



Fundação Instituto Brasileiro de Geografia e Estatística - IBGE

Diretoria de Pesquisas - DPE

Departamento de Emprego e Rendimento - DEREN

Departamento de Metodologia – DEMET

Aspectos de amostragem da Pesquisa de Economia Informal Urbana 97

Rosângela Antunes Pereira Almeida

Zélia Magalhães Bianchini

RIO DE JANEIRO

Junho/98

APRESENTAÇÃO

Este texto apresenta uma descrição dos aspectos relativos à amostra da Pesquisa de Economia Informal Urbana, realizada em outubro de 1997, enfocando os aspectos ligados ao plano amostral, aos procedimentos de seleção, ao acompanhamento da amostra e ao processo de expansão e de estimação da precisão das estimativas.

AGRADECIMENTOS

Agradecemos a contribuição de Half Hussmanns (consultor da OIT), Sonia Albieri e Pedro Luís do Nascimento Silva (ambos do DEMET), pelas sugestões e comentários apresentados.

SUMÁRIO

1. INTRODUÇÃO	7
2. PLANO AMOSTRAL.....	7
2.1 ESTRATIFICAÇÃO DAS UNIDADES PRIMÁRIAS.....	8
2.2 DIMENSIONAMENTO DA AMOSTRA	10
3. SELEÇÃO DA AMOSTRA.....	16
3.1 SELEÇÃO DOS SETORES	16
3.2 SELEÇÃO DOS DOMICÍLIOS	18
4. ACOMPANHAMENTO DA AMOSTRA	23
5. EXPANSÃO DA AMOSTRA E ESTIMAÇÃO DA PRECISÃO DAS ESTIMATIVAS	24
BIBLIOGRAFIA.....	27
ANEXO	29

1. INTRODUÇÃO

A Pesquisa de Economia Informal Urbana de 1997 (ECINF97) é uma pesquisa domiciliar por amostragem que foi conduzida pelo IBGE a nível nacional, visando captar o papel e a dimensão do setor informal na economia brasileira, através da identificação dos proprietários de negócios informais e da investigação das características de funcionamento das unidades produtivas¹. A ECINF97 tem como população objetivo as pessoas residentes na área urbana que trabalhavam por conta própria ou como empregadores com até cinco empregados, em pelo menos uma situação de trabalho, de atividades não-agrícolas (trabalhadores domésticos foram excluídos). As informações foram coletadas tendo como referência o mês de outubro de 1997. Em 1994 foi realizada uma pesquisa piloto no município do Rio de Janeiro que abrangeu todas as etapas previstas para a implantação da pesquisa a nível nacional.

Este trabalho enfoca os aspectos relativos à amostra da ECINF97, em especial, aqueles que diferenciam uma pesquisa do setor informal de uma pesquisa domiciliar tradicional. Essas diferenças surgem pelo fato de se tratar de populações raras, que tendem a ser heterogêneas em virtude dos diferentes tipos de atividade e de suas irregularidades na dispersão da população. Todos esses fatores contribuem para aumentar a complexidade do desenho amostral, incluindo a preparação do cadastro de seleção, a seleção da amostra e os procedimentos de estimação.

Tais aspectos relativos à amostra da ECINF97 encontram-se descritos nos seguintes capítulos:

- capítulo 2 - que aborda os aspectos ligados ao plano amostral, com destaque para a estratificação adotada e o dimensionamento da amostra;
- capítulo 3 - que descreve os procedimentos de seleção;
- capítulo 4 - que trata do acompanhamento da amostra;
- capítulo 5 - que descreve o processo de expansão e estimação da precisão das estimativas.

2. PLANO AMOSTRAL

A ECINF97 foi realizada através uma amostra probabilística de domicílios, obtida em dois estágios de seleção, com estratificação das unidades primárias (setores urbanos) e seleção com probabilidade proporcional ao total de domicílios ocupados existentes na época do Censo Demográfico de 1991 (CD/91), e teve como unidades secundárias os domicílios com moradores ocupados como conta própria ou como empregadores com até cinco empregados. Esses domicílios foram estratificados por grupo de atividade objeto da pesquisa e selecionados com equiprobabilidade em cada estrato.

¹ Ver Jorge (1996).

Na medida em que se pretende obter resultados para cada uma das Unidades da Federação e, também, para as Regiões Metropolitanas de Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Vitória, Rio de Janeiro, São Paulo, Curitiba e Porto Alegre, além do Município de Goiânia, o plano amostral foi aplicado de forma independente para cada uma dessas áreas, que foram definidas como as áreas da pesquisa.

2.1 Estratificação das Unidades Primárias

ESTRATIFICAÇÃO GEOGRÁFICA

Os setores urbanos foram estratificados, primeiramente, por sua condição geográfica, buscando, desta forma, o espalhamento da amostra para garantir a representação das diversas áreas que compõem as áreas da pesquisa. Assim, em cada Unidade da Federação, foram definidos dois ou três estratos geográficos, conforme a existência ou não de Região Metropolitana, assim estabelecidos:

- estrato A - setores urbanos pertencentes ao município da capital;
- estrato B - setores urbanos pertencentes aos demais municípios da Região Metropolitana;
- estrato C - setores urbanos pertencentes aos municípios restantes.

A única exceção ocorreu no Pará, onde não existiu o estrato exclusivo para o município da capital, uma vez que somente dois municípios compõem a Região Metropolitana de Belém.

ESTRATIFICAÇÃO PELA RENDA

A segunda etapa do processo de estratificação das unidades primárias foi realizada dentro de cada estrato geográfico e considerou a média da renda domiciliar de cada setor, convertida em salários mínimos, (que na época era de Cr\$36.161,60), obtida a partir dos resultados definitivos do questionário da amostra do Censo Demográfico de 1991.

A utilização da variável renda na estratificação dos setores objetivava garantir a inclusão na amostra de proprietários do setor informal (conta própria e empregadores com até cinco empregados) provenientes de diversas classes de renda.

Para obtenção dos estratos de renda em cada partição geográfica, utilizou-se a PROC FASTCLUS, do SAS, a qual determina, a partir do número de estratos que se pretende, os conjuntos de setores com maior homogeneidade para a variável definida.

Para a pesquisa foram definidos três estratos de renda com limites diferenciados para cada partição geográfica: renda alta, média e baixa. Entretanto, verificou-se que, para algumas partições geográficas, o número de setores alocados em determinado estrato de

renda era muito pequeno. Optou-se, então, por agregá-lo ao de renda mais próxima. Desta forma, em algumas partições temos apenas dois estratos de renda ou até mesmo um, o que significa que não houve estratificação dos setores pela renda, como é o caso dos estratos referentes à capital de Tocantins e ao resto da Unidade da Federação de Roraima.

O quantitativo de setores em cada estrato, por Unidade da Federação, é apresentado na tabela 1.

Tabela 1 - Número de setores urbanos por estrato, segundo as Unidades da Federação

