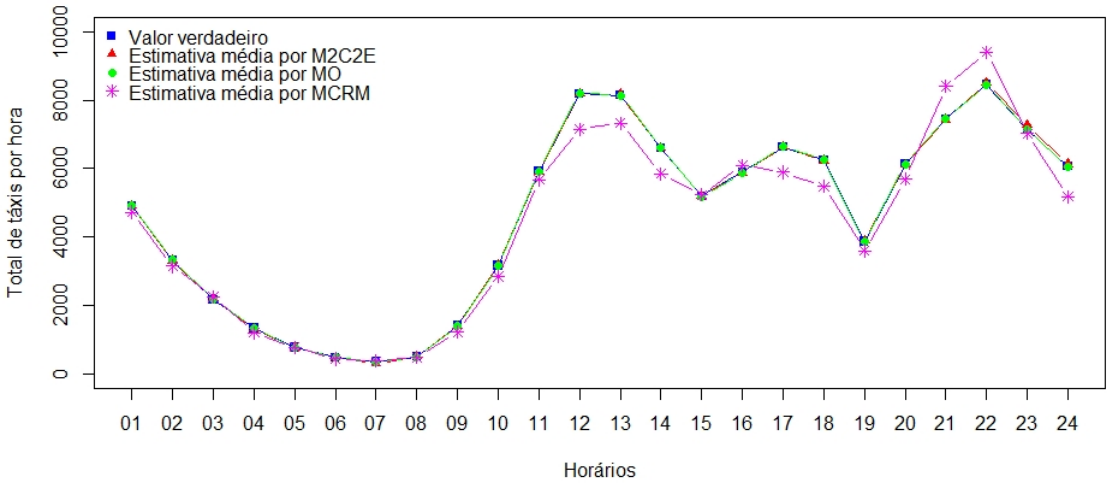


Figura 7.8: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 3\%N$.

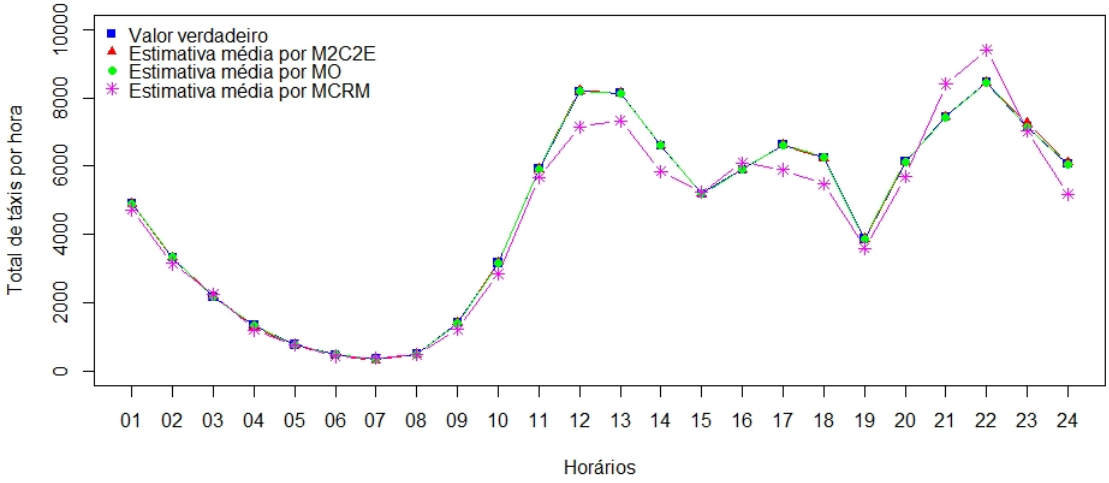


Figura 7.9: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 5\%N$.

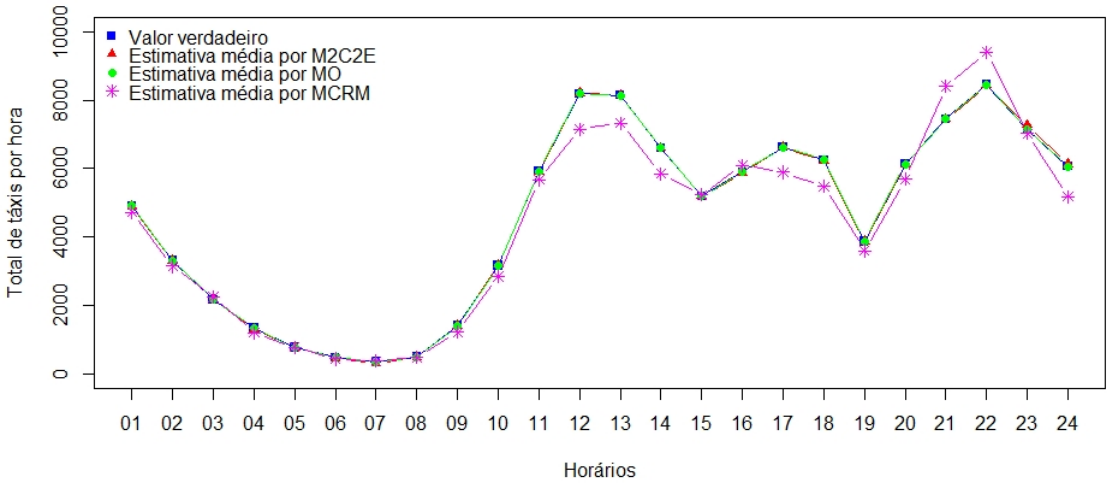


Figura 7.10: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 10\%N$.

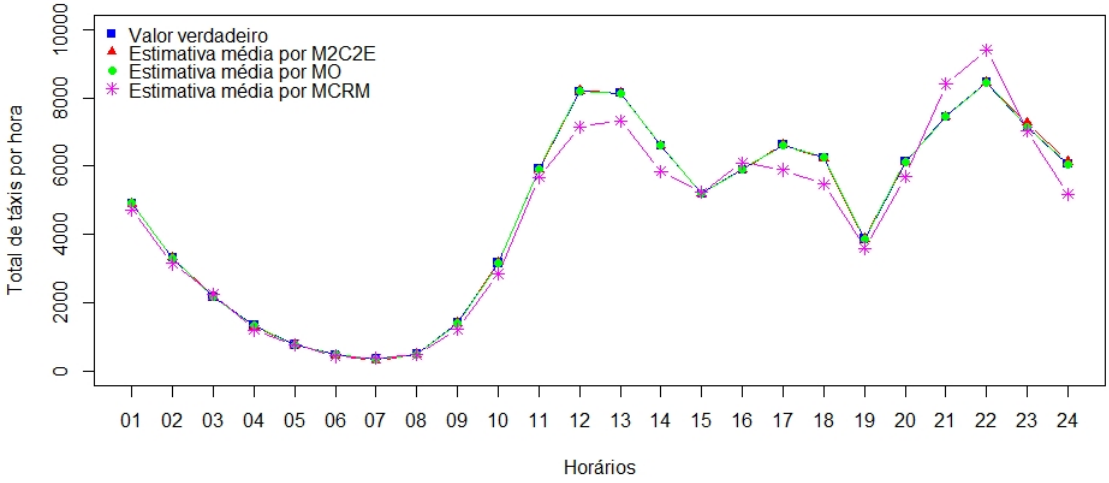


Figura 7.11: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 1 e $z_1 = 15\%N$.

