DPE
Doação. CR$ 800,00

TEXTOS PARA DISCUSSÃO, ISSN 0103-6661

91

UMA PROPOSTA DE METODOLOGIA PARA A EXPANSÃO DA AMOSTRA DO CENSO DEMOGRÁFICO DE 1991

NÚMERO 62 OUTUBRO DE 1993

Presidente da República
Itamar Franco

Ministro-Chefe da Secretaria de Planejamento, Orçamento e Coordenação
Alexis Stepanenko

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Silvio Augusto Minciotti

Diretor de Planejamento e Coordenação
Maurício de Souza Rodrigues Ferrão

ÓRGÃOS TÉCNICOS SETORIAIS

Diretoria de Pesquisas
Tereza Cristina Nascimento Araújo

Diretoria de Geociências
Sergio Bruni

Diretoria de Informática
Francisco Quental

Centro de Documentação e Disseminação de Informações
Nelson de Castro Senra

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE
DIRETORIA DE PESQUISAS - DPE

UMA PROPOSTA DE METODOLOGIA PARA A EXPANSÃO DA AMOSTRA DO CENSO DEMOGRÁFICO DE 1991

PEDRO LUIS DO NASCIMENTO SILVA
Bacharel e Mestre em Estatística
ZÉLIA MAGALHÃES BIANCHINI
Matemática e Mestre em Estatística
SONIA ALBIERI
Bacharel e Mestre em Estatística

RIO DE JANEIRO

1993

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Av. Franklin Roosevelt, 166 - Centro

CEP 20 271-201 - Rio de janeiro, RJ - Brasil

DIRETORA DE PESQUISAS

TEREZA CRISTINA NASCIMENTO ARAÚJO

DIRETORA-ADJUNTA DE PESQUISAS

MARIA MARTHA MALARD MAYER (respondendo)

CHEFE DA DIVISÃO DE METODOLOGIA

ZÉLIA MAGALHÃES BIANCHINI

GE000012599-3

IBGE - CDDI/DEDOC
REDE DE BIBLIOTECAS
N.º de Reg.: 41
Data: 21/02/94

311.213.1(81)1991
S586p
DPE

© 1990, 91, 92, 93

TEXTOS PARA DISCUSSÃO

Série publicada pela Diretoria de Pesquisas do IBGE, com o objetivo de divulgar ensaios, estudos e outros trabalhos técnicos nas áreas econômica, social e demográfica, elaborados no âmbito da Diretoria.

Edição: Divisão de Documentação e Disseminação da Diretoria de Pesquisas. (DDI/DPE)

Silva, Pedro Luis do Nascimento
Uma proposta de metodologia para a expansão da amostra do censo demográfico de 1991 / Pedro Luis do Nascimento Silva, Zélia Magalhães Bianchini, Sonia Albieri. - Rio de Janeiro : IBGE, Diretoria de Pesquisas, 1993.
106 p. - (Textos para discussão, ISSN 0103-6661 ; n.62)

Inclui bibliografia.

ISBN 85-240-0472-X

1. Censo - Metodologia. 2. Brasil - Censo Demográfico, 1991 - Metodologia. I. Biachini, Zélia Magalhães. II. Albieri, Sonia. III. IBGE. Diretoria de Pesquisas. IV. Título. V Série.

IBGE.CDDI. Dep. de Documentação e Biblioteca
RJ/IBGE-93/30 CDU 311.213.1:312(81)"1991"
EST

Informações: Biblioteca Setorial da Diretoria de Pesquisas - Rua Visconde de Niterói, 1.246, Bloco B, sala 1.211-B, Mangueira
Telefone: (021) 284-8938 / 284-3322 - ramal 303

AGRADECIMENTOS

A realização deste trabalho não seria possível sem a colaboração de algumas pessoas, a quem deixamos aqui registrado nosso mais sincero agradecimento.

A John Kovar, Mike Hidiroglou, Michael Bankier, Stephen Rathwell e Susana Rubin, todos do *Statistics Canada*, pela ajuda através da cessão de bibliografia, dos programas de computador e pela sempre pronta disposição de atender nossos pedidos e consultas.

Ao Prof. Djalma Galvão Carneiro Pessoa da ENCE, que nos ensinou o *caminho das pedras* e participou conosco das primeiras discussões sobre a metodologia de expansão a ser usada no Censo, indicando rumos para nossa pesquisa, ministrando curso de Análise Multivariada de Dados Discretos do qual participamos com grande interesse, e que esteve junto conosco quando empreendemos a primeira leitura do artigo de Bankier(1990) que expõe a metodologia de estimação que se vai aqui examinar com detalhes.

A Jelke Bethlehem, Wouter Keller, Marly Odekerken e Robert Göttgens do *Central Bureau of Statistics* da Holanda, que também nos ajudaram fornecendo literatura e sistemas computacionais, que embora não tenhamos proposto usar na expansão da amostra do Censo 91, estão servindo a outras aplicações não menos importantes.

A Lenildo Fernandes Silva e Teresa Cristina Nascimento Araújo, diretores da Diretoria de Pesquisas do IBGE, que deram seu apoio mantendo os técnicos que participaram deste trabalho unidos e dedicados ao tema durante os últimos 3 anos e fornecendo as condições necessárias à realização viagens de intercâmbio e outras atividades não menos importantes para a conclusão do trabalho.

A Eduardo Robson Tardin Costa, que forneceu suporte técnico indispensável à solução de alguns problemas computacionais enfrentados durante a realização do trabalho.

APRESENTAÇÃO

Este documento pretende apresentar para exame, debate e decisão uma proposta de nova metodologia para ser aplicada na expansão da amostra do Censo Demográfico de 1991. Tal proposta resulta dos estudos e da pesquisa que vêm sendo conduzidos pela equipe da Divisão de Metodologia desde 1989, e baseia-se essencialmente na metodologia desenvolvida e aplicada para a expansão do Censo canadense de 1991, ora em curso naquele país. Pretende-se com essa metodologia substituir a que foi empregada pelo IBGE no Censo Demográfico de 1980 e a que foi proposta e estudada por Costa(1987a, 1988, 1989, 1991) e Costa e Lima(1991) com aplicação aos dados do Censo Experimental de Limeira de 1988.