UNIDADES DA FEDERAÇÃO	TOTAL DE SETORES URBANOS	ESTRATO A - MUNICÍPIO DA CAPITAL				ESTRATO B - RESTO DA REGIÃO METROPOLITANA				ESTRATO C - RESTO DA UNIDADE DA FEDERAÇÃO			
		TOTAL	RENDA BAIXA	RENDA MÉDIA	RENDA ALTA	TOTAL	RENDA BAIXA	RENDA MÉDIA	RENDA ALTA	TOTAL	RENDA BAIXA	RENDA MÉDIA	RENDA ALTA
BRASIL	102.778	32.872	25.063	6.563	1.246	15.097	13.264	1.684	149	54.809	44.959	9.040	810
Rondônia (1)	620	212	162	50	0	0	0	0	0	408	304	104	0
Acre (1)	197	140	101	39	0	0	0	0	0	57	29	28	0
Amazonas (1)	1.073	740	611	106	23	0	0	0	0	333	284	49	0
Roraima (1)	101	83	43	40	0	0	0	0	0	18	0	18	0
Pará	1.847	0	0	0	0	666	499	132	35	1.181	1.012	169	0
Amapá (1)	152	99	54	45	0	0	0	0	0	53	28	25	0
Tocantins (1)	371	13	0	13	0	0	0	0	0	358	274	84	0
Maranhão (1)	1.331	192	154	38	0	0	0	0	0	1.139	960	179	0
Piauí (1)	1.119	476	410	66	0	0	0	0	0	643	449	166	28
Ceará	4.024	1.776	1.470	271	35	413	207	162	44	1.835	1.682	153	0
Rio Grande do Norte (1)	1.352	527	410	117	0	0	0	0	0	825	751	74	0
Paraíba (1)	1.827	421	306	86	29	0	0	0	0	1.406	1.166	218	22
Pernambuco	4.090	1.076	820	207	49	1.062	978	84	0	1.952	1.491	416	45
Alagoas (1)	1.150	504	407	97	0	0	0	0	0	646	447	173	26
Sergipe (1)	840	368	258	75	35	0	0	0	0	472	354	105	13
Bahia	5.710	1.716	1.353	289	74	369	336	33	0	3.625	3.068	505	52
Minas Gerais	10.781	1.976	1.439	406	131	1.089	948	141	0	7.716	6.140	1.347	229
Espírito Santo	1.717	242	152	76	14	626	462	152	12	849	711	138	0
Rio de Janeiro	12.945	6.252	4.863	1.237	152	3.972	3.733	239	0	2.721	2.131	523	67
São Paulo	28.053	9.585	7.230	1.943	412	4.876	4.504	372	0	13.592	11.765	1.811	16
Paraná	5.808	1.377	911	354	112	409	330	79	0	4.022	3.190	751	81
Santa Catarina (1)	3.107	262	166	65	31	0	0	0	0	2.845	2.221	524	100
Rio Grande do Sul	7.967	1.764	1.397	367	0	1.615	1.267	290	58	4.588	3.795	781	12
Mato Grosso do Sul (1)	1.239	420	313	91	16	0	0	0	0	819	620	161	38
Mato Grosso (1)	1.158	300	254	46	0	0	0	0	0	858	670	176	12
Goiás (1)	2.668	820	628	142	50	0	0	0	0	1.848	1.417	362	69
Distrito Federal (2)	1.531	1.531	1.151	297	83	0	0	0	0	0	0	0	0

(1) - Não existe Região Metropolitana na Unidade da Federação. (2) - Município único

2.2 Dimensionamento da Amostra

Para a determinação do tamanho da amostra, de cada área da pesquisa, estabeleceu-se como variável de dimensionamento o *total de proprietários de unidades produtivas do setor informal*, que deveria ser estimado com um erro de amostragem associado à estimativa a ser fixado.

Foi definido, também, que o número de domicílios a serem selecionados por setor, seria mantido constante em todas as áreas da pesquisa. A princípio, fixou-se em 16 domicílios por setor.

O dimensionamento da amostra foi estabelecido com base nos resultados do Censo Demográfico de 1991 (CD/91).

ESTIMADORES UTILIZADOS

A partir do desenho amostral pretendido, um estimador não tendencioso para o **total** de uma característica y qualquer é dado por:

$$\hat{Y} = \sum_{h=1}^L \hat{Y}_h = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{m_h} \frac{N_{hi}}{P_{hi}} \frac{1}{n_{hi}} \sum_{j=1}^{n_{hi}} y_{hij}$$

onde:

L é o número de estratos da área da pesquisa;

\hat{Y}_h é o total estimado da variável y no estrato h ;

m_h é o número de setores da amostra no estrato h ;

N_{hi} é o número de domicílios com proprietários do setor informal no setor i do estrato h ;

P_{hi} é a probabilidade de seleção, num sorteio, do setor i do estrato h ;

$$P_{hi} = \frac{D_{hi}}{D_h}$$

D_{hi} é o total de domicílios ocupados existentes no setor i do estrato h , obtido pelo CD/91;

$$D_h = \sum_{i=1}^{M_h} D_{hi}$$

M_h é o número de setores no universo do estrato h ;

n_{hi} é o número de domicílios com proprietários do setor informal a serem selecionados no setor i do estrato h ;

y_{hij} é o valor da variável y no j -ésimo domicílio selecionado do setor i do estrato h .

A **variância** para este estimador de total é dada por:

$$V(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} P_{hi} \left[\frac{Y_{hi}}{P_{hi}} - Y_h \right]^2 + \sum_{i=1}^{M_h} \frac{N_{hi}}{P_{hi}} \left[\frac{N_{hi} - n_{hi}}{n_{hi}} \right] S_{hi}^2 \right\}$$

onde:

Y_{hi} é o total da variável y no setor i do estrato h ;

Y_h é o total da variável y no estrato h ;

S_{hi}^2 é a variância entre os domicílios da variável y no setor i do estrato h ;

$$S_{hi}^2 = \frac{1}{N_{hi} - 1} \sum_{j=1}^{N_{hi}} (Y_{hij} - \bar{Y}_{hi})^2$$

Y_{hij} é o valor da variável y no domicílio j do setor i do estrato h ;

\bar{Y}_{hi} é a média entre os domicílios da variável y no setor i do estrato h .

Para o cálculo da **variância** necessitava-se dos dados populacionais para a característica de interesse, que no caso era o número de proprietários de unidades produtivas do setor informal. Porém, os dados mais recentes disponíveis para tal variável eram os obtidos na amostra do Censo Demográfico de 1991. Utilizou-se, então o seguinte estimador consistente da variância, cuja justificativa encontra-se em anexo.

$$v(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} P_{hi} \left[\frac{\hat{Y}'_{hi}}{P_{hi}} - \hat{Y}'_h \right]^2 - \sum_{i=1}^{M_h} \left[\frac{1}{P_{hi}} - 1 \right] N_{hi}^2 \left[1 - \frac{d_{hi}}{N_{hi}} \right] \frac{s_{hi}^2}{d_{hi}} \right\} +$$

$$+ \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{\hat{N}_{hi}^2}{P_{hi}} \left[1 - \frac{n_{hi}}{\hat{N}_{hi}} \right] \frac{s_{hi}^{\prime 2}}{n_{hi}}$$

onde:

\hat{Y}'_{hi} é o estimador do total da variável y , obtido a partir da amostra do CD/91, no setor i do estrato h ;

$$\hat{Y}'_{hi} = \frac{D_{hi}}{d_{hi}} \sum_{j=1}^{d_{hi}} y'_{hij}$$

$$\hat{Y}'_h = \sum_{i=1}^{M_h} \hat{Y}'_{hi}$$

y'_{hij} é o valor da variável y no j -ésimo domicílio do setor i do estrato h na amostra do CD/91;

d_{hi} é o número de domicílios ocupados na amostra do CD/91, no setor i do estrato h ;

s_{hi}^2 é o estimador da variância da variável y no setor i do estrato h , obtido a partir da amostra do CD/91;

$$s_{hi}^2 = \frac{1}{d_{hi} - 1} \sum_{j=1}^{d_{hi}} (y'_{hij} - \bar{y}_{hi})^2$$

$$\bar{y}_{hi} = \frac{1}{d_{hi}} \sum_{j=1}^{d_{hi}} y'_{hij}$$

\hat{N}_{hi} é a estimativa de domicílios com proprietários de unidades produtivas do setor informal, obtida a partir da amostra do CD/91, no setor i do estrato h ;

$$\hat{N}_{hi} = D_{hi} \frac{n'_{hi}}{d_{hi}}$$

n'_{hi} é o número de domicílios com proprietários de unidades produtivas do setor informal, na amostra do CD/91, no setor i do estrato h ;

$s'_{hi}{}^2$ é o estimador da variância entre os domicílios com proprietários do setor informal, obtido pela amostra do CD/91, no setor i do estrato h ;

$$s'_{hi}{}^2 = \frac{1}{n'_{hi} - 1} \sum_{j=1}^{n'_{hi}} (y'_{hij} - \bar{y}'_{hi})^2$$

$$\bar{y}'_{hi} = \frac{1}{n'_{hi}} \sum_{j=1}^{n'_{hi}} y'_{hij}$$

Para determinação do número de setores a serem selecionados por área da pesquisa, foi usado o estimador $v(\hat{Y})$ e consideradas ainda as duas condições citadas a seguir:

- o número de domicílios com proprietários do setor informal a serem selecionados deveria ser constante em todos os setores da amostra:

$$n_{hi} = \bar{n} \quad (\forall h, \forall i)$$

- o número de setores a serem selecionados em cada estrato deveria ser proporcional ao número de domicílios com proprietários do setor informal existentes no estrato:

$$m_h = m \frac{\hat{N}_h}{\hat{N}} \quad \text{sendo} \quad m = \sum_{h=1}^L m_h \quad \text{e} \quad \hat{N} = \sum_{h=1}^L \hat{N}_h = \sum_{h=1}^L \sum_{i=1}^{M_h} \hat{N}_{hi} .$$

Além de substituir esses valores em $v(\hat{Y})$ foi considerado, para a determinação de m (o número de setores a serem selecionados em cada área da pesquisa), o estimador do coeficiente de variação associado ao estimador de total da variável y , $cv(\hat{Y})$, definido por:

$$cv(\hat{Y}) = \frac{\sqrt{v(\hat{Y})}}{\hat{Y}'} \quad \text{onde:} \quad \hat{Y}' = \sum_{h=1}^L \sum_{i=1}^{M_h} \frac{D_{hi}}{d_{hi}} \sum_{j=1}^{d_{hi}} y'_{hij} .$$

DETERMINAÇÃO DO NÚMERO DE SETORES NA AMOSTRA

Fixando-se os valores para o número de domicílios a serem selecionados por setor (\bar{n}) e para o $cv(\hat{Y})$, sendo y o número de proprietários de unidades produtivas do setor informal da área da pesquisa, o total de setores a serem selecionados por área foi determinado pela seguinte expressão:

$$m = \frac{\hat{N}}{(\hat{Y}' cv(\hat{Y}))^2} \sum_{h=1}^L \frac{1}{\hat{N}_h} \left\{ \sum_{i=1}^{M_h} P_{hi} \left[\frac{\hat{Y}'_{hi}}{P_{hi}} - \hat{Y}'_h \right]^2 - \sum_{i=1}^{M_h} \left[\frac{1}{P_{hi}} - 1 \right] D_{hi}^2 \left[1 - \frac{d_{hi}}{D_{hi}} \right] \frac{s_{hi}^2}{d_{hi}} + \sum_{i=1}^{M_h} \frac{\hat{N}_{hi}^2}{P_{hi}} \left[1 - \frac{\bar{n}}{\hat{N}_{hi}} \right] \frac{s_{hi}'^2}{\bar{n}} \right\}$$

Para se definir o número de setores na amostra em cada área de pesquisa, foram avaliadas alternativas quanto ao número de domicílios a serem selecionados por setor: 16, 20 e 24. Além disso, investigou-se o efeito de utilizar o coeficiente de variação de 5% ou 6% para estimar o número de proprietários do setor informal em cada área. O resultado destes estudos é dado na tabela 2 a seguir.

Tabela 2 - Número de setores a serem selecionados, segundo o nível de precisão desejado e o número de domicílios selecionados por setor, por área da pesquisa

ÁREAS DA PESQUISA	CV=5%			CV=6%
	16 DOMICÍLIOS POR SETOR	20 DOMICÍLIOS POR SETOR	24 DOMICÍLIOS POR SETOR	16 DOMICÍLIOS POR SETOR
Rondônia	66	65	64	46
Acre	71	71	70	50
Amazonas	83	82	82	58
Roraima	47	47	46	33
Pará (1)	92	91	90	64
RM de Belém	59	57	57	41
Amapá	68	66	66	47
Tocantins	52	51	51	36
Maranhão	72	71	71	50
Piauí	74	74	73	52
Ceará (1)	99	97	97	69
RM de Fortaleza	70	69	69	49
Rio Grande do Norte	86	85	84	59
Paraíba	104	103	102	72
Pernambuco (1)	123	122	121	85
RM de Recife	83	82	81	57
Alagoas	113	112	111	78
Sergipe	96	95	95	67
Bahia (1)	131	130	129	91
RM de Salvador	90	89	88	62
Minas Gerais (1)	110	108	108	76
RM de Belo Horizonte	70	69	68	49
Espírito Santo (1)	94	93	92	65
RM de Vitória	72	72	71	50
Rio de Janeiro (1)	98	97	96	68
RM do Rio de Janeiro	87	86	86	40
São Paulo (1)	109	108	107	75
RM de São Paulo	82	81	80	57
Paraná (1)	107	106	105	75
RM de Curitiba	71	70	70	49
Santa Catarina	80	80	79	56
Rio Grande do Sul (1)	99	98	97	69
RM de Porto Alegre	70	70	69	49
Mato Grosso do Sul	54	53	53	38
Mato Grosso	62	61	60	43
Goiás (2)	95	93	92	66
Goiânia	60	60	59	42
Distrito Federal	93	92	92	65

(1): O número de setores a serem selecionados na Unidade da Federação inclui o da Região Metropolitana.

(2): O número de setores a serem selecionados na Unidade da Federação inclui o do município de Goiânia.

Pode-se observar que para o nível de precisão de 5%, a variação no número de domicílios a serem selecionados por setor produz alterações insignificantes, donde se pode concluir que a maior parte da variância da estimativa do total de proprietários do setor informal vem da variação entre os setores. Já a variação no número de setores, quando se altera de 5% para 6% o nível de precisão da estimativa, é bastante significativa.

A decisão sobre que número de setores selecionar baseou-se no fator custo da investigação. Assim, optou-se pelo número de setores em cada área obtido através do

cálculo que fixou em 16 o número de domicílios a serem selecionados por setor e o coeficiente de variação de 5%. Excepcionalmente, por motivo de custo, para as áreas da Região Norte a opção foi pelo número de setores obtido com o coeficiente de variação de 6%. Tais opções resultaram numa amostra de 2.233 setores e um número esperado de 35.728 entrevistas na amostra.

Com o intuito de avaliar o ganho da estratificação dos setores pela variável renda para estimar o número de proprietários de unidades produtivas do setor informal, foi calculado o número de setores na amostra utilizando-se somente a estratificação geográfica para $\bar{n} = 16$ e coeficientes de variação fixados em 6% para as áreas da Região Norte e 5% para as demais áreas da pesquisa.

Desses cálculos resultou em 2.296 o número de setores da amostra, o que implicaria em aumentar, para todo o País, 63 setores na amostra e 1.008 entrevistas em relação ao plano amostral que considerou a estratificação dos setores pela variável renda. Cabe registrar que para a maioria das áreas, a diferença no total de setores da amostra foi de apenas 1 ou 2 setores, e no caso de São Paulo e Rio Grande do Sul não houve diferença. Porém, no Distrito Federal a diferença foi de 14 setores (15%), implicando no aumento de 224 entrevistas; no Ceará e na Bahia a diferença foi de 6 setores.

O ganho da estratificação dos setores pela renda não é tão expressivo para estimar o número de proprietários de unidades produtivas do setor informal, para a maioria das áreas e a contribuição para a redução da amostra é de cerca de 1.000 entrevistas. Porém, optou-se por essa estratificação para assegurar a inclusão na amostra de setores provenientes de diversas classes de renda média.

ALOCAÇÃO DA AMOSTRA DE SETORES NOS ESTRATOS

Conhecido o tamanho da amostra de setores, m , foi feita a alocação por estrato de cada área da pesquisa.

A distribuição dos setores entre os estratos foi feita proporcionalmente à estimativa do total de domicílios com proprietários do setor informal existentes no estrato, segundo o CD/91. Para se obter um número inteiro de setores a serem selecionados procedeu-se ao seguinte processo de arredondamento.

$$m_h = \text{parte inteira de} \left(m \cdot \frac{\hat{N}_h}{\hat{N}} \right) + 1$$

Além disso, estabeleceu-se que todo estrato teria, minimamente, 2 setores selecionados, $m_h = 2$.