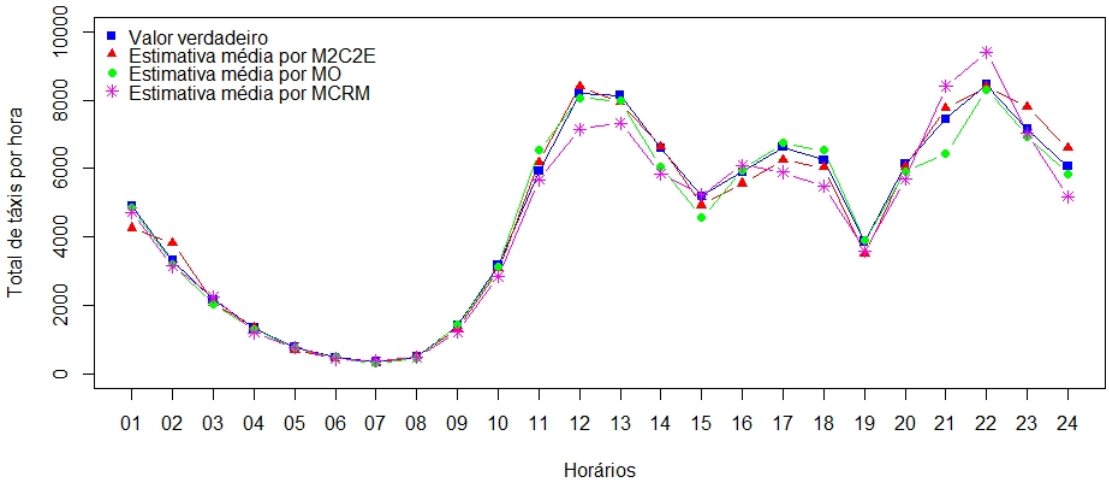


Figura 7.12: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 1\%N$.

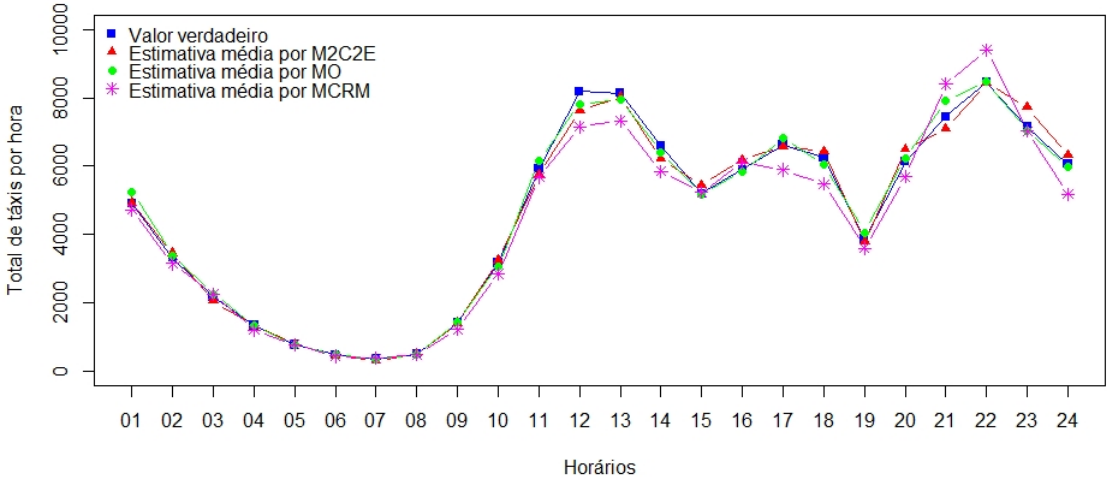


Figura 7.13: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 3\%N$.

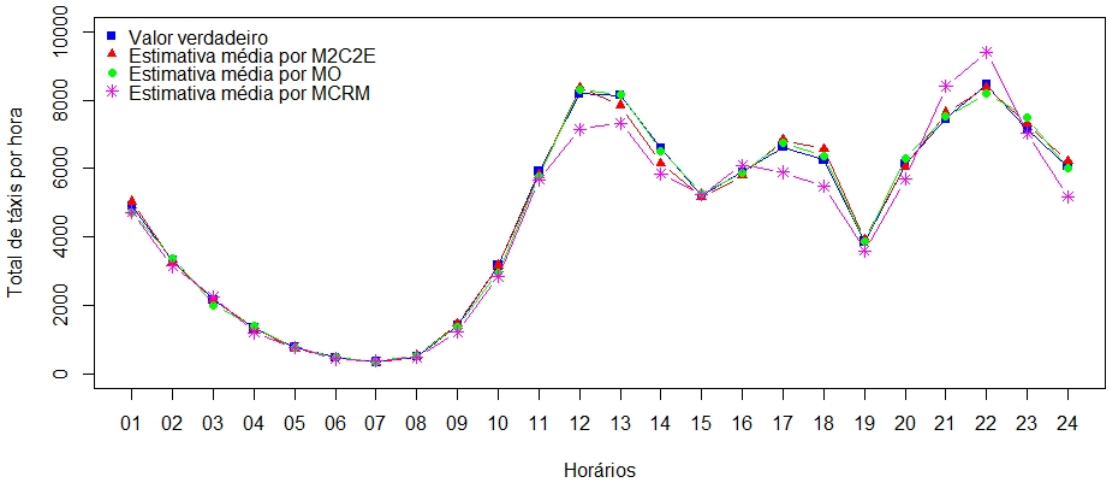


Figura 7.14: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 5\%N$.