SUMÁRIO

1. Introdução 9

2. Processos de estimação empregados nos censos demográficos passados 13

2.1 Introdução e notação básica 13

2.2 Estimadores empregados nos censos de 60 e 70 16

2.3 Processo de estimação empregado no censo de 80 19

3. Processo de estimação grupo de controle iterativo 27

4. Um resumo das alternativas de métodos de estimação consideradas 34

5. A metodologia de estimação proposta para o censo 91 44

5.1 Introdução 44

5.2 Descrição geral da metodologia GLSEP para o caso univariado 44

5.3 Descrição geral da metodologia GLSEP para o caso multivariado 47

5.4 Detalhamento da metodologia de estimação proposta para o censo 91 58

5.4.1 Estágio preparatório 62

5.4.2 Primeiro estágio de ajuste dos pesos 75

5.4.3 Segundo estágio de ajuste dos pesos 79

5.5 Razões que recomendam a adoção da metodologia MQG2 84

6. Uma aplicação experimental da metodologia MQG2 89

7. Pendências e linhas de trabalho para completar a especificação da metodologia a ser usada na expansão da amostra do censo 91 101

8. Referências 104

1. INTRODUÇÃO

O desenho amostral adotado na coleta de parte dos dados dos Censos Demográficos do Brasil (1960, 1970, 1980 e 1991) compreende a seleção sistemática e com equiprobabilidade, dentro de cada setor censitário, de uma amostra dos domicílios particulares e das famílias ou componentes de grupos conviventes recenseados em domicílios coletivos, com fração amostral constante para setores censitários de um mesmo município. Nos Censos Demográficos de 60, 70 e 80 a fração amostral foi fixada em 25% para todos os municípios, enquanto que para o Censo Demográfico de 1991 (Censo 91) foram usadas duas frações amostrais distintas: 10% para os municípios com população superior a 15.000 habitantes e 20% para os demais municípios[1].

Na coleta das informações do Censo 91 foram usados dois modelos de questionário: um questionário básico (denominado CD-1.01) aplicado nas unidades não selecionadas para a amostra e contendo as perguntas referentes às características que foram investigadas para 100% da população, e um questionário de amostra (denominado CD-1.02) contendo, além das perguntas que também constam do questionário básico, outras perguntas mais detalhadas sobre características do domicílio e das pessoas (tais como religião, migração, escolaridade, fecundidade, mão-de-obra, rendimento).

As informações da amostra precisam ser "expandidas" para que se possa obter resultados para o total da população pesquisada, isto é, os dados coletados exclusivamente no questionário da amostra vão alimentar um processo de estimação que vai dar origem às estimativas de totais das características de domicílios, de famílias e de pessoas investigadas por amostragem.

---

[1] Ver Metodologia do Censo Demográfico de 1980, Silva & Bianchini (1990) e Pinheiro & Lima (1988, 1990).

Uma primeira alternativa que se tem disponivel para a metodologia de expansão da amostra seria o emprego dos estimadores naturais correspondentes ao desenho amostral adotado. No caso das características de domicílios, a amostra do Censo é estratificada por setor censitário, com seleção sistemática de domicílios em cada setor censitário. Por outro lado, para as características de pessoas e de famílias, também há um estágio de conglomeração, já que todas as pessoas ou famílias de cada domicílio selecionado são pesquisadas.

Acontece que, no Censo Demográfico, são conhecidos os totais da população para um subconjunto das características investigadas (as que são pesquisadas a 100%), o que coloca um problema: se os estimadores naturais forem aplicados à amostra para estimar os totais dessas características cujos valores "verdadeiros"[2] são conhecidos na população, os valores das estimativas assim calculadas apresentarão diferenças de tais valores, o que é indesejável. Por outro lado, o conhecimento desses totais na população torna viável a aplicação de métodos alternativos para o processo de expansão da amostra, os quais procuram, entre outras finalidades, satisfazer os seguintes objetivos principais:

a) buscar a consistência das estimativas baseadas somente nas informações da amostra dos totais das características investigadas a 100% com os totais conhecidos da população para essas mesmas características;

b) obter estimativas mais precisas para os totais das características investigadas por amostragem;

c) controlar na estimação possíveis tendências ocorridas durante a coleta dos dados, especialmente na aplicação do mecanismo de seleção da amostra de domicílios e pessoas.

---

[2] A menos dos erros alheios à amostragem.

Desde 1960 que a metodologia de estimação adotada para a expansão da amostra dos Censos Demográficos brasileiros faz uso das informações disponíveis para o total da população para permitir o emprego de estimadores mais apropriados (e complexos) que os estimadores naturais do desenho. O capítulo 2 apresenta uma breve restrospectiva dos métodos de estimação empregados nos censos anteriores, com maior ênfase e detalhamento na descrição do processo de estimação adotado no Censo Demográfico de 1980. É também efetuada uma análise crítica da metodologia empregada no Censo de 1980, apontando-se os principais "inconvenientes" dessa metodologia que conduziram à procura de outras metodologias alternativas.

No capítulo 3 é descrito e analisado criticamente o método denominado Grupo de Controle Iterativo (GCI), sugerido por Costa (1987a) e experimentado por aquele autor com dados do Censo Demográfico Experimental de Limeira de 1988 - ver Costa(1988, 1989, 1991) e Costa e Lima(1991). Tal análise crítica baseia-se essencialmente em Bianchini(1991). Essa análise conduziu à rejeição da idéia de adotar o método GCI para a expansão da amostra do Censo Demográfico de 1991, e justificou a realização de um esforço de pesquisa para identificar outros métodos alternativos que pudessem ser empregados para expandir a amostra do Censo 91. Uma descrição das alternativas consideradas e do resultado dessa pesquisa de alternativas é apresentada no capítulo 4, que se baseia principalmente em Albieri(1991a,b).