Conseqüentemente, em função destes procedimentos, o número final de setores a serem selecionados por área de pesquisa sofreu pequenos ajustes, resultando em 2.340

setores e 37.440 entrevistas esperadas, significando um acréscimo de 107 setores em relação ao número determinado através do cálculo da variância.

Nas Unidades da Federação (UF) que têm Região Metropolitana (RM), foi determinado o número de setores da amostra para a UF e para a RM, independentemente. A partir do tamanho da amostra da RM foi feita a alocação dos setores, de forma proporcional ao total de domicílios com proprietários do setor informal, nos estratos de renda do município da capital (estrato A) e nos demais municípios da RM (estrato B). O número de setores da amostra para o conjunto dos municípios de fora da RM (estrato C) foi obtido pela diferença entre o número de setores da UF e o número de setores da RM. A partir daí foi feita a alocação por estrato de renda proporcional à estimativa do total de proprietários do setor informal existentes no estrato C, segundo o CD/91.

Na tabela 3, mais adiante, consta o número de setores selecionados e o número esperado de domicílios com proprietários do setor informal na amostra, por área da pesquisa.

3. SELEÇÃO DA AMOSTRA

3.1 Seleção dos Setores

A seleção dos setores foi feita, de forma sistemática, com probabilidade proporcional ao total de domicílios ocupados existentes na época do Censo Demográfico de 1991, dentro de cada estrato.

No processo de seleção dos setores estabeleceu-se que não haveria coincidências com os setores selecionados na Pesquisa Nacional por Amostra de Domicílios - PNAD e na Pesquisa Mensal de Emprego - PME, ambas pesquisas domiciliares que estariam em fase de entrevista na mesma época que a ECINF97. A possibilidade de 2 ou até 3 pesquisas domiciliares sendo realizadas em um mesmo setor ou, pior, em um mesmo domicílio poderia provocar uma elevação na taxa de não entrevistados devida à recusa do informante.

O total de setores selecionados, o número esperado de domicílios na amostra, o número de municípios com setores na amostra, bem como a média de setores selecionados por município da amostra constam da tabela 3 apresentada a seguir. Houve um grande espalhamento da amostra, que se pode verificar pela média de setores selecionados por município, causado pelo modelo em dois estágios (setores e domicílios) empregado, isto é, sem a seleção prévia de municípios. A opção pela não inclusão do estágio de seleção de municípios deveu-se ao aumento no número de setores a serem selecionados que teria provocado, o que não seria conveniente devido à complexidade do processo de listagem dos domicílios que é usado pela pesquisa, como se verá mais adiante.

Tabela 3 - Total de setores selecionados, de domicílios esperados na amostra, número de municípios com setores na amostra e média de setores selecionados por município, por área de pesquisa

ÁREAS DE PESQUISA	SETORES SELECIONADOS	DOMICÍLIOS ESPERADOS NA AMOSTRA	MUNICÍPIOS COM SETORES NA AMOSTRA	MÉDIA DE SETORES SELECIONADOS POR MUNICÍPIO
BRASIL	2.340	37.440	753	3,1
RONDÔNIA	48	768	14	3,4
ACRE	52	832	8	6,5
AMAZONAS	61	976	14	4,4
RORAIMA	34	544	3	11,3
PARÁ (1)	68	1.088	22	3,1
RM DE BELÉM	43	688	2	21,5
AMAPÁ	51	816	7	7,3
TOCANTINS	38	608	17	2,2
MARANHÃO	75	1.200	45	1,7
PIAUÍ	76	1.216	37	2,1
CEARÁ (1)	105	1.680	31	3,4
RM DE FORTALEZA	75	1.200	6	12,5
RIO GRANDE DO NORTE	88	1.408	42	2,1
PARAÍBA	108	1.728	50	2,2
PERNAMBUCO (1)	129	2.064	42	3,1
RM DE RECIFE	86	1.376	10	8,6
ALAGOAS	116	1.856	34	3,4
SERGIPE	100	1.600	34	2,9
BAHIA (1)	138	2.208	37	3,7
RM DE SALVADOR	93	1.488	7	13,3
MINAS GERAIS (1)	114	1.824	44	2,6
RM DE BELO HORIZONTE	72	1.152	10	7,2
ESPÍRITO SANTO (1)	101	1.616	22	4,6
RM DE VITÓRIA	78	1.248	5	15,6
RIO DE JANEIRO (1)	103	1.648	17	6,1
RM DO RIO DE JANEIRO	90	1.440	9	10,0
SÃO PAULO (1)	115	1.840	40	2,9
RM DE SÃO PAULO	84	1.344	16	5,3
PARANÁ (1)	113	1.808	39	2,9
RM DE CURITIBA	73	1.168	8	9,1
SANTA CATARINA	84	1.344	48	1,8
RIO GRANDE DO SUL (1)	106	1.696	35	3,0
RM DE PORTO ALEGRE	74	1.184	13	5,7
MATO GROSSO DO SUL	58	928	23	2,5
MATO GROSSO	65	1.040	25	2,6
GOIÁS (2)	99	1.584	22	4,5
GOIÂNIA	62	992	1	62,0
DISTRITO FEDERAL	95	1.520	1	95,0

(1): Os valores apresentados para a Unidade da Federação incluem os da Região Metropolitana.

(2): Os valores apresentados para a Unidade da Federação incluem os do município de Goiânia.

3.2 Seleção dos Domicílios

Uma abordagem particularmente especial no caso da ECINF diz respeito à preparação do cadastro da população objetivo da pesquisa para a seleção dos domicílios, isto é, os domicílios ocupados com algum morador proprietário de unidades produtivas do setor informal (conta própria ou empregador com até cinco empregados em pelo menos uma situação de trabalho). Além disso, era necessário garantir a presença na amostra de proprietários de unidades econômicas do setor informal de cada grupo de atividade objeto da pesquisa:

- grupo 1 - indústria da transformação e extrativa mineral;
- grupo 2 - indústria da construção;
- grupo 3 - comércio de mercadorias;
- grupo 4 - serviços de alojamento e alimentação;
- grupo 5 - serviços de transporte;
- grupo 6 - serviços de reparação, pessoais, domiciliares e de diversões;
- grupo 7 - serviços técnicos e auxiliares; e
- grupo 8 - outros serviços.

De acordo com Kalton e Anderson (1986) a varredura (“screening”) é uma das estratégias a ser considerada para a investigação de populações raras. Quando a população objetivo de uma pesquisa for “rara” e estiver espalhada pelos conglomerados, uma estratégia eficiente para investigar esse tipo de população, deve considerar a varredura completa dos conglomerados da amostra, para identificação da população objetivo e construção do cadastro de seleção das unidades de último estágio, apenas com as unidades identificadas como parte da população objetivo da pesquisa.

A varredura dos setores da amostra da ECINF foi feita através da operação de *listagem*. Neste aspecto a ECINF se diferencia de uma pesquisa domiciliar tradicional, cujo objetivo da *listagem* dos setores selecionados é apenas produzir uma lista completa dos endereços das unidades domiciliares, para a seleção dos domicílios. Na ECINF a operação de *listagem* teve um custo muito mais elevado, pois além de produzir a lista de endereços dos domicílios, envolveu a realização de uma entrevista em cada domicílio, para identificar as atividades econômicas desenvolvidas pelos moradores.

Para garantir a presença na amostra de proprietários de unidades econômicas do setor informal dos diversos grupos de atividades, foi decidido introduzir uma estratificação secundária dos domicílios da população objetivo por grupo de atividade, a partir dos resultados da *listagem*.

Por estas razões, a *listagem* da ECINF consistiu numa varredura dos setores da amostra com uma pequena entrevista para coletar basicamente as seguintes informações:

- quais moradores do domicílio, de 10 anos ou mais de idade, trabalhavam no período de referência?
- entre os moradores ocupados, quais eram proprietários de unidades econômicas do setor informal, em pelo menos uma situação de trabalho?
- quais as atividades que esses proprietários do setor informal desenvolviam?