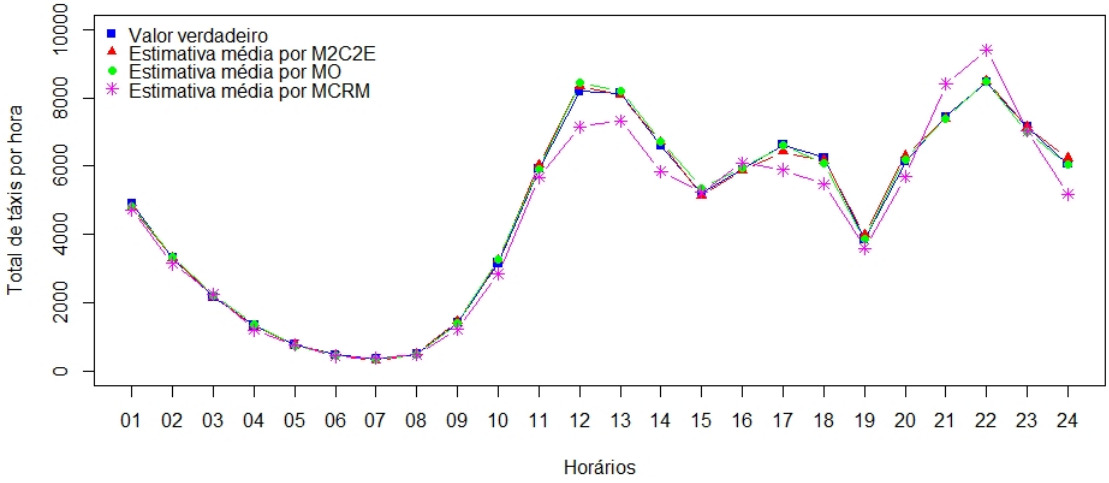


Figura 7.15: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 10\%N$.

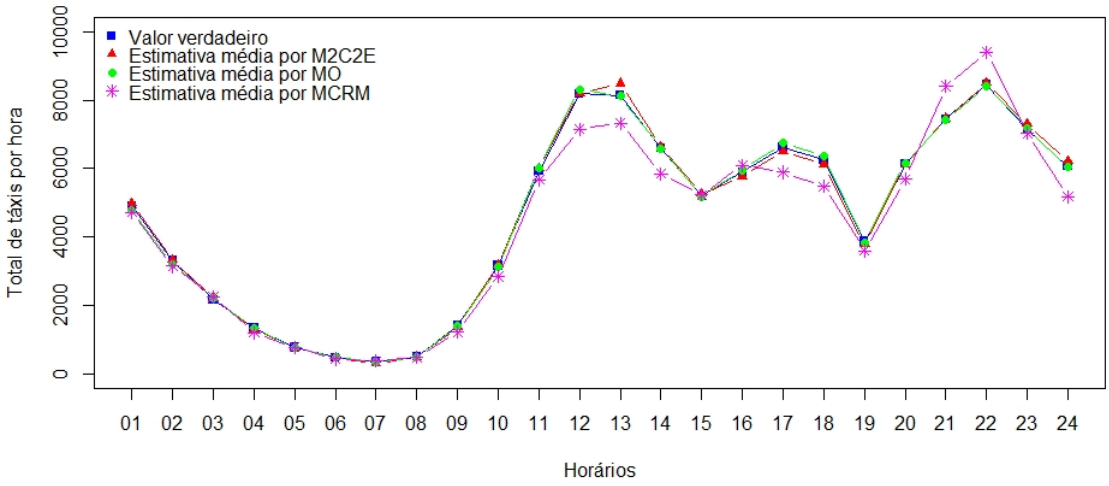


Figura 7.16: Estimativas médias para número total de táxis por Método Ótimo - MO, por Método 2-Camadas e 2-Estimadores - M2C2E e por Método de Captura e Recaptura Múltipla - MCRM na configuração 2 e $z_1 = 15\%N$.

A Figura 7.17 contém os boxplots dos erros relativos referentes as estimativas médias para os três métodos (MO, M2C2E e MCRM) e as duas configurações. Os valores superiores ao limite superior $=$ 3° quartil $+ 1,5 *$ (3° quartil - 1° quartil) ou inferiores ao limite inferior $=$ 1° quartil $- 1,5 *$ (3° quartil - 1° quartil) são chamados de outliers. Em particular, os valores superiores a 1,96%, 6,63%, 4,06% e 8,66% são outliers para o MO_{HT}, MO_{HH}, $M2C2E_{HT_mod}$ e $M2C2E_{HH_mod}$, respectivamente. A distribuição do erro relativo do MCRM não apresentou outliers, contudo esse método tem a maior amplitude interquartílica e concentração em valores mais elevados de erro relativo em comparação com os demais métodos. Na Tabela 7.12, é possível encontrar as medidas resumo referente a Figura 7.17.

Medidas Resumo	MO_{HT}	MO_{HH}	$M2C2E_{HT_mod}$	$M2C2E_{HH_mod}$	MCRM
Mínimo	0,00%	0,00%	0,00%	0,01%	0,80%
1° Quartil	0,04%	0,66%	0,16%	1,16%	3,64%
Mediana	0,17%	1,74%	0,45%	2,26%	7,15%
Média	0,79%	2,35%	1,46%	3,10%	7,74%
3° Quartil	0,81%	3,05%	1,73%	4,16%	11.64%
Máximo	7,73%	13,43%	10,45%	15,94%	17.34%

Tabela 7.12: Estatística descritiva dos erros relativos referentes as estimativas médias do Método Ótimo, do Método 2-Camadas e 2-Estimadores e do Método de Captura e Recaptura Múltipla nas configurações 1 e 2.

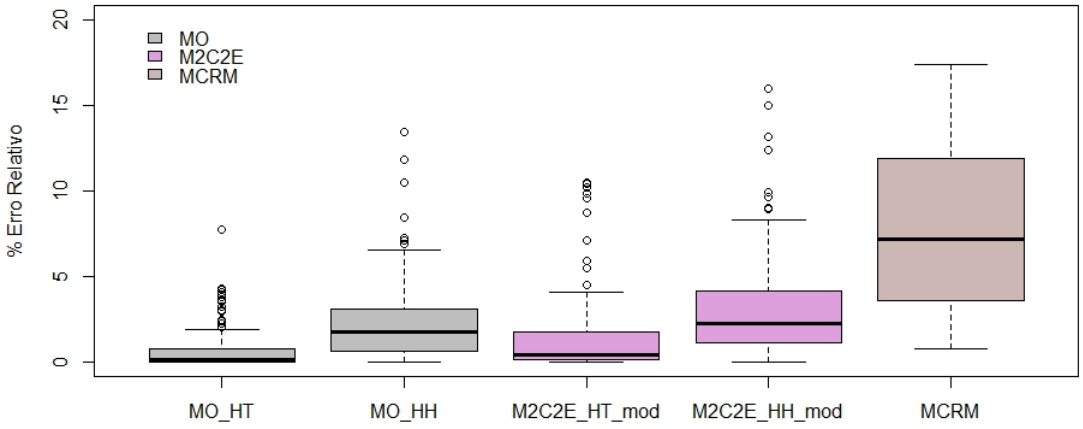


Figura 7.17: Boxplots dos erros relativos referentes as estimativas médias do método ótimo usando o estimador de Horvitz-Thompson - MO com HT e estimador de Hansen-Hurwitz - MO com HH, do framework proposto usando o estimador de Horvitz-Thompson - M2C2E com HT_mod e estimador de Hansen-Hurwitz - M2C2E com HH_mod e MCRM.