No capítulo 5 é apresentada com detalhes a metodologia desenvolvida no *Statistics Canada* para a expansão da amostra do Censo Demográfico de 1991 daquele país - ver Bankier(1989, 1990) e Rathwell(1991). Essa metodologia, denominada *Generalized Least Squares Estimation Procedure - GLSEP* - se baseia no emprego de estimadores de Mínimos Quadrados Generalizados em dois estágios. A metodologia GLSEP é a que se propõe adotar para a expansão da amostra do Censo Demográfico de 1991.

O capítulo 6 descreve uma aplicação experimental da metodologia GLSEP aos dados coletados no Censo Demográfico Experimental de Limeira de 1988, e apresenta uma análise preliminar dos resultados dessa aplicação.

O capítulo 7 apresenta uma análise das demais componentes da metodologia necessária para a expansão da amostra do Censo Demográfico de 1991, servindo como uma agenda de trabalho para orientar os esforços e o trabalho ainda necessários para concluir a especificação completa da metodologia de expansão a ser adotada.

O texto contém ainda uma seção dedicada à apresentação das referências bibliográficas citadas no texto e usadas no trabalho (seção 8) e uma seção dedicada à apresentação dos anexos necessários à complementação do texto (seção 9).

2. PROCESSOS DE ESTIMAÇÃO EMPREGADOS NOS CENSOS DEMOGRÁFICOS PASSADOS

2.1. Introdução e Notação Básica

A expansão da amostra dos Censos Demográficos brasileiros anteriores sempre foi feita separadamente para cada área de ponderação[3]. O conjunto das áreas de ponderação é definido em cada Censo de acordo com a respectiva base geográfica, sempre de modo que todo setor pertença a alguma área de ponderação e que as áreas de ponderação reunidas cubram todo o território nacional. Assim, para obter estimativas referentes a uma área qualquer, formada por algum conjunto de áreas de ponderação, basta agregar as estimativas das áreas ponderação que a integram (no caso de estimativas de totais). Por esse processo de agregação é que são calculadas as estimativas a nível de Unidades da Federação e do total do país.

Nos Censos de 1960 e 1970, uma área de ponderação foi definida como o conjunto de setores de cada município existente nas respectivas bases geográficas, sendo que o número de municípios era de 3.952 em 1970. No caso do Censo de 1970, houve uma única exceção à regra de formação das áreas de ponderação já que, no município do Rio de Janeiro, foram formadas áreas de ponderação separadas para cada uma das suas Regiões Administrativas. No Censo de 1980, as áreas de ponderação foram definidas como sendo o município, mas os municípios mais populosos foram subdivididos para formar áreas de ponderação, tendo sido bem maior o número de municípios subdivididos que em 1970. No total, a base geográfica do Censo de 1980 apresentava 3.991 municípios, dos quais 3.944 formaram uma área de ponderação cada um e 47 foram subdivididos em 2 ou mais áreas de ponderação, tendo sido ao todo definidas e utilizadas 4.219 áreas de ponderação para a expansão da amostra.

---

[3]Área de ponderação é a menor área geográfica para a qual são calculadas estimativas e divulgados resultados baseados nas informações do questionário da amostra do Censo Demográfico. Funciona como estrato geográfico. Uma área de ponderação é sempre um conjunto de setores censitários. Cada setor censitário pertence a uma única área de ponderação.

No restante desse texto, serão apresentadas as descrições de diferentes processos de estimação efetuados a nível de cada uma das áreas de ponderação, já que a expansão da amostra a nível de área de ponderação forma a base para a obtenção de todos os resultados derivados das informações pesquisadas exclusivamente no questionário de amostra do Censo Demográfico. Em primeiro lugar, descreve-se o desenho amostral efetivamente usado para selecionar a amostra do Censo. Em seguida, são apresentadas as hipóteses consideradas na construção dos processos de estimação.

Numa área de ponderação qualquer, a amostra do Censo é uma amostra estratificada de domicílios, com a estratificação definida pelos setores censitários e seleção sistemática dos domicílios com equiprobabilidade dentro de cada setor, sendo que a fração de amostragem é constante ao longo dos setores, o que equivale a uma alocação proporcional da amostra nos estratos (setores censitários). Para as características de pessoas (e de famílias), é preciso considerar ainda um estágio de conglomeração, já que todas as pessoas (e famílias) residentes em um domicílio selecionado são pesquisadas usando o questionário da amostra.

Para o esquema de amostragem definido anteriormente, sejam:

M o número de domicílios da área ponderação considerada;

T o número de setores censitários (estratos) da área de ponderação considerada;

M_t o número de domicílios do setor t, para $t=1,\ldots,T$;

m o número de domicílios da amostra da área de ponderação considerada;

m_t o número de domicílios da amostra do setor t, para $t=1,\ldots,T$;

$d = m/M$ a fração de amostragem de domicílios observada da área de ponderação considerada;

$d_t = m_t/M_t$ a fração de amostragem de domicílios observada do setor t, para $t=1,\ldots,T$;

N o número de pessoas da área de ponderação considerada;

n o número de pessoas da amostra da área de ponderação considerada;

N_t o número de pessoas do setor t, para $t=1,\ldots T$;

n_t o número de pessoas da amostra do setor t, para $t=1,\ldots T$;

N_{tI} o número de pessoas do I-ésimo domicílio do setor t, para $I=1,\ldots,M_t$ e $t=1,\ldots T$;

N_{ti} o número de pessoas do i-ésimo domicílio da amostra do setor t, para $i=1,\ldots,m_t$ e $t=1,\ldots T$;

$p = n/N$ a fração de amostragem de pessoas observada da área de ponderação considerada;

$p_t = n_t/N_t$ a fração de amostragem de pessoas observada do setor t, para $t=1,\ldots T$.

Pode-se derivar diretamente das definições acima que:

$$N_t = \sum_{I=1}^{M_t} N_{tI}$$

$$n_t = \sum_{i=1}^{m_t} N_{ti}$$

$$N = \sum_{t=1}^{T} N_t$$

$$n = \sum_{t=1}^{T} n_t$$

Para a montagem dos processos de estimação que foram empregados nos Censos Demográficos passados, apesar de ter sido aplicado para a seleção da amostra o esquema que se acabou de descrever, admitiu-se um desenho amostral distinto, mais simples, que não leva em conta os efeitos da estratificação por setores censitários, da seleção sistemática dos domicílios e da conglomeração das pessoas por domicílios, a fim de simplificar a análise das propriedades dos estimadores empregados[4].