As informações coletadas na *listagem* da ECINF nos setores selecionados foram utilizadas com os seguintes objetivos:

- identificar e produzir a lista completa dos endereços das unidades domiciliares;
- identificar os domicílios ocupados com algum morador proprietário de unidades produtivas do setor informal;
- gerar um cadastro dos domicílios com proprietários de unidades produtivas do setor informal, onde se identificavam os endereços e as atividades econômicas desenvolvidas por seus moradores;
- associar cada domicílio com proprietários do setor informal a cada estrato definido por grupo de atividade.

Reconhecendo que num mesmo domicílio podem morar proprietários do setor informal que desenvolvem atividades separadas e distintas, e, até mesmo, um único morador exercer mais de uma atividade do setor informal, foi necessário classificar cada domicílio em apenas um estrato de grupo de atividade, para fazer a seleção da amostra de domicílios. A atividade do domicílio foi escolhida entre aquelas desenvolvidas por seus moradores, proprietários do setor informal, de acordo com uma ordenação de prioridade estabelecida para os grupos de atividades, cujo objetivo era dar chance de seleção aos grupos de atividades mais rarefeitos, caso contrário, sabia-se que as atividades de prestação de serviços e comércio, que são as mais frequentes entre as pessoas ocupadas e, em especial no caso dos conta própria e pequenos empregadores, prevaleceriam. Desse modo, o domicílio foi alocado ao estrato por determinada atividade, embora na entrevista tenham sido consideradas todas as atividades exercidas por seus moradores.

A seleção dos domicílios estava limitada àqueles do cadastro dos domicílios com proprietários de unidades produtivas do setor informal, gerado com os resultados da *listagem*. Portanto, os domicílios em que não foram identificadas atividades do setor informal não tiveram chance de seleção. Por isso, tanto a má identificação de proprietários do setor informal na listagem, bem como a defasagem entre o período de tempo da listagem e a entrevista puderam gerar problemas de cobertura e de desatualização do cadastro de seleção. Quanto mais próxima fosse a entrevista da listagem, menor seria a chance de ocorrerem trocas dos moradores dos domicílios ou mudanças de atividades econômicas dos moradores. Isto porque no setor informal a pessoa muda de atividade com mais frequência que nas relações de trabalho formais. Como consequência imediata dos problemas na desatualização do cadastro têm-se a não realização da entrevista, a redução da amostra e a perda na precisão das estimativas e na qualidade das informações.

A defasagem entre a listagem e a entrevista foi de cerca de 2 meses. Pois, além da operação de campo da listagem ser mais complexa, implicou na apuração das informações coletadas para a seleção dos domicílios.

A alocação dos 16 domicílios a serem selecionados por setor foi feita proporcionalmente entre os grupos de atividades existentes no setor, ou seja:

$$n_{hij} = \frac{N_{hij}^*}{N_{hi}^*} * 16$$

onde:

n_{hij} é o total de domicílios com proprietários do setor informal a serem selecionados no grupo de atividade j no setor i do estrato h ;

N_{hij}^* é o total de domicílios classificados como pertencentes ao grupo de atividade j no setor i do estrato h , obtido pela listagem;

$$N_{hi}^* = \sum_{j=1}^8 N_{hij}^*$$

Após a distribuição proporcional do total de domicílios a serem selecionados pelos grupos de atividades, foram feitos os seguintes ajustes:

$$n_{hij} = \text{parte inteira de } (n_{hij} + 0,5)$$

$$n_{hij} \geq 2 \quad \text{onde } N_{hij}^* \geq 2$$

$$n_{hij} = N_{hij}^* \quad \text{onde } N_{hij}^* < 2$$

Esses procedimentos de correção do total de domicílios selecionados por grupo de atividade em cada setor provocou um crescimento médio de cerca de 30% em cada área de pesquisa, o que favorece o levantamento na medida em que esse acréscimo pode ser usado para cobrir as eventuais entrevistas não realizadas.

De posse do total de domicílios a serem selecionados por grupo de atividade em cada setor da amostra, procedeu-se à seleção sistemática dos domicílios, independentemente por grupo de atividade.

As tabelas 4 e 5, a seguir, apresentam a quantidade de domicílios com proprietários do setor informal listados e a distribuição final da amostra pelos grupos de atividade, respectivamente.

Cabe registrar que foram selecionados 2 setores, um no Maranhão e outro em São Paulo, em que não foram encontradas unidades produtivas para serem selecionadas.

Tabela 4 - Domicílios com proprietários do setor informal listados, total e por grupo de atividade, segundo as áreas de pesquisa

ÁREAS DA PESQUISA	DOMICÍLIOS COM PROPRIETÁRIOS DO SETOR INFORMAL LISTADOS								
	TOTAL	GRUPO 1	GRUPO 2	GRUPO 3	GRUPO 4	GRUPO 5	GRUPO 6	GRUPO 7	GRUPO 8
BRASIL	297.696	30.576	47.123	80.436	22.912	18.368	68.602	26.982	2.697
RONDÔNIA	5.947	735	1.073	1.448	461	468	1.240	475	47
ACRE	6.964	603	1.198	1.882	502	435	1.641	400	303
AMAZONAS	7.505	846	1.341	2.515	529	463	1.328	271	212
RORAIMA	3.698	325	608	1.192	289	224	842	178	40
PARÁ (1)	10.132	985	1.174	2.898	1.241	502	2.171	1.074	87
RM DE BELÉM	6.560	636	782	1.793	839	277	1.254	939	40
AMAPÁ	6.971	417	1.097	2.087	577	551	1.875	316	51
TOCANTINS	6.463	756	1.050	1.688	575	447	1.351	562	34
MARANHÃO	13.540	1.561	2.025	4.086	927	941	3.233	635	132
PIAUI	9.932	1.252	1.905	2.883	749	502	1.971	544	126
CEARÁ (1)	12.248	1.890	1.346	3.698	899	644	2.718	936	117
RM DE FORTALEZA	9.156	1.305	972	2.683	768	458	2.024	868	78
RIO GRANDE DO NORTE	11.516	1.119	1.749	3.584	866	765	2.774	617	42
PARÁIBA	13.961	1.754	1.652	4.572	1.139	736	2.957	1.104	47
PERNAMBUCO (1)	16.702	1.734	2.052	4.944	1.130	975	3.997	1.686	184
RM DE RECIFE	11.432	864	1.487	3.193	812	631	2.934	1.374	137
ALAGOAS	13.593	1.267	1.746	4.370	1.012	966	3.160	957	115
SERGIPE	12.332	1.431	1.689	3.405	1.232	939	2.690	858	88
BAHIA (1)	18.459	1.309	2.829	4.738	1.898	1.107	4.765	1.722	91
RM DE SALVADOR	12.271	843	1.947	2.979	1.222	691	3.236	1.297	56
MINAS GERAIS (1)	13.342	1.827	2.130	3.371	927	1.006	2.409	1.527	145
RM DE BELO HORIZONTE	8.618	1.123	1.332	2.186	584	675	1.528	1.069	121
ESPÍRITO SANTO (1)	13.784	1.263	2.721	3.118	1.243	665	2.730	2.008	36
RM DE VITÓRIA	10.797	810	2.083	2.407	1.003	529	2.223	1.711	31
RIO DE JANEIRO (1)	10.934	663	1.905	2.794	874	570	2.667	1.394	67
RM DO RIO DE JANEIRO	9.374	480	1.627	2.442	750	497	2.323	1.203	52
SÃO PAULO (1)	11.819	983	1.992	2.670	974	829	2.928	1.312	131
RM DE SÃO PAULO	8.326	681	1.319	1.890	675	588	2.161	925	87
PARANÁ (1)	15.193	1.419	3.211	3.410	793	971	3.735	1.499	155
RM DE CURITIBA	9.848	918	2.038	2.162	498	614	2.553	989	76
SANTA CATARINA	10.768	1.144	1.754	2.399	521	625	3.206	983	136
RIO GRANDE DO SUL (1)	10.804	839	1.666	2.568	546	586	3.006	1.551	42
RM DE PORTO ALEGRE	7.155	566	997	1.703	375	364	2.008	1.115	27
MATO GROSSO DO SUL	9.422	575	2.112	2.484	568	447	2.186	986	64
MATO GROSSO	11.331	1.523	2.212	2.564	813	803	2.400	893	123
GOIÁS (2)	13.031	1.719	1.923	3.217	866	760	2.993	1.485	68
GOIÂNIA	8.038	1.276	960	2.036	490	386	1.739	1.122	29
DISTRITO FEDERAL	7.305	637	963	1.851	761	441	1.629	1.009	14

(1): o total de domicílios listados apresentados na Unidade da Federação inclui o da Região Metropolitana.