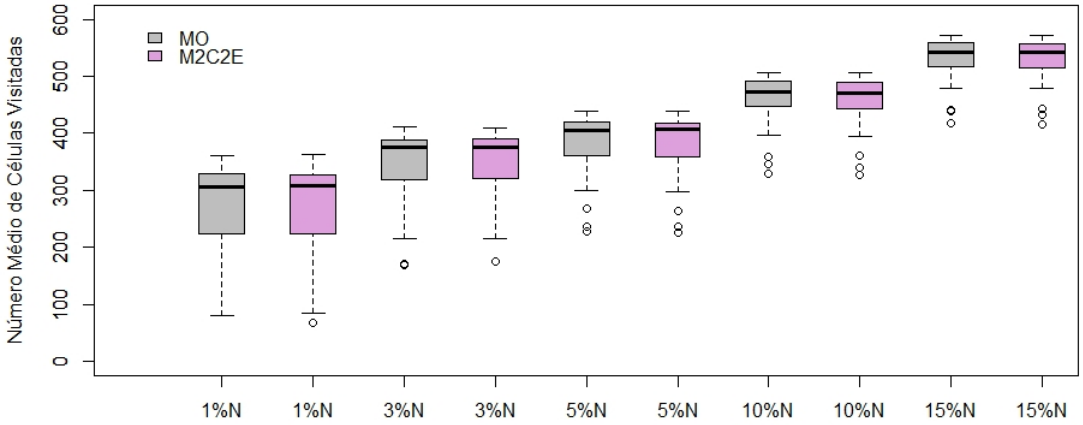


Figura 7.18: Boxplots do número médio de células visitadas do método ótimo - MO e do framework proposto - M2C2E, por número de células visitadas inicialmente $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$.

A Figura 7.18 contém os boxplots com a amostra inicial de células visitadas $z_1 = \{1\%N, 3\%N, 5\%N, 10\%N, 15\%N\}$ no eixo X e o número médio de células visitadas no eixo Y para o MO e para o M2C2E. Cada um deles é formado pelas 24 horas de cada tamanho amostral inicial z_1, ou seja, os primeiros boxplots são $1\%N$ com todos os horários (01:00hs, 02:00hs, ... , 24:00hs) até os últimos que são $15\%N$ com todos os horários

(01:00hs, 02:00hs, ... , 24:00hs). Os valores inferiores a 99,31; 232,10; 281,75; 384,80 e 460,25 são outliers para o MO e os valores inferiores a 94,34; 229,10; 278,20; 377,40 e 453,50 são outliers para o M2C2E.

7.2.4 Teste de Hipótese para Comparação dos Métodos

Usando o teste de hipótese não paramétrico Wilcoxon Mann-Whitney para duas amostras independentes, é possível notar que as distribuições do erro relativo na Figura 7.17 apresentam diferenças significativas entre os métodos MO e M2C2E com p-valores menor do que o nível de significância de 5% iguais a 0,00022 entre MO_{HT} e o $M2C2E_{HT_mod}$; e 0,04201 entre MO_{HH} e o $M2C2E_{HH_mod}$. Tal fato aponta que o $M2C2E_{HT_mod}$ e o $M2C2E_{HH_mod}$ erra mais do que o MO_{HT} e MO_{HH}, mas isso é esperado, pois o framework proposto utiliza-se de estimativa enquanto o MO de parâmetro dentro das células selecionadas. O MCRM não foi testado em relação ao erro relativo com os outros métodos, porque a sua mediana difere dos demais.

Aplicando o mesmo teste de hipótese não paramétrico, tem-se que as distribuições do número médio de células visitadas na Figura 7.18 não contém diferença significativa entre os dois métodos com p-valores iguais a 0,94; 0,95; 0,89; 0,81 e 0,97 para os valores de $z_1 = \{1\%N, 3\%N, 5\%N, 10\%, 15\%N\}$, respectivamente. Isso significa que ao nível de significância de 5%, não há evidências de que essas distribuições sejam diferentes em pares, ou seja, $1\%N$ para M2C2E e MO, $3\%N$ para M2C2E e MO, $5\%N$ para M2C2E e MO, 10% para M2C2E e MO, e 15% para M2C2E e MO.

O teste de hipótese T^2 Hotelling para verificar a igualdade entre os vetores de estimativas médias do método MO e do método M2C2E foi realizado sob as seguintes hipóteses:

Para a configuração 1:

$$\text{H0: } \vec{\mu}_{M2C2E_{HT_mod}} = \vec{\mu}_{MO_{HT}} \text{ versus H1: } \vec{\mu}_{M2C2E_{HT_mod}} \neq \vec{\mu}_{MO_{HT}}$$

Para a configuração 2:

$$\text{H0: } \vec{\mu}_{M2C2E_{HH_mod}} = \vec{\mu}_{MO_{HH}} \text{ versus H1: } \vec{\mu}_{M2C2E_{HH_mod}} \neq \vec{\mu}_{MO_{HH}}$$

Esse teste foi realizado considerando as estimativas médias obtidas pelo método MO e pelo método M2C2E por horário em todos os horários independente do tamanho inicial z_1, por exemplo, na configuração 1 para o horário de 01:00hs, tem-se que:

$$H0: \begin{pmatrix} 4724,41 \\ 4912,43 \\ 4900,42 \\ 4891,55 \\ 4901,56 \end{pmatrix} = \begin{pmatrix} 4880,82 \\ 4911,01 \\ 4905,12 \\ 4919,63 \\ 4909,56 \end{pmatrix} \quad \text{versus} \quad H1: \begin{pmatrix} 4724,41 \\ 4912,43 \\ 4900,42 \\ 4891,55 \\ 4901,56 \end{pmatrix} \neq \begin{pmatrix} 4880,82 \\ 4911,01 \\ 4905,12 \\ 4919,63 \\ 4909,56 \end{pmatrix}$$

Na configuração 1, os horários 03:00hs, 07:00hs, 11:00hs, 14:00hs, 16:00hs, 22:00hs e 23:00hs, ao nível de significância de 5%, há evidências para rejeitar que $\vec{\mu}_{M2C2E_{HT_mod}} = \vec{\mu}_{MO_{HT}}$ segundo os p-valores iguais a 0,01; 0,00; 0,00; 0,00; 0,00; 0,01 e 0,00, respectivamente. Portanto, nos demais horários não há evidências para rejeitar que $\vec{\mu}_{M2C2E_{HT_mod}} = \vec{\mu}_{MO_{HT}}$.

Por outro lado, na configuração 2, apenas nos horários de 07:00hs e de 24:00hs com p-valor de 0,01 em ambos, ao nível de significância de 5%, há evidências para rejeitar que $\vec{\mu}_{M2C2E_{HH_mod}} = \vec{\mu}_{MO_{HH}}$, para os demais horários não há evidências para rejeitar que $\vec{\mu}_{M2C2E_{HH_mod}} = \vec{\mu}_{MO_{HH}}$, ao nível de significância de 5%.