[4]Uma análise dos efeitos dessa hipótese sobre as propriedades dos estimadores empregados considerando-se o esquema de seleção efetivamente adotado será objeto de outros trabalhos. Por exemplo, Silva e Albieri(1991) examinam essa questão com respeito à estimação de características de domicílios, empregando dados do Censo Demográfico Experimental de Limeira de 1988.

Assim, para a montagem dos processos de estimação e análise das propriedades dos estimadores adotados, admitiu-se que uma amostra aleatória simples de unidades da população (domicílios ou pessoas) foi investigada em cada área de ponderação para obter as informações mais detalhadas de interesse.

Por outro lado, é sabido que no Censo Demográfico estão disponíveis informações sobre algumas características pesquisadas no questionário da amostra cujos totais são conhecidos para 100% das unidades da população (domicílios ou pessoas) existentes na área de ponderação considerada. Essas informações formam a base para o emprego dos estimadores que se vai passar a descrever nas seções seguintes.

2.2 Estimadores Empregados nos Censos de 60 e 70

Nos Censos Demográficos de 1960 e 1970 foram empregados processos de estimação bastante semelhantes, baseados no uso de pós-estratificação. Tais processos de estimação foram batizados de "grupos de controle". A principal diferença entre os processos de estimação usados nesses Censos foi o número de pós-estratos empregados em cada um. No Censo de 1970, o número inicial de pós-estratos formados em cada área de ponderação foi de 46, número maior que o de "grupos de controle" empregados no Censo de 1960.

A idéia básica dos processos de estimação de "grupos de controle" aplicados nos Censos de 1960 e 1970 foi estratificar as pessoas da amostra segundo algumas das variáveis pesquisadas a 100% (variáveis auxiliares). No Censo de 70, as variáveis auxiliares usadas na pós-estratificação foram: condição de presença (morador presente, morador ausente, não morador presente), situação do domicílio (urbano ou rural), condição no domicílio (chefe ou não chefe), sexo e idade (5 grupos etários). Com base nas variáveis auxiliares acima indicadas, foram então formados inicialmente os 46 pós-estratos mencionados[5]. No que segue, apresenta-se a descrição do processo de estimação adotado no Censo de 70, análogo ao que foi usado no Censo de 60.

---

[5] A definição dos 46 pós-estratos usados no Censo de 70 pode ser encontrada na Metodologia do Censo Demográfico de 1980, página 207.

O estimador empregado nos Censos de 60 e 70 para obter o total de uma certa característica de interesse y a nível da área de ponderação foi:

$$\hat{Y}^{GC} = \sum_{h=1}^{H} N_h \cdot \bar{y}_h$$

onde

$\hat{Y}^{GC}$ é o estimador do método de Grupos de Controle para o total da variável y a nível da área de ponderação;

$\bar{y}_h = \frac{1}{n_h} \cdot \sum_{j=1}^{n_h} y_{hj}$ é a média amostral da variável y no pós-estrato h, para $h=1,\ldots,H$;

y_{hj} é o valor da variável de interesse y para a j-ésima pessoa da amostra da área de ponderação pertencente ao pós-estrato h, para $j=1,\ldots,n_h$ e $h=1,\ldots,H$;

N_h é o número total de pessoas da área de ponderação pertencentes ao pós-estrato h, $h=1,\ldots,H$;

n_h é o número de pessoas da amostra pertencentes ao pós-estrato h na área de ponderação, $h=1,\ldots,H$.

A alocação das pessoas nos H pós-estratos formados era feita com base nas variáveis auxiliares mencionadas, permitindo assim que fossem efetuadas as contagens do número de pessoas na amostra (n_h) e na população (N_h) da área de ponderação em cada pós-estrato ou grupo de controle.

Além disso, o estimador $\hat{Y}^{GC}$ apresentado acima pode ser reescrito para evidenciar a definição dos pesos ou fatores de expansão que foram posteriormente armazenados para cada pessoa, de modo a simplificar a obtenção das estimativas de totais para as variáveis pesquisadas por amostragem. Outra razão para reexpressar o estimador $\hat{Y}^{GC}$ é a conveniência da notação que mais tarde será usada para definir outros processos de estimação mais complexos, como o que foi aplicado no Censo de 1980 e o que se propõe aplicar para o Censo de 1991. Assim, pode-se reescrever o estimador $\hat{Y}^{GC}$ como:

$$\hat{Y}^{GC} = \sum_{h=1}^{H} w_h \cdot \sum_{j=1}^{n_h} y_{hj} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} w_h \cdot y_{hj}$$

onde

$w_h = \frac{N_h}{n_h}$ é o peso (ou fator de expansão) aplicado para todas as pessoas do pós-estrato h, $h=1,\ldots,H$.

Os pesos calculados para cada pessoa da amostra foram arredondados para inteiros, para evitar a obtenção de estimativas fracionárias para algum total. Os totais das características de famílias e domicílios foram estimados empregando-se os pesos correspondentes aos respectivos chefes (da família e do domicílio).

Vale ainda notar que o estimador acima foi aplicado em cada área de ponderação sobre os pós-estratos (ou grupos de controle) resultantes de um processo de reagrupamento dos 46 pós-estratos inicialmente definidos, reagrupamento esse que foi aplicado sempre que, em algum pós-estrato inicial, uma das situações abaixo era observada:

$N_h \leq 100$ ou $w_h \geq 16$.

Assim, o número final de pós-estratos empregados para a estimação em cada área de ponderação podia variar em função das particularidades da população e da amostra de cada área, devido à eventual necessidade de reagrupamento dos pós-estratos inicialmente definidos.

Não é demais destacar também que o estimador $\hat{Y}^{GC}$ é condicionalmente não-viciado, sempre que $n_h > 0$ para todo h, $h=1,\ldots,H$. As condições de tamanho impostas para a formação dos pós-estratos não garantiam diretamente esta condição, mas como sempre se exigia que $N_h > 100$ para que um pós-estrato fosse mantido separado e como a fração amostral empregada era de 1/4 ou 25%, a probabilidade de ocorrência de alguma situação em que $N_h > 100$ e $n_h = 0$ era extremamente baixa. Não se tem registro de nenhuma área de ponderação em que tal situação tenha sido observada na expansão das amostras dos Censos de 60 ou 70.