(2): o total de domicílios listados apresentados na Unidade da Federação inclui o do município de Goiânia.

Tabela 5 - Domicílios selecionados, total e por grupo de atividade, segundo as áreas de pesquisa

ÁREAS DA PESQUISA	DOMICÍLIOS SELECIONADOS								
	TOTAL	GRUPO 1	GRUPO 2	GRUPO 3	GRUPO 4	GRUPO 5	GRUPO 6	GRUPO 7	GRUPO 8
BRASIL	48.934	5.394	6.665	11.264	4.761	4.458	9.486	5.742	1.164
RONDÔNIA	990	123	150	217	101	97	180	93	29
ACRE	1.103	104	171	254	106	101	213	103	51
AMAZONAS	1.283	136	192	360	121	120	204	102	48
RORAIMA	731	72	105	192	68	62	142	66	24
PARÁ (1)	1.422	149	160	343	175	128	264	167	36
RM DE BELÉM	892	92	100	207	113	81	147	129	23
AMAPÁ	1.091	99	151	276	103	107	231	89	35
TOCANTINS	796	98	115	177	81	77	144	85	19
MARANHÃO	1.569	191	200	404	141	148	291	137	57
PIAUI	1.634	206	264	380	161	144	275	145	59
CEARÁ (1)	2.191	319	233	551	216	195	400	217	60
RM DE FORTALEZA	1.574	213	160	382	171	138	286	184	40
RIO GRANDE DO NORTE	1.853	206	241	486	183	171	368	171	27
PARAÍBA	2.253	280	252	616	223	202	408	240	32
PERNAMBUCO (1)	2.737	289	323	690	257	237	536	325	80
RM DE RECIFE	1.829	164	221	426	179	155	383	242	59
ALAGOAS	2.371	247	280	653	229	233	460	226	43
SERGIPE	2.092	253	270	491	236	196	387	207	52
BAHIA (1)	2.862	272	379	620	318	271	607	352	43
RM DE SALVADOR	1.946	176	260	396	218	185	427	257	27
MINAS GERAIS (1)	2.372	298	356	518	226	231	376	311	56
RM DE BELO HORIZONTE	1.496	186	218	324	142	143	239	207	37
ESPIRITO SANTO (1)	2.107	224	352	418	212	186	365	322	28
RM DE VITÓRIA	1.633	154	266	323	168	146	287	266	23
RIO DE JANEIRO (1)	2.155	190	321	462	217	187	447	293	38
RM DO RIO DE JANEIRO	1.886	157	283	409	192	163	400	252	30
SÃO PAULO (1)	2.329	224	299	491	212	212	506	320	65
RM DE SÃO PAULO	1.708	165	201	360	152	154	389	240	47
PARANÁ (1)	2.386	256	398	466	200	216	481	310	59
RM DE CURITIBA	1.542	161	241	297	125	134	334	214	36
SANTA CATARINA	1.795	200	241	331	151	160	444	210	58
RIO GRANDE DO SUL (1)	2.244	221	296	454	183	191	500	367	32
RM DE PORTO ALEGRE	1.575	153	191	315	129	132	359	275	21
MATO GROSSO DO SUL	1.207	108	206	276	103	106	222	157	29
MATO GROSSO	1.393	182	228	277	133	127	252	142	52
GOIÁS (2)	2.079	269	262	438	195	188	408	280	39
GOIÂNIA	1.299	188	147	275	119	112	239	199	20
DISTRITO FEDERAL	1.889	178	220	423	210	165	375	305	13

(1): o total de domicílios selecionados apresentados na Unidade da Federação inclui o da Região Metropolitana.

(2): o total de domicílios selecionados apresentados na Unidade da Federação inclui o do município de Goiânia.

4. ACOMPANHAMENTO DA AMOSTRA

A situação verificada na coleta das entrevistas nem sempre corresponde ao que era esperado com base nos resultados da listagem, ou seja, nem todos os domicílios selecionados são, efetivamente, entrevistados.

Embora a população objetivo tenha sido composta, exclusivamente, pelos domicílios com proprietários do setor informal, devido à defasagem entre a época da listagem e a das entrevistas (cerca de dois meses), alguns domicílios mudaram a sua condição de ser ou não objeto de pesquisa (por terem perdido seus moradores proprietários do setor informal, porque mudaram de residência ou deixaram de pertencer a essa categoria, ou, até mesmo, por erro de identificação ocorrido durante a listagem). Existe, também, a possibilidade de um domicílio, que na época da listagem estava ocupado, tornar-se um domicílio vago, ter sido demolido, ou deixar de ser um domicílio particular para se transformar numa unidade não residencial. Além disso, a entrevista pode não ser realizada por não ser possível encontrar os moradores (domicílios fechados), ou por recusa

No caso de se ter realizado a entrevista num determinado domicílio, duas situações podem ter ocorrido: havia moradores proprietários do setor informal (situação A1) ou não (situação A2). Todos os casos restantes são não-entrevistas (situação B). A tabela 6 apresenta a situação final dos domicílios selecionados quanto à realização das entrevistas. Como se pode observar, o total de entrevistas realizadas do tipo A1, entrevistas realizadas em domicílios com proprietários do setor informal, ficou, na maioria das Unidades da Federação, próximo do número esperado de entrevistas apresentado na tabela 3. A taxa de resposta, definida pela razão entre o número de entrevistas do tipo A1 pelo número de domicílios selecionados, foi de 75,63% a nível nacional. Portanto, a taxa de não-resposta foi compensada pelo aumento de cerca de 30% no total de domicílios selecionados, provocado pelo ajuste do processo de seleção, porém com variabilidade entre as áreas; como se pode notar Roraima teve a maior taxa de não-resposta (37,89%) e Sergipe a menor (16,11%).

Tabela 6 - Situação de entrevista dos domicílios selecionados, por Unidade da Federação

UNIDADES DA FEDERAÇÃO	DOMICÍLIOS SELECIONADOS						
	TOTAL	ENTREVISTA REALIZADA				ENTREVISTA NÃO REALIZADA	
		A1		A2		B	
		TOTAL	%	TOTAL	%	TOTAL	%
BRASIL	48.934	37.010	75,63	8.815	18,01	3.109	6,35
RONDÔNIA	990	692	69,90	232	23,43	66	6,67
ACRE	1.103	814	73,80	208	18,86	81	7,34
AMAZONAS	1.283	919	71,63	244	19,02	120	9,35
RORAIMA	731	454	62,11	107	14,64	170	23,26
PARÁ	1.422	1.170	82,28	180	12,66	72	5,06
AMAPÁ	1.091	763	69,94	238	21,81	90	8,25
TOCANTINS	796	659	82,79	101	12,69	36	4,52
MARANHÃO	1.569	1.120	71,38	336	21,41	113	7,20
PIAUÍ	1.634	1.279	78,27	242	14,81	113	6,92
CEARÁ	2.191	1.744	79,60	342	15,61	105	4,79
RIO GRANDE DO NORTE	1.853	1.464	79,01	326	17,59	63	3,40
PARAÍBA	2.253	1.782	79,09	384	17,04	87	3,86
PERNAMBUCO	2.737	2.033	74,28	545	19,91	159	5,81
ALAGOAS	2.371	1.744	73,56	456	19,23	171	7,21
SERGIPE	2.092	1.755	83,89	229	10,95	108	5,16
BAHIA	2.862	2.213	77,32	419	14,64	230	8,04
MINAS GERAIS	2.372	1.856	78,25	381	16,06	135	5,69
ESPÍRITO SANTO	2.107	1.541	73,14	443	21,03	123	5,84
RIO DE JANEIRO	2.155	1.513	70,21	508	23,57	134	6,22
SÃO PAULO	2.329	1.668	71,62	467	20,05	194	8,33
PARANÁ	2.386	1.854	77,70	418	17,52	114	4,78
SANTA CATARINA	1.795	1.337	74,48	327	18,22	131	7,30
RIO GRANDE DO SUL	2.244	1.730	77,09	420	18,72	94	4,19
MATO GROSSO DO SUL	1.207	935	77,46	197	16,32	75	6,21
MATO GROSSO	1.393	1.035	74,30	271	19,45	87	6,25
GOIÁS	2.079	1.623	78,07	360	17,32	96	4,62
DISTRITO FEDERAL	1.889	1.313	69,51	434	22,98	142	7,52

5. EXPANSÃO DA AMOSTRA E ESTIMAÇÃO DA PRECISÃO DAS ESTIMATIVAS

A estimação de totais prevê a utilização do estimador natural do desenho amostral, após a reponderação para compensar a não-resposta dentro de cada estrato. Os domicílios listados como tendo proprietários do setor informal e que na entrevista não tinham mais esta característica (situação de entrevista do tipo A2) são incluídos na estimação de totais, mas recebem o valor zero para cada variável de interesse da investigação.