Dessa forma, o teste de hipótese T^2 Hotelling para comparação dos vetores de estimativas médias dos métodos M2C2E e MO com os dados reais de táxis no município do Rio de Janeiro no dia 22 de junho de 2016, ao nível de significância de 5%, evidenciam que a utilização do estimador de Hansen-Hurwitz aproxima as estimativas médias do M2C2E ao MO mais do que o estimador de Horvitz-Thompson.

Capítulo 8

Considerações Finais

Essa dissertação visa quantificar redes formadas por população rara e agrupada. O objetivo principal é solucionar uma lacuna da AAC no momento de obter o número de elementos dentro da célula, pois todos os elementos deveriam ser encontrados, por isso a AAC foi chamada de Método Ótimo - MO, devido ao fato do parâmetro ser conhecido. Contudo, afirmar a possibilidade de encontrar todos elementos pode não ser viável para todos os casos.

Uma possível solução encontrada ao problema é a estimação dentro de cada célula selecionada, através do MCRM o qual é considerado relevante por prover estimativas relevantes, ele tem a vantagem de consumir menos tempo e ser mais barato do que um censo populacional onde todos os elementos precisam ser observados, por exemplo. Entretanto, para utilizar o MCRM seria necessário determinar o critério de parada para o número de recapturas. Para tal problema foi apresentado o critério de parada proposto por Singham et al. [40] [41] [42], o qual foi possível pela adequação do valor de δ como sendo $\widehat{n}_1\varepsilon$, quando o valor de k^* for obtido, retornaria a estimativa na k^*-ésima recaptura.

O framework proposto recebeu o nome de Método 2-Camadas e 2-Estimadores - M2C2E e esse método procura tornar a AAC um plano amostral mais realista através das estimativas trazidas pelo MCRM dentro das células selecionadas. Os estudos com dados sintéticos mostraram que o critério de parada está relacionado com o valor máximo da meia largura δ e o número de elemento coletado a cada recaptura $n_1, n_2, ..., n_{k^*}$.

Vale ressaltar que o MCRM pode apresentar dificuldades, que limitam o critério de parada, nos casos em que o número de elementos na rede é extremamente pequeno, pelo fato de não encontrar elementos de interesse e em células muito densas, pois com o número de recapturados igual a zero, matematicamente, as estimativas não estão bem definidas

pelo estimador de Schnabel, por exemplo.

Na aplicação do M2C2E e do MO à população de táxis conectada a um aplicativo através de uma rede móvel no município do Rio de Janeiro, não houve o problema de células muito densas do MCRM, porque o número máximo de elementos dentro de uma célula não ultrapassou 400 táxis. Note que o problema no critério de parada devido a célula muito densa foi gerado com 10000 elementos em uma célula, o que seria um caso bem extremo.

Em relação ao tempo computacional, Townsend et al. [50] afirmaram que para obter estimativas precisas no contexto de população rara e agrupada, usando implementação computacional, podem exigir tempos elevados de execução. No software RStudio Cloud, cada algoritmo foi implantado em um contêiner no qual foi alocado 1 GB de RAM dedicada e 1 CPU Xeon E5-2666 V3. Sendo assim, as simulações com dados sintéticos nas grades de 10x10 e de 20x20 para 10 mil replicações levaram em torno de 5 minutos e 1 hora e 5 minutos, respectivamente, para cada um dos tamanhos iniciais z_1. Dessa forma, conclui-se que o número de células na grade é um fator decisivo para o tempo de processamento.

Os dados reais ao sobrepor uma grade de 40x40 com 100 replicações, o processamento dos dados levou em torno de 1 hora e 15 minutos para cada um dos tamanhos iniciais z_1 no RStudio Cloud com 1 GB de RAM dedicada e 1 CPU Xeon E5-2666 V3. Levando em consideração que foram dois métodos MO e M2C2E em duas configurações, o tempo totalizou 25 dias de processamento sem interrupções. Caso fosse utilizado o RStudio em um computador doméstico com uma configuração igual a Core I3 e 4 GB de RAM, levariam, aproximadamente, 52 dias de processamento sem interrupções para realizar o Capítulo 7.

Os estudos com dados sintéticos auxiliaram as análises sobre o critério de parada e evidenciam que o M2C2E apresenta-se melhor em comparação ao MO, quando o total populacional aumenta e o tamanho da grade diminui, mas partindo de um grau de agrupamento de $\phi = 0,1$. Por outro lado, a aplicação a dados reais configura um caso de uso aos sistemas distribuídos analisados por hora com uma variabilidade maior do total populacional entre 357 e 8451 elementos em comparação aos estudos com dados sintéticos. 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 Hansen-Hurwitz aproxima as estimativas médias do M2C2E ao MO mais do que o estimador de Horvitz-Thompson.

Os resultados permitem concluir que o M2C2E apresenta boas 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 a 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 a população do interior de cada célula, como pressuposto pela AAC, pois estima apenas a partir da amostra de alguns elementos.

8.1 Contribuições

As contribuições desta dissertação, do ponto de vista teórico, são:

- Adaptação do critério de parada proposto por Singham et al. [40] [41] [42] ao contexto do MCRM;
- Implementação do método M2C2E o qual visa contornar a lacuna da AAC; e,
- Modificação dos estimadores clássicos de Horvitz-Thompson e Hansen-Hurwitz.

8.2 Trabalhos Futuros

Os trabalhos futuros importantes a serem desenvolvidos seriam: analisar a eficiência de outros critérios de parada para o MCRM e a AAC os quais poderão ser implementados através da abordagem bayesiana, confeccionar um método adaptativo local de refinamento da grade e estudar o fator de agrupamento entre os elementos da rede, ou seja, saber o quanto essa variável influencia nos métodos. Nesta dissertação, por exemplo, o estudo com rede sintética apresentava-se mais agrupada do que os dados reais da rede de táxis.

Outro ponto a ser levado em consideração é conseguir modelar computacionalmente a proporção da área para o MCRM, tal necessidade ocorre devido ao ambiente computacional ser mais controlado do que em situações reais, pois durante o estudo com dados sintéticos e com dados reais todos os elementos têm a mesma chance de serem capturados independentemente do tamanho da área na qual os elementos de interesse estão contidos, por exemplo, na prática capturar um elemento em 1 m^2 não é igual a capturar em 10000 km^2. Portanto, o trabalho futuro seria a criação de peso a prior em função da área da célula na qual o elemento está inserido.

Referências

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

[2] Affonso, L. H. T. A. Alguns métodos de amostragem para populações raras e agrupadas. Master's thesis, Universidade de São Paulo.

[3] Akanda, M.; Salam, A. A generalized estimating equations approach to capture-recapture closed population models: methods.