2.3 Processo de Estimação Empregado no Censo de 80

A exemplo do que foi feito na seção anterior, aqui também se vai descrever todo o processo de estimação para uma área de ponderação qualquer, já que o processo de estimação empregado também se baseou na mesma idéia de estimação separada para cada área de ponderação. Vale destacar apenas que houve um aumento no número de áreas de ponderação definidas, que passou de 3.952 em 70 para 4.219 em 80, como já mencionado.

A seção 2.2 resumiu os esquemas de estimação adotados nos Censos de 1960 e 1970 pelo IBGE. Para o Censo Demográfico de 1980 decidiu-se modificar o processo de estimação, tendo em vista os avanços metodológicos realizados nessa questão em outros países (tais como Canadá e Estados Unidos), bem como um interesse em ampliar o uso das informações auxiliares disponíveis para a construção de pós-estratos, mediante o aumento do número de pós-estratos inicialmente definidos. Além disso, a modificação do processo de estimação tinha por objetivo aumentar a "consistência" das estimativas amostrais para as características pesquisadas a 100% com os respectivos valores "verdadeiros" conhecidos para um número maior de variáveis, tanto para as variáveis referentes a domicílios como para as referentes a pessoas.

Com esses objetivos em mente, foi empregado um esquema de estimação denominado Processo Iterativo de Estimação por Totais Marginais (PIETOM), baseado no *Univariate Raking Ratio Estimation Procedure (RREP)*. Tal processo de estimação foi aplicado separadamente para a estimação de características de pessoas e de domicílios, porém de maneira análoga. Devido a essa separação do processo de estimação em dois (um para domicílios e outro para pessoas), será descrito aqui com detalhes e notação própria somente o caso da estimação de características de pessoas. Mas vale destacar que a adaptação dessa notação para o caso do processo de estimação de características de domicílios é imediata.

O princípio básico em que se apóia o processo de estimação denominado RREP consiste em adotar uma pós-estratificação definida por uma tabela de contingência[6] com R linhas e C colunas. Em seguida, um peso não viciado $w_{rck}^{(0)}$ é atribuído à k-ésima pessoa da amostra pertencente à célula rc da tabela. Esse peso inicial geralmente é igual ao inverso da probabilidade de seleção de cada pessoa.

O método RREP procura determinar os pesos finais $w_{rck} = a_{rc} \cdot w_{rck}^{(0)}$ (ou os fatores de ajuste a_{rc}) de tal forma que seja minimizada a função de discriminação de informação (função de perda ou função de distância)[7] para o estimador *"Raking" Univariado (RU)*:

$$G^{RU} = \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{X}_{rc}^{(0)} \cdot \log_n \left(\frac{\hat{X}_{rc}}{\hat{X}_{rc}^{(0)}} \right)$$

sujeito às restrições de consistência nas marginais da tabela

$\hat{X}_{r.} = X_{r.}$ para $r=1,\ldots,R$;

$\hat{X}_{.c} = X_{.c}$ para $c=1,\ldots,C$;

onde

G^{RU} é a função de perda correspondente à metodologia de *Raking Ratio Estimation* univariada;

$$X_{rc} = \sum_{k=1}^{N_{rc}} X_{rck}$$

N_{rc} é o número de pessoas da população na célula rc da tabela de pós-estratificação da área de ponderação considerada;

$$\hat{X}_{rc} = \sum_{k=1}^{n_{rc}} w_{rck} \cdot x_{rck}$$

n_{rc} é o número de pessoas da amostra na célula rc da tabela de pós-estratificação da área de ponderação considerada;

---

[6] Também usualmente denominada matriz de pós-estratificação. Neste texto dar-se-á preferência ao uso do termo tabela de pós-estratificação.

[7] Vale observar que a definição usada aqui para a função de perda implica, pela tentativa de minimizar essa função, em obter os pesos finais tais que as restrições de consistência sejam satisfeitas, mas de modo a manter mínima a distância entre as estimativas não-viciadas baseadas nos pesos iniciais $w_{rck}^{(0)}$ e as estimativas baseadas nos pesos finais ajustados w_{rck}.

$$\hat{X}_{rc}^{(0)} = \sum_{k=1}^{n_{rc}} w_{rck}^{(0)} \cdot x_{rck}$$

X_{rck} é o valor da variável auxiliar x para a k-ésima pessoa da população na célula rc da tabela de pós-estratificação da área de ponderação considerada;

x_{rck} é o valor da variável auxiliar x para a k-ésima pessoa da amostra na célula rc da tabela de pós-estratificação da área de ponderação considerada.

Admite-se, na definição da função de perda G^{RU}, que a somação inclui apenas os pós-estratos (células da tabela de contingência) nos quais $\hat{X}_{rc}^{(0)} \neq 0$. Para X_{rc}, $\hat{X}_{rc}$ e $\hat{X}_{rc}^{(0)}$ a somação sobre algum dos subíndices é indicada através da substituição desse subíndice por um ponto (.).

No Censo Demográfico de 1980, X_{rck} é igual a 1 se a k-ésima pessoa da população pertence à célula rc da tabela de pós-estratificação definida e 0 caso contrário. Desta forma, $X_{rc} = N_{rc} = \sum_{k=1}^{N_{rc}} X_{rck}$ será igual ao número de pessoas na célula rc da tabela de pós-estratificação na população da área de ponderação considerada. De maneira análoga, x_{rck} é igual a 1 se a k-ésima pessoa da amostra da área de ponderação considerada pertence à célula rc da tabela de pós-estratificação, e 0 em caso contrário, e se pode escrever que $x_{rc} = n_{rc} = \sum_{k=1}^{n_{rc}} x_{rck}$.