Para representar o estimador de totais das variáveis para o domínio I , definido pelo conjunto de domicílios com proprietários de unidades produtivas do setor informal, é conveniente definir a variável:

$$y_{hijk}(I) = \begin{cases} y_{hijk} & \text{se a entrevista no domicílio } k, \text{ do grupo } j, \text{ setor } i, \text{ estrato } h, \text{ for do tipo A1} \\ 0 & \text{caso contrário} \end{cases}$$

Usando essa notação, o estimador para o total de uma característica y , associada aos proprietários de unidades produtivas do setor informal, numa determinada área da pesquisa é dado por:

$$\hat{Y}_I = \sum_{h=1}^L \sum_{i=1}^{m_h} \sum_{j=1}^8 \sum_{k=1}^{n_{hij}^*} w_{hij} \times y_{hijk} (I)$$

onde:

w_{hij} é o fator de expansão ou peso associado aos proprietários de unidades produtivas do setor informal do grupo j , setor i , estrato h ; é obtido pelo inverso da probabilidade de inclusão de cada unidade da amostra ajustado pelo inverso da taxa de resposta de cada estrato de grupo de atividade, ou seja,

$$w_{hij} = \frac{1}{m_h P_{hi}} \times \frac{N_{hij}^*}{n_{hij}^*} = \frac{1}{m_h P_{hi}} \times \frac{N_{hij}^*}{n_{hij}} \times \frac{n_{hij}}{n_{hij}^*}$$

e,

m_h é o número de setores selecionados no estrato h ;

P_{hi} é a probabilidade de seleção do setor i do estrato h ;

N_{hij}^* é o total de domicílios com proprietários do setor informal do grupo j no setor i do estrato h , obtido pela listagem;

n_{hij}^* é o total de domicílios com entrevista realizada do tipo A1 ou A2 no grupo j no setor i do estrato h ;

n_{hij} é o total de domicílios selecionados no grupo j no setor i do estrato h ;

y_{hijk} é o valor da característica y no domicílio k do grupo j no setor i do estrato h ;

$y_{hijk} = 0$ se a entrevista no domicílio k do grupo j no setor i do estrato h não for do tipo A1;

O refinamento da estratificação dos domicílios por grupos de atividade e a realidade da situação da entrevista pode nos levar a situações em que o número de entrevistas (A1+A2) é nulo em determinado grupo de atividade. Neste caso é recomendável agrupar os estratos de grupos de atividade para o cálculo dos pesos ou fatores de expansão dos proprietários de unidades produtivas do setor informal.

Além dos pesos para estimação das características dos proprietários de unidades produtivas do setor informal, é necessário considerar, também, o peso associado à unidade produtiva propriamente dita, ou seja, o que será usado para estimar as características da

unidade produtiva. Isto se justifica em função de que uma única unidade pode ser propriedade de um ou mais sócios, sendo então necessário aplicar uma fator de correção para evitar a superestimação. Este segundo peso é dado por:

$$w_{hijk}^* = w_{hij} \times fator_{hijk}$$

onde:

$fator_{hijk}$ é o inverso do número de sócios da unidade produtiva.

Neste caso, o estimador para o total de uma característica y , associada à unidade produtiva é dado por:

$$\hat{Y}_I = \sum_{h=1}^L \sum_{i=1}^{m_h} \sum_{j=1}^8 \sum_{k=1}^{n_{hij}^*} w_{hijk}^* \times y_{hijk}(I)$$

A estimativa da variância de, \hat{Y}_I , usando Ultimate Cluster, pode ser escrita como:

$$v(\hat{Y}_I) = \sum_{h=1}^L \frac{m_h}{(m_h - 1)} \sum_{i=1}^{m_h} \left(\hat{Y}_{hi}(I) - \hat{Y}_h(I) \right)^2$$

onde:

$$\hat{Y}_{hi}(I) = \sum_{j=1}^8 \sum_{k=1}^{n_{hij}^*} w_{hijk} \times y_{hijk}(I) \quad \text{ou} \quad \hat{Y}_{hi}(I) = \sum_{j=1}^8 \sum_{k=1}^{n_{hij}^*} w_{hijk}^* \times y_{hijk}(I)$$

dependendo de ser característica de proprietário ou de unidade produtiva, respectivamente;

e

$$\hat{Y}_h(I) = \frac{1}{m_h} \sum_{i=1}^{m_h} \hat{Y}_{hi}(I)$$

OBS: se existir setor i no estrato h , tal que $N_{hij}^* = 0 \quad \forall j = 1, 2, \dots, 8$, então devemos considerar: $\hat{Y}_{hi}(I) = 0$.

O estimador do coeficiente de variação, associado ao estimador de total da característica y no domínio I , é definido por:

$$cv(\hat{Y}_I) = \frac{\sqrt{v(\hat{Y}_I)}}{\hat{Y}_I}$$

BIBLIOGRAFIA

Cochran, William G. *Sampling Techniques (3rd edition)*, John Wiley & Sons, New York, 1977.

Jorge, Angela F. *Pesquisa de Economia Informal Urbana*, Rio de Janeiro: IBGE, 1996. 17 p.
(Artigo apresentado no Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais).

Kalton, G. e Anderson, D.W. *Sampling Rare Populations. The Journal of the Royal Statistical Society A*, 149, part 1, pp 65-82, 1986.

Verma, V. *Methods of Data Collection on the Informal sector*. (Tanzânia), 1992.

ANEXO

Obtenção do estimador da variância para o dimensionamento da amostra

A partir do desenho amostral definido no capítulo 2, um estimador não viciado para o **total** de uma característica y qualquer é dado por:¹

$$\hat{Y} = \sum_{h=1}^L \hat{Y}_h = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{m_h} \frac{N_{hi}}{P_{hi}} \frac{1}{n_{hi}} \sum_{j=1}^{n_{hi}} y_{hij}$$

e a expressão da variância de \hat{Y} é dada por:

$$V(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} P_{hi} \left(\frac{Y_{hi}}{P_{hi}} - Y_h \right)^2 + \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{N_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}} \right) \frac{S_{hi}^2}{n_{hi}} \quad (1)$$

$$V(\hat{Y}) = V_1(\hat{Y}) + V_2(\hat{Y})$$

$$V_1(\hat{Y}) \text{ é a primeira componente da variância} \quad V_1(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} P_{hi} \left(\frac{Y_{hi}}{P_{hi}} - Y_h \right)^2$$

$$V_2(\hat{Y}) \text{ é a segunda componente da variância} \quad V_2(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{N_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}} \right) \frac{S_{hi}^2}{n_{hi}}$$

Um primeiro estimador proposto para a primeira componente da variância, $V_1(\hat{Y})$, é definido substituindo os valores populacionais por valores estimados pela amostra do CD/91 dado por:

$$\hat{V}_1(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} P_{hi} \left(\frac{\hat{Y}'_{hi}}{P_{hi}} - \hat{Y}'_h \right)^2$$