[4] Arnold, H. Permutation support for multivariate techniques. *Biometrika 51*, 1-2 (1964), 65–70.

[5] Bailey, N. T. On estimating the size of mobile populations from recapture data. *Biometrika* (1951), 293–306.

[6] Blower, J. G.; Bishop, J. A.; Cook, L. M. *Estimating the size of animal populations.* 1981.

[7] Briand, L. C.; El Emam, K.; Freimut, B. G.; Laitenberger, O. A comprehensive evaluation of capture-recapture models for estimating software defect content. *IEEE Transactions on Software Engineering 26*, 6 (2000), 518–540.

[8] Brown, J.; Manly, B. Restricted adaptive cluster sampling. *Environmental and Ecological Statistics 5*, 1 (1998), 49–63.

[9] Brown, J. A.; Salehi, M.; Moradi, M.; Panahbehagh, B.; Smith, D. R. Adaptive survey designs for sampling rare and clustered populations. *Mathematics and Computers in Simulation 93* (2013), 108–116.

[10] Budrys, E.; Budrienė, A.; Pakalniškis, S. Population size assessment using mark-release-recapture of 12 species of orthoptera, diptera and hymenoptera: a comparison of methods. *Latvijas entomologs 41* (2004), 32–43.

[11] Castagliola, P.; Achouri, A.; Taleb, H.; Celano, G.; Psarakis, S. Monitoring the coefficient of variation using control charts with run rules. *Quality Technology & Quantitative Management 10*, 1 (2013), 75–94.

[12] Chao, A.; Yang, M. C. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika 80*, 1 (1993), 193–201.

[13] Cochran, W. G. *Sampling techniques.* John Wiley & Sons, 2007.

[14] Coeli, C. M.; Veras, R. P.; Coutinho, E. d. S. F. Metodologia de captura-recaptura: uma opção para a vigilância das doenças não transmissíveis na população idosa. *Cadernos de Saúde Pública 16* (2000), 1071–1082.

[15] Costanza, M. C.; Afifi, A. Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of the American Statistical Association 74*, 368 (1979), 777–785.

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

[17] DG, C. Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics 1* (1951), 131–160.

[18] Diggle, P. J.; Besag, J.; Gleaves, J. T. Statistical analysis of spatial point patterns by means of distance methods. *Biometrics* (1976), 659–667.

[19] Domhan, T.; Springenberg, J. T.; Hutter, F. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015).

[20] Dunn, J.; Andreoli, S. B. Método de captura e recaptura: nova metodologia para pesquisas epidemiológicas. *Revista de Saúde Pública 28* (1994), 449–453.

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

[22] Gattone, S. A.; Di Battista, T. Adaptive cluster sampling with a data driven stopping rule. *Statistical Methods & Applications 20*, 1 (2011), 1–21.

[23] Golovin, D.; Solnik, B.; Moitra, S.; Kochanski, G.; Karro, J.; Sculley, D. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), ACM, pp. 1487–1495.

[24] Hall, R. E. Analysis of the capture-recapture method of determining fish population size in a pond community.

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

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

[27] Hwang, W.-H.; Chao, A.; Yip, P. S. Theory & methods: Continuous-time capture-recapture models with time variation and behavioural response. *Australian & New Zealand Journal of Statistics 44*, 1 (2002), 41–54.

[28] Johnson, R. A.; Wichern, D. W., et al. *Applied multivariate statistical analysis*, vol. 5. Prentice hall Upper Saddle River, NJ, 2002.

[29] Jorge, M. C.; Alpizar-Jara, R. Captura-recaptura: um estudo de simulação para avaliar a performance de estimadores do tipo lincoln-petersen.

[30] King, R.; Overstall, A. Population size estimation and capture–recapture methods.

[31] Lincoln, F. C. *Calculating waterfowl abundance on the basis of banding returns.* No. 118. US Department of Agriculture, 1930.

[32] Mares, M.; Streilein, K.; Willig, M. Experimental assessment of several population estimation techniques on an introduced population of eastern chipmunks. *Journal of Mammalogy 62*, 2 (1981), 315–328.

[33] Mills, H. On the statistical validation of computer programs. *IBM FSD Report* (1972).

[34] Oliveira, C. I. F. *Método de captura e recaptura para a estimação da abundância de uma população: aplicação da metodologia Bootstrap.* Tese de Doutorado, 2007.

[35] Otis, D. L.; Burnham, K. P.; White, G. C.; Anderson, D. R. Statistical inference from capture data on closed animal populations. *Wildlife monographs*, 62 (1978), 3–135.

[36] Peng, S.-L.; Li, S.-S.; Liao, X.-K.; Peng, Y.-X.; Xiao, N. Estimation of a population size in large-scale wireless sensor networks. *Journal of Computer Science and Technology 24*, 5 (2009), 987–997.

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

[38] Schumacher, F.; Eschmeyer, R. The estimation of fish populations in lakes and ponds. *Journal of the Tennessee Academy of Science 18*, 228 (1943).

[39] Singh, H. P.; Yadav, A. A class of estimators for estimating the population mean and variance using auxiliary information under adoptive cluster sampling in sample surveys. *Bulletin of Pure & Applied Sciences-Mathematics and Statistics 38*, 1 (2019), 171–175.

[40] Singham, D. I. *Analysis of Sequential Stopping Rules for Simulation Experiments.* Tese de Doutorado, University of California, Berkeley, USA, 2010.

[41] Singham, D. I.; Schruben, L. W. Analysis of sequential stopping rules. In *Proceedings of the 2009 Winter Simulation Conference (WSC)* (2009), IEEE, pp. 723–730.

[42] Singham, D. I.; Schruben, L. W. Finite-sample performance of absolute precision stopping rules. *INFORMS Journal on Computing 24*, 4 (2012), 624–635.

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

[44] Solberg, K. H.; Bellemain, E.; Drageset, O.-M.; Taberlet, P.; Swenson, J. E. An evaluation of field and non-invasive genetic methods to estimate brown bear (ursus arctos) population size. *Biological Conservation 128*, 2 (2006), 158–168.

[45] Streeter, M. Bayes optimal early stopping policies for black-box optimization. *arXiv preprint arXiv:1902.08285* (2019).

[46] Tallmon, D. A.; Koyuk, A.; Luikart, G.; Beaumont, M. A. Onesamp: a program to estimate effective population size using approximate bayesian computation. *Molecular Ecology Resources 8*, 2 (2008), 299–301.

[47] Teo, J. Exploring dynamic self-adaptive populations in differential evolution. *Soft Computing 10*, 8 (2006), 673–686.

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

[49] Thompson, W. L. Estimating abundance of rare or elusive species. *Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters 389* (2004).

[50] Townsend, J. K.; Haraszti, Z.; Freebersyser, J. A.; Devetsikiotis, M. Simulation of rare events in communications networks. *IEEE Communications Magazine 36*, 8 (1998), 36–41.