O sistema de equações que precisa ser resolvido para determinar os pesos finais w_{rck} ou os correspondentes coeficientes de ajuste a_{rc} é não-linear. Por essa razão, um procedimento iterativo tem que ser usado para resolver tal sistema - veja Arora e Brackstone(1977a,b) e Brackstone e Rao(1979). No caso do Censo Demográfico brasileiro de 1980, foi usada uma versão desse procedimento iterativo que compreendia a execução de apenas 4 iterações (PIETOM), por razões de economia de tempo computacional e para simplificação da estimação da precisão. Assim, o cálculo dos pesos finais empregados na expansão da amostra do Censo Demográfico de 1980 foi efetuado segundo o algoritmo descrito a seguir:

Passo 1 - calcula os pesos iniciais $w_{rck}^{(0)}$ por

$$w_{rck}^{(0)} = \frac{N}{n} \qquad \text{para todo } k=1,\ldots,n_{rc},\ r=1,\ldots,R,\ \text{e } c=1,\ldots,C;$$

Passo 2 - inicia processo iterativo, fazendo o contador de iterações $u=1$;

Passo 3 - se u é ímpar, calcula o peso ajustado na iteração u por:

$$w_{rck}^{(u)} = w_{rck}^{(u-1)} \cdot \frac{N_{.c}}{\sum_{a=1}^{R} n_{ac} \cdot w_{ack}^{(u-1)}}$$

Passo 4 - se u é par, calcula o peso ajustado na iteração u por:

$$w_{rck}^{(u)} = w_{rck}^{(u-1)} \cdot \frac{N_{r.}}{\sum_{b=1}^{C} n_{rb} \cdot w_{rbk}^{(u-1)}}$$

Passo 5 - incrementa o contador de iterações u, fazendo $u = u + 1$; se $u \leq 4$ então repete os passos 3 ou 4 conforme o caso; se $u > 4$ encerra o processo.

Ao final do processo, os totais de características de interesse de pessoas serão estimados usando-se o estimador $\hat{Y}^{PIETOM}$, dado por:

$$\hat{Y}^{PIETOM} = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{k=1}^{n_{rc}} w_{rck}^{(4)} \cdot y_{rck}$$

onde

y_{rck} é o valor da variável de interesse y observado referente à k-ésima pessoa da amostra da célula rc da tabela de pós-estratificação da área de ponderação considerada.

Deve ser notado que, devido ao fato do processo iterativo de estimação ter sido interrompido após a quarta iteração, os pesos $w_{rck}^{(4)}$ representam apenas uma aproximação dos pesos w_{rck} que resultariam da solução exata da minimização da função G^{RU} sujeito as restrições de consistência impostas. Vale ainda notar que, como os pesos não-viciados inicialmente empregados foram constantes para todas as pessoas na área de ponderação considerada (já que $w_{rck}^{(0)} = N/n$ para qualquer k, r e c), os pesos finais $w_{rck}^{(4)}$ são constantes para todas as pessoas de uma dada célula rc da tabela de pós-estratificação, sendo possível, a princípio, descartar o subíndice k.

Isto só não foi feito porque foi empregado um processo de arredondamento dos pesos para que fossem obtidos pesos inteiros, que implicou na obtenção de pesos diferentes para pessoas de uma mesma célula da tabela de pós-estratificação, muito embora tais diferenças nunca possam superar 1 unidade dentro de cada célula da tabela.

É importante aqui notar que o processo de estimação adotado no Censo Demográfico de 1980 pode ser usado para produzir estimativas dos totais de pessoas em cada uma das células da tabela de contingência empregada para definir a pós-estratificação. Chamando de $\hat{N}_{rc}^{PIETOM}$ o número total estimado de pessoas na célula rc da tabela de pós-estratificação na área de ponderação considerada, para $r=1,\ldots,R$ e $c=1,\ldots,C$, pode-se obter $\hat{N}_{rc}^{PIETOM}$ por:

$$\hat{N}_{rc}^{PIETOM} = \sum_{k=1}^{n_{rc}} w_{rck}^{(4)} \quad \text{para } r=1,\ldots,R \text{ e } c=1,\ldots,C.$$

Caso não se tivesse empregado arredondamento dos pesos $w_{rck}^{(4)}$ para pesos inteiros, é possível mostrar que $\hat{N}_{r.}^{PIETOM} = N_{r.}$ para $r=1,\ldots,R$, isto é, o processo de estimação adotado garante consistência perfeita com os totais marginais das linhas da tabela de pós-estratificação adotada. Além disso, apesar de não garantir a consistência exata dos totais marginais das colunas da tabela de pós-estratificação, o estimador de $\hat{N}_{.c}^{PIETOM}$ aproxima bastante bem os totais das marginais de coluna conhecidos $N_{.c}$, isto é, se pode afirmar que $\hat{N}_{.c}^{PIETOM} \cong N_{.c}$ para $c=1,\ldots,C$. Por outro lado, o processo não garante consistência perfeita das estimativas de totais de pessoas dentro de cada célula rc da tabela (isto é, o processo não garante que $\hat{N}_{rc}^{PIETOM} = N_{rc}$), embora utilize as informações disponíveis para o total da população em cada célula da tabela durante o processo de ajuste dos pesos. Em particular, é digno de nota o caso das células não vazias na população mas sem amostra (células não vazias sem amostra), para as quais ocorre que $N_{rc} > 0$ com $n_{rc} = 0$, situação em que a "consistência" (igualdade) da estimativa amostral com a contagem populacional conhecida é impossível.

No Censo Demográfico de 1980 as variáveis auxiliares utilizadas para definir a pós-estratificação de pessoas foram:

a) sexo, situação do domicílio (urbano ou rural) e idade (12 grupos etários), cujos cruzamentos definiram as linhas da tabela de pós-estratificação; e

b) condição de presença (morador ou não morador), espécie do domicílio (particular ou coletivo), condição no domicílio (chefe ou não chefe) e tamanho do domicílio (em 6 categorias), cujos cruzamentos definiram as colunas da tabela de pós-estratificação.

Com isso, a tabela de pós-estratificação inicialmente formada para cada área de ponderação tinha 48 linhas (R = 48) e 15 colunas (C = 15) no caso da expansão da amostra de pessoas, totalizando 720 células. No caso da expansão de domicílios, esses números eram de 16 linhas e 20 colunas, totalizando 320 células[8]. De forma semelhante ao que ocorreu nos processos de estimação dos Censos de 60 e 70, foram definidas algumas condições de controle sobre os tamanhos dos pós-estratos inicialmente formados em cada área de ponderação.