Uma análise das propriedades deste estimador é feita através da obtenção da expressão para a esperança do estimador, demonstrada a seguir:

$$E(\hat{V}_1(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} E \left[\sum_{i=1}^{M_h} P_{hi} \sum_{i=1}^{M_h} P_{hi} \left(\frac{\hat{Y}'_{hi}}{P_{hi}} - \hat{Y}'_h \right)^2 \right] = \sum_{h=1}^L \frac{1}{m_h} E \left[\sum_{i=1}^{M_h} \frac{\hat{Y}'_{hi}^2}{P_{hi}} - \hat{Y}'_h^2 \right]$$

¹ A notação é a mesma que foi adotada no item 2.2.

$$E(\hat{Y}'_1(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} \left[\sum_{i=1}^{M_h} \frac{1}{P_{hi}} E(\hat{Y}'_{hi}{}^2) - E(\hat{Y}'_h{}^2) \right]$$

Mas

$$E(\hat{Y}'_{hi}{}^2) = E(D_{hi}^2 \bar{y}_{hi}^{*2}) = D_{hi}^2 E(\bar{y}_{hi}^{*2}) = D_{hi}^2 \left(V(\bar{y}_{hi}^*) + E^2(\bar{y}_{hi}^*) \right)$$

supondo que a amostra de domicílios ocupados do CD/91 no setor i do estrato h , de tamanho d_{hi} , é aleatória simples sem reposição¹, tem-se:

$$E(\hat{Y}'_{hi}{}^2) = D_{hi}^2 \left[\left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} + \bar{Y}_{hi}^2 \right]$$

sendo:

$$\bar{y}_{hi}^* = \frac{1}{d_{hi}} \sum_{j=1}^{d_{hi}} y_{hij}$$

$$S_{hi}^{*2} = \frac{1}{D_{hi} - 1} \sum_{j=1}^{D_{hi}} (Y_{hij} - \bar{Y}_{hi})^2$$

e

$$E(\hat{Y}'_h{}^2) = V(\hat{Y}'_h) + E^2(\hat{Y}'_h)$$

$$E(\hat{Y}'_h{}^2) = V\left(\sum_{i=1}^{M_h} \hat{Y}'_{hi}\right) + E^2\left(\sum_{i=1}^{M_h} \hat{Y}'_{hi}\right) = \sum_{i=1}^{M_h} V(\hat{Y}'_{hi}) + \left(\sum_{i=1}^{M_h} E(\hat{Y}'_{hi})\right)^2$$

$$E(\hat{Y}'_h{}^2) = \sum_{i=1}^{M_h} D_{hi}^2 \left[\left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} \right] + \left(\sum_{i=1}^{M_h} Y_{hi} \right)^2$$

$$E(\hat{Y}'_h{}^2) = \sum_{i=1}^{M_h} D_{hi}^2 \left[\left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} \right] + Y_h^2$$

Então:

¹ Na realidade a amostra do CD/91 é sistemática em cada setor.

$$E(\hat{V}_1(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} \frac{1}{P_{hi}} \left(D_{hi}^2 \left[\left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} + \bar{Y}_{hi}^2 \right] \right) - \left(\sum_{i=1}^{M_h} D_{hi}^2 \left[\left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} \right] + Y_h^2 \right) \right\}$$

$$E(\hat{V}_1(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} \frac{D_{hi}^2}{P_{hi}} \left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} + \sum_{i=1}^{M_h} \frac{D_{hi}^2}{P_{hi}} \bar{Y}_{hi}^2 - \sum_{i=1}^{M_h} D_{hi}^2 \left[\left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} \right] - Y_h^2 \right\}$$

$$E(\hat{V}_1(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} \frac{D_{hi}^2}{P_{hi}} \left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} + \sum_{i=1}^{M_h} \frac{Y_{hi}^2}{P_{hi}} - \sum_{i=1}^{M_h} D_{hi}^2 \left[\left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} \right] - Y_h^2 \right\}$$

$$E(\hat{V}_1(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} \frac{Y_{hi}^2}{P_{hi}} - Y_h^2 + \sum_{i=1}^{M_h} \left(\frac{1}{P_{hi}} - 1 \right) D_{hi}^2 \left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} \right\}$$

$$E(\hat{V}_1(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} P_{hi} \left(\frac{Y_{hi}}{P_{hi}} - Y_h \right)^2 + \sum_{i=1}^{M_h} \left(\frac{1}{P_{hi}} - 1 \right) D_{hi}^2 \left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}} \right\}$$

Podemos então notar que:

$$\sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \left(\frac{1}{P_{hi}} - 1 \right) D_{hi}^2 \left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{S_{hi}^{*2}}{d_{hi}}$$

constitui-se numa parcela de vício da 1ª componente do estimador da variância $\hat{V}_1(\hat{Y})$.

Neste caso, foi então considerado o estimador não viciado para a primeira componente da variância, $V_1(\hat{Y})$, definido por: $v_1(\hat{Y})$:

$$v_1(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} P_{hi} \left(\frac{\hat{Y}'_{hi}}{P_{hi}} - \hat{Y}'_h \right)^2 - \sum_{i=1}^{M_h} \left(\frac{1}{P_{hi}} - 1 \right) D_{hi}^2 \left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{s_{hi}^2}{d_{hi}} \right\} \quad (2)$$

Um estimador consistente proposto para a segunda componente da variância, $V_2(\hat{Y})$, é definido substituindo os valores populacionais por valores estimados pela amostra do CD/91, e é dado por:

$$v_2(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{\hat{N}_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{\hat{N}_{hi}} \right) \frac{s_{hi}'^2}{n_{hi}} \quad (3)$$

Diante da complexidade para a obtenção do vício desse estimador para a 2ª componente da variância, fez-se uma aproximação de que $\hat{N}_{hi} = N_{hi} \quad \forall i, \forall j$, ou seja, de

que a estimativa do total de domicílios com proprietários do setor informal proveniente da amostra do CD/91 corresponde ao valor do respectivo total no universo, para qualquer setor de qualquer estrato.

Neste caso, $v_2(\hat{Y})$ fica não viciado para estimar $V_2(\hat{Y})$, pois:

$$v_2(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{N_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}}\right) \frac{s_{hi}'^2}{n_{hi}}$$

$$E(v_2(\hat{Y})) = E\left(\sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{N_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}}\right) \frac{s_{hi}'^2}{n_{hi}}\right) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{N_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}}\right) \frac{E(s_{hi}'^2)}{n_{hi}}$$

supondo que a amostra de domicílios ocupados do CD/91 no setor i do estrato h , de tamanho d_{hi} , é aleatória simples sem reposição, então as n'_{hi} unidades da amostra do CD/91 com proprietários do setor informal também constituem uma amostra aleatória simples sem reposição de tamanho n'_{hi} da subpopulação dos domicílios com proprietários do setor informal, e portanto, $E(s_{hi}'^2) = S_{hi}^2$. Logo,

$$E(v_2(\hat{Y})) = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{N_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}}\right) \frac{S_{hi}^2}{n_{hi}} = V_2(\hat{Y})$$

Portanto, juntando o estimador obtido na expressão (2) com o estimador definido na expressão (3), tem-se o estimador consistente $v(\hat{Y})$, que é o adotado para estimar a variância definida em (1), e é dado por:

$$v(\hat{Y}) = v_1(\hat{Y}) + v_2(\hat{Y})$$

$$v(\hat{Y}) = \sum_{h=1}^L \frac{1}{m_h} \left\{ \sum_{i=1}^{M_h} P_{hi} \left(\frac{\hat{Y}'_{hi}}{P_{hi}} - \hat{Y}'_h \right)^2 - \sum_{i=1}^{M_h} \left(\frac{1}{P_{hi}} - 1 \right) D_{hi}^2 \left(1 - \frac{d_{hi}}{D_{hi}} \right) \frac{s_{hi}^2}{d_{hi}} \right\} + \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{\hat{N}_{hi}^2}{P_{hi}} \left(1 - \frac{n_{hi}}{\hat{N}_{hi}} \right) \frac{s_{hi}'^2}{n_{hi}}$$