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

[52] Yang, M. C.; Chao, A. Reliability-estimation and stopping-rules for software testing, based on repeated appearances of bugs. *IEEE Transactions on Reliability 44*, 2 (1995), 315–321.

[53] Yao, Y.; Rosasco, L.; Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation 26*, 2 (2007), 289–315.

[54] Zielinski, K.; Weitkemper, P.; Laur, R.; Kammeyer, K.-D. Examination of stopping criteria for differential evolution based on a power allocation problem. In *Proceedings of the 10th International Conference on Optimization of Electrical and Electronic Equipment* (2006), vol. 3, pp. 149–156.

APÊNDICE A – Aspectos Computacionais

A.1 População Rara e Agrupada: Geração e Contagem

```
#######################################
# Confecção de grade
#######################################
# Grade 20X20, ou seja, N = 400 células
x = 0
ladox = 20     #para definir o tamanho da grade
ladoy = 20     #para definir o tamanho da grade
y = seq(0,ladoy,l = ladoy+1)
grid = expand.grid(x,y)
x11(4,4)
par(mar=c(1,1,1,1))
plot(grid,col="black",lwd=1.7,axes=F,ylim=c(0,ladoy),xlim=c(0,ladox),
type="l",ylab=NA,xlab=NA,cex.lab=1.4)
for (i in 1:ladoy)
{x = i; grid = expand.grid(x,y); lines(grid,lwd=1.7)}
x = seq(0,ladox,l=ladox+1)
for (i in 0:ladox)
{y = i; grid = expand.grid(x,y); lines(grid,lwd=1.7)}
#######################################
# Geração da população rara e agrupada
#######################################
# Para os pontos "pais"
Nv = rpois(10,ladox) # Poisson, para obtemos valores inteiros
pais = length(Nv) #número de pais
```

```
copx=runif(pais,1,ladox-1) #coordenadas x do ponto pais
copy=runif(pais,1,ladoy-1) #coordenadas y do ponto pais
points(copx,copy,bg=25,pch=21,cex=0.45) # plota os pontos pais
coordp = matrix(c(copx,copy),pais,2,byrow = F)
# Para os pontos "filhos"
coordf = NULL
tamanho = 9  # número de filhos
sigma = 0.1   #para concentrar mais os filhos ou não
for(i in 1:pais){
cofx = rnorm(9,copx[i],sigma) #coordenadas x dos pontos filhos
while(any((cofx>ladox)|(cofx<0))){
aux = which((cofx>ladox)|(cofx<0))
cofx[aux] = rnorm(9,copx[i],sigma)}
#cofx[aux] = rnorm(length(aux),copx[i],sigma)}
cofy = rnorm(tamanho,copy[i],sigma) #coordenadas y dos pontos filhos
while(any((cofy>ladoy)|(cofy<0))){
aux = which((cofy>ladoy)|(cofy<0))
cofy[aux] = rnorm(9,copy[i],sigma)}
#cofy[aux] = rnorm(length(aux),copy[i],sigma)}
points(cofx,cofy,bg=25,pch=21,cex=0.45) #plota os pontos filhos
coord = matrix(c(cofx,cofy),9,2,byrow = F)
coordf = rbind(coordf,coord)
}
########################################
# Para identificar os elementos
########################################
coordenada_popul1 = rbind(coordp,coordf)
coordenada_popul1
coordx = coordenada_popul1[,1] #coordenada X
coordy = coordenada_popul1[,2] #coordenada Y
id_elemento = seq(1,length(coordenada_popul1)/2,1)
com_id_elemento = cbind(coordx,coordy,id_elemento)
```

```
########################################
# Contagem das variáveis de interesse nas células
########################################
x = coordenada_popul1[,1]
y = coordenada_popul1[,2]
matriz = array(0,c(20,20))
contagem = function(x,y){
for(j in 0:19){
for(i in 0:19){
cont = 0
for (k in 1:length(x)){
if(((x[k]<=(i+1))&&(x[k]>i))&&((y[k]<=(j+1))&&(y[k]>j)))
{
cont = cont+1
matriz[i+1,20-j]= cont
}
}
}
}
return(matriz)
}
pop_rara_agrupada = t(contagem(x,y))
pop_rara_agrupada
```

A.2 Método de Captura e Recaptura Múltipla: Estimador de Schnabel e Critério de Parada

```
########################################
# Matriz de Captura e Matrizes de Recaptura
########################################
elementos_por_celula1 = com_id_elemento_id_unidade[id_unidade==1301,1:4]
elementos_por_celula1
#apresenta todos os elementos dentro da célula selecionada.

captura_1cel = array(0,c(50,10)) #50 CAPTURAS E 10 ELEMENTOS
for(i in 1:50){
captura_1cel[i,] = sample(elementos_por_celula1[,3],10, replace = FALSE)
}
captura_1cel #NO INTERIOR O ID DO ELEMENTO

recaptura_1cel1 = array(0,c(50,10))
recaptura_1cel2 = array(0,c(50,10))
até...recaptura_1ce49 = array(0,c(50,10))

for(i in 1:50){
recaptura_1cel1[i,] = sample(elementos_por_celula1[,3],10,replace = FALSE)
recaptura_1cel2[i,] = sample(elementos_por_celula1[,3],10,replace = FALSE)
até...recaptura_1cel49[i,]
}
########################################
# Estimador de Schnabel
########################################
n_1 = ... = n_50 = 10 ; m_1 = rep(0,50); m_2 = ... = m_50 = 10;
u1 = ... = u50 = NULL; M_1 = ... = M_50 = NULL
n_hat_schnabel11 = ... = n_hat_schnabel150 = NULL

#recapturados
for(i in 1:50){
m_2[i] = recapturados1_celula[i,1]
```

```
m_3[i] = recapturados1_celula[i,2]
    até...m_50[i] = recapturados1_celula[i,49]

u1[i] = n_1 - m_1[i]
u2[i] = n_2 - m_2[i]
    até...u50[i] = n_50 - m_50[i]

M_1=0
M_2[i] = u1[i]
M_3[i] = u1[i]+u2[i]
    até...M_50[i]

n_hat_schnabel11[i] = sum(n_2*M_2[i])/sum(m_2[i])
n_hat_schnabel12[i] = sum(n_2*M_2[i],n_3*M_3[i])/sum(m_2[i],m_3[i])
    até...n_hat_schnabel149[i]
}
########################################
# Critério de Parada
########################################
matriz_schnabel = matrix(c(n_hat_schnabel11,...,n_hat_schnabel149),
50,49,byrow = FALSE)
i = 1
média = NULL

média[1]= mean(c(matriz_schnabel[i,2],matriz_schnabel[i,3]))
média[2]= mean(c(matriz_schnabel[i,2],matriz_schnabel[i,3],
matriz_schnabel[i,4]))
até...média[48]

variância = NULL

variância[1]= var(c(matriz_schnabel[i,2],matriz_schnabel[i,3]))
variância[2]= var(c(matriz_schnabel[i,2],matriz_schnabel[i,3],
matriz_schnabel[i,4]))
até...variância[48]
```

```
k = seq(2,50,1)
LS = NULL
LI = NULL
MK = NULL
eta = 0.95

for(i in 1:48){
LS[i] = média[k[i]]+qt(eta,k-1)[i]*(sqrt(variância[k[i]]/k[i]))
LI[i] = média[k[i]]-qt(eta,k-1)[i]*(sqrt(variância[k[i]]/k[i]))
MK[i] = qt(eta,k-1)[i]*(sqrt(variância[k[i]]/k[i]))
}
parada_k* = which(MK <= média[1]*0.05)[1]
estimativa_k* = round(matriz_schnabel[1,parada], digits = 0)
```