Entretanto, como a pós-estratificação passou a ser definida por uma tabela de contingência de 2 dimensões, as condições de controle do tamanho tiveram que ser aplicadas de forma a garantir que os totais de unidades da amostra nas marginais dessa tabela de pós-estratificação fossem satisfatórios. Assim, foi efetuado reagrupamento de linhas ou de colunas da tabela de pós-estratificação definida inicialmente em cada área de ponderação sempre que ocorresse uma das condições:

a) se $n_{r.} < 5$ ou $N_{r.}/n_{r.} > 8$ reagrupava-se a linha r com uma das linhas adjacentes;

b) se $n._c < 5$ ou $N._c/n._c > 8$ reagrupava-se a coluna c com uma das colunas adjacentes.

---

[8]As tabelas de pós-estratificação iniciais de pessoas e de domicílios usadas no Censo Demográfico de 1980 podem ser encontradas na Metodologia do Censo Demográfico de 1980, páginas 211 e 225, respectivamente.

Esse processo de reagrupamento implicou que o número de células, linhas e colunas da tabela de pós-estratificação final empregada variasse de uma área de ponderação para outra, tal como fora o caso da expansão da amostra dos Censos de 60 e 70. Porém, como o número de pós-estratos inicialmente definidos aumentou muito (de 46 para 720) do Censo de 70 para o de 80, aumentou também a variabilidade do número final de pós-estratos efetivamente empregado no cálculo dos pesos para a expansão da amostra.

Um dos motivos que levou à adoção do PIETOM como processo de estimação para os dados da amostra do Censo 80 foi exatamente o fato de que são empregadas condições restritivas sobre o tamanho dos pós-estratos apenas nas marginais da tabela de pós-estratificação, permitindo assim que um número muito maior de pós-estratos seja formado.

Entretanto, esse é também um dos pontos fracos dessa metodologia. O processo de estimação no qual se baseia o PIETOM, o RREP, foi proposto para estimar N_{rc} em situações nas quais esses valores não são conhecidos (estimação em tabelas de contingência), o que não é o caso do Censo, no qual tais valores são conhecidos. Além disso, o emprego de condições restritivas apenas nas marginais da tabela de pós-estratificação permite a ocorrência de situações nas quais $n_{rc} = 0$ com $N_{rc} > 0$, ou então de células com valores de n_{rc} muito baixos (células rarefeitas na amostra).

Como o processo de reagrupamento da tabela só dependia das condições restritivas sobre as marginais, podem ter ocorrido em algumas áreas de ponderação casos de tabelas com amostras rarefeitas ou nulas em algumas células. A solução considerada ideal teria sido a aplicação de reagrupamento das células individuais da tabela, mas então se estaria retornando ao processo de estimação do Censo de 70 e certamente se teria que abrir mão da intenção de aumentar tanto o número de pós-estratos empregados na expansão, além do que se estaria complicando o processo de reagrupamento.

Com a intenção de contornar esse problema, foi sugerido por Costa(1987a) um método denominado Grupo de Controle Iterativo (GCI) que é analisado com detalhes na seção 3 desse texto. Outra alternativa é um método híbrido proposto Oh e Scheuren(1987), que combina o uso de estimadores de pos-estratificação usuais nas células não-rarefeitas da tabela de pós-estratificação com o estimador baseado no "*Raking Univariado*" para as células rarefeitas. Tal processo híbrido de estimação não será aqui considerado.

Uma outra motivação que levou à busca de processos de estimação alternativos ao PIETOM vem do fato de que seu estimador de precisão tem expressão complexa, o que implicou no uso de programas de computador de elaboração e execução caras no Censo de 80. Mais ainda, tais programas foram perdidos, e caso se pretendesse aplicar o PIETOM para a expansão da amostra do Censo Demográfico de 91, teriam que ser elaborados novos programas de estimação da precisão do PIETOM, tarefa complexa e que custou grande esforço na preparação do Censo de 80, ou se teria que investigar métodos aproximados de menor custo e complexidade computacional para a estimação da precisão.

3. PROCESSO DE ESTIMAÇÃO GRUPO DE CONTROLE ITERATIVO

A primeira tentativa de apresentar um método alternativo ao PIETOM para a expansão da amostra do Censo Demográfico de 1991 surge no trabalho de Costa(1987a) que, partindo da mesma tabela de pós-estratificação inicial (48 linhas por 15 colunas) usada para formar os pós-estratos em cada área de ponderação no Censo de 80, propôs um novo método de cálculo para obtenção dos pesos iniciais $w_{rck}^{(0)}$ baseado no estimador de pós-estratificação simples e de ajuste desses pesos iniciais com base nas marginais da tabela, sendo que o ajuste dos pesos iniciais só é efetuado quando ocorre alguma célula não vazia sem amostra na tabela, isto é, alguma célula rc tal que $N_{rc} > 0$ com $n_{rc} = 0$. Costa(1987a) batizou esse novo método de "Grupo de Controle Iterativo" (GCI).

O método GCI é, na verdade, uma modificação da pós-estratificação simples no caso em que ocorrem células não vazias sem amostra, com um esquema de ajuste dos pesos das outras células pelas marginais similar ao do PIETOM, destinado a distribuir as unidades da população nas células não vazias sem amostra nas células onde essa situação não ocorre. Segundo Costa(1987a), esse método foi concebido com a finalidade de assegurar as vantagens conferidas pela pós-estratificação mais detalhada adotada pelo PIETOM e de aumentar (comparativamente ao PIETOM) a consistência das estimativas dos totais de pessoas em cada célula ($\hat{N}_{rc}$) com os totais conhecidos da população (N_{rc}). Vale destacar que o PIETOM não busca essa consistência das estimativas nas células da tabela de pós-estratificação, mas tão somente nas marginais de tal tabela.