A.3 Amostragem Adaptativa por Conglomerados

```
AAC=function(popul,z1,c)
{
indices=array(1:(nrow(popul)*ncol(popul)),c(nrow(popul),ncol(popul)))
d=(nrow(popul)*ncol(popul))  #número total de células na grade
soma=integer(0)
somaj=integer(0)
divisao=integer(0)
alpha=integer(0)
amostrat=sample(indices,z1) #retira uma amostra de tamanho z1
empilha=array(NA,c(length(amostrat),d))
for (l in 1:length(amostrat))
{
amostra=amostrat[l]
referencia=amostra
vizinho=function(x) #varredura nos vizinhos
{
#determina a posição do elemento da referência
lca=array(0,c(length(x),2))
for (i in 1:length(x))
{lca[i,]=which(indices==x[i], arr.ind=TRUE)}
linha=lca[1,1]
coluna=lca[1,2]
cima=integer(0)
baixo=integer(0)
esquerda=integer(0)
direita=integer(0)
#cria vizinhos
if (popul[linha,coluna]>c)
{
if (linha!=1){cima=indices[lca[1,1]-1,lca[1,2]]}
if (linha!=nrow(popul)){baixo=indices[lca[1,1]+1,lca[1,2]]}
if (coluna!=1){esquerda=indices[lca[1,1],lca[1,2]-1]}
if (coluna!=ncol(popul)) {direita=indices[lca[1,1],lca[1,2]+1]}
}
```

```
b=integer(0)
b=cbind(cima,baixo,esquerda,direita)
}
b=vizinho(referencia)
#cria um vetor contendo as células já visitadas
dados=array(NA,c(d,1))
#cria um vetor indicando se os vizinhos daquele elemento
já estão contemplados
explore=array(0,c(d,1))
#coloca o elemento amostrado na matriz de dados
dados[referencia]=popul[which(indices==referencia, arr.ind=TRUE)]
visita=amostra[which(popul[amostra]>c)] # os visitados no passo 1
while (length(visita)>0)
{
referencia=min(visita)
b=vizinho(referencia)
for (i in 1:d)
{if (length(b[which(b==i)])>0) dados[i]=popul[i]}
#indica que os vizinhos da referência já foram explorados
explore[referencia]=1
#elementos a serem explorados
visita=which(dados>c)
explorados=which(explore==1)
comuns=array(0,c(length(visita),length(explorados)))
for (i in 1:length(visita))
for (j in 1:length(explorados))
{if (visita[i]==explorados[j]) comuns[i,j]=i}
visita=visita[-comuns]
}
empilha[l,]=dados
}
#conglomerados distintos
conglomerados=unique(empilha)
#redes distintas
redes=array(NA,c(nrow(conglomerados),ncol(conglomerados)))
```

```
#redes de tamanho 1
for (i in 1:nrow(conglomerados)){
for (j in 1:ncol(conglomerados)){
if ((is.na(conglomerados[i,j])==T|(conglomerados[i,j]<=c)))
redes[i,j]=NA
else redes[i,j]=conglomerados[i,j]
if (length(which(is.na(conglomerados[i,])==F))==1)
redes[i,j]=conglomerados[i,j]
}
}
x=integer(0)
for (i in 1:ncol(conglomerados))
{
x[i]=1-min(is.na(conglomerados[,i]))
}
amtotal=which(x==1)
w=integer(0)

var_HT=0
pi=matrix(0,length(amostrat),length(amostrat))

for (k in 1:nrow(redes))
{
rede=redes[k,which(is.na(redes[k,])==F)]
lr=max(1,length(rede))
soma[k]=sum(rede)
w[k]=mean(rede)
alpha[k]=1-(choose((d-lr),length(amostrat))/choose(d,length(amostrat)))
divisao[k]=soma[k]/alpha[k]
for (j in 1:nrow(redes))
{
lrj=max(1,length(rede))
rede=redes[j,which(is.na(redes[j,])==F)]
somaj[j]=sum(rede)
pi[k,j]= 1-(choose(d-lr,length(amostrat))+choose(d-lrj,length(amostrat))
```

```
-choose(d-lr-lrj,length(amostrat)))/choose(d,length(amostrat))
}
}
for (k in 1:nrow(redes))
{
for (j in 1:nrow(redes))
{
var_HT=var_HT+(1/d^2)*(soma[k]*somaj[j])*(pi[k,j]-pi[k,k]*pi[j,j])
/(pi[k,k]*pi[j,j]*pi[k,j])
}
}
for (k in 1:nrow(redesr))
{
reder=redesr[k,which(is.na(redesr[k,])==F)]
w[k]=mean(reder)
}
HT=sum(divisao)
N = d
HH=N*mean(w)
size=length(amtotal)
result=cbind(HT,var_HT,size,HH)
}
HT = integer(0)
tam = integer(0)
HH = integer(0)
DES = integer(0)
tamDES = integer(0)
junto = 0
c = 0
N =  nrow(read.table("populacao_real.txt"))*ncol(read.table
("populacao_real.txt"))
populacao = matrix(scan("populacao_real.txt"),nrow=40,ncol=40,byrow=T)

for(z1 in c(1/100*N, 3/100*N,5/100*N,10/100*N,15/100*N))
{
```

```
for(r in 1:10000)
{
aacobs = AAC(populacao,z1,c)
HT[r] = aacobs[1,1]
tam[r] = aacobs[1,3]
HH[r] = aacobs[1,4]
print(paste("r=",r,"---","z1=",z1))
}
EHT = mean(HT)
VHT = var(HT)
EHH = mean(HH)
VHH = var(HH)
Eeta = mean(tam)
estatisticas = c(z1,EHT,VHT,EHH,VHH,Eeta)
junto = rbind(junto,estatisticas)
}
write.table(junto,paste("Resultados_N=",N,".txt",sep=""),col.names=FALSE,
row.names=FALSE)
```