O método GCI proposto por Costa(1987a) supõe também a ocorrência de reagrupamentos nas linhas ou colunas da tabela de pós-estratificação inicial formada em cada área de ponderação, devido à imposição das mesmas condições sobre os tamanhos das células marginais da tabela que foram impostas no Censo 80 quando o PIETOM foi empregado. Tais condições foram:

a) se $n_{r.} < 5$ ou $N_{r.}/n_{r.} > 8$ a linha r deve ser agrupada com uma das linhas adjacentes;

b) se $n._c < 5$ ou $N._c/n._c > 8$ a coluna c deve ser agrupada com uma das colunas adjacentes.

Ainda segundo Costa(1987a), o método GCI seria também uma forma de contornar o problema que surge na pós-estratificação baseada na tabela de duas entradas do PIETOM quando da ocorrência de células não vazias sem amostra, causado ou pelo esgotamento das regras de reagrupamento ou pelo fato de que as condições de tamanho são aplicadas a nível das marginais como no PIETOM. Segundo aquele autor, o GCI permitiria tirar proveito da pós-estratificação baseada na tabela de duas entradas mais plenamente que no PIETOM.

O estimador de total da variável de interesse y, baseado no método GCI, é definido de tal forma que os pesos iniciais são obtidos através do estimador da pós-estratificação, quando possível, e posteriormente esses pesos são ajustados pelas marginais como nas iterações 1 e 2 do PIETOM. Assim, o estimador $\hat{Y}^{GCI}$ do total Y da variável de interesse y a nível da área de ponderação é definido por:

$$\hat{Y}^{GCI} = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{k=1}^{n_{rc}} w_{rck}^{GCI} \cdot y_{rck}$$

onde

$$w_{rck}^{GCI} = w_{rck}^{(2)}$$

$w_{rck}^{(u)}$ é o valor do peso na iteração u, para $u=0,1,2$.

$$w_{rck}^{(o)} = \begin{cases} \dfrac{N_{rc}}{n_{rc}} & \text{se } n_{rc} > 0 \\ 0 & \text{se } n_{rc} = 0 \end{cases}$$

$$w_{rck}^{(1)} = w_{rck}^{(o)} \cdot \frac{N_{.c}}{\sum_{a=1}^{R} n_{ac} \cdot w_{ack}^{(o)}}$$

$$w_{rck}^{(2)} = w_{rck}^{(1)} \cdot \frac{N_{r.}}{\sum_{b=1}^{C} n_{rb} \cdot w_{rbk}^{(1)}}$$

para todo $k=1,\ldots,n_{rc}$, $r=1,\ldots,R$, e $c=1,\ldots,C$.

Vale observar que, no caso particular de tabelas de pós-estratificação que não possuam células não vazias sem amostra, o estimador do GCI coincide com o estimador de pós-estratificação usual. Um outro aspecto que vale a pena comentar sobre o estimador do GCI é o fato de que as células rarefeitas na amostra (células com valores muito pequenos de n_{rc}, digamos $0 < n_{rc} \leq 5$), até mesmo células com uma única unidade na amostra, terão pesos calculados pelo estimador da pós-estratificação simples, fazendo prever nesses casos que, mesmo que a consistência a nível dessas células seja assegurada, a precisão das estimativas para as contagens por células seja muito pequena.

Em Costa(1987a) foi feita uma aplicação e análise comparativa de três métodos alternativos para a expansão da amostra de 25% do Censo Demográfico de 1980 de 4 municípios do Estado do Espírito Santo, a saber: pós-estratificação simples (ou Grupos de Controle), PIETOM e GCI. Em Costa(1989) foi proposta e em Costa e Lima(1991) foram analisados os resultados da aplicação do PIETOM e do GCI para a expansão da amostra de 10% do Censo Demográfico Experimental de Limeira de 1988, considerando 9 áreas de ponderação simuladas com tamanhos variados e 6 tabelas de pós-estratificação alternativas.

Dos experimentos realizados na tentativa de comparar o GCI com o PIETOM, Costa e Lima(1991) apresentaram como conclusão as seguintes recomendações:

- a utilização do método GCI para a expansão da amostra do Censo Demográfico de 91, em substituição ao PIETOM;

- o emprego de uma das tabelas de pós-estratificação de pessoas examinadas, e o uso dos pesos dos respectivos chefes para estimar os totais das variáveis de domicílios e de famílias, simplificando o cálculo dos pesos devido ao uso de apenas uma tabela de pós-estratificação, a exemplo do que se fez no Censo de 70;

- a adoção de um limite inferior de 5.000 pessoas na população para a formação de uma área de ponderação, de forma semelhante ao que já fora feito no Censo de 80.

Por outro lado, deve ser considerado que os métodos de estimação usualmente empregados pelo IBGE ou por outras instituições congêneres em situação semelhante (expansão da amostra de coleta dos Censos de população) têm suas propriedades bastante conhecidas e descritas com detalhes na literatura disponível. Em Costa e Lima(1991) são apresentadas as análises efetuadas com a aplicação do estimador GCI aos dados reais do Censo Demográfico Experimental de Limeira de 1988. Tais análises, entretanto, não consideraram o cálculo nem do vício nem de estimativas para a variância do estimador GCI, mesmo que esse cálculo fosse efetuado através do emprego de métodos indiretos de estimação.

Bianchini(1991) investigou as propriedades do estimador baseado no GCI, através da obtenção de expressões para o cálculo do vício e da tentativa de obtenção de expressões para a variância do estimador de total baseado no GCI. A seguir são apresentadas algumas das conclusões que foram alcançadas com essa análise do método GCI:

a) o GCI foi apresentado como método alternativo para solucionar o problema de células não vazias sem amostra ou rarefeitas, sem necessariamente agrupá-las, na tentativa de obter estimativas mais consistentes com as informações dos totais das características investigadas a 100% com os totais conhecidos da população em cada célula do que as obtidas com a aplicação do PIETOM; esse, entretanto, não era um objetivo pretendido pelo PIETOM, nem tampouco pode ser considerado como um objetivo inquestionável a ser perseguido, pois a maior consistência das estimativas das contagens populacionais por célula implicará numa possível perda da precisão do estimador devido à ocorrência das células rarefeitas na amostra; nesse sentido, a proposta do PIETOM, que se baseia na minimização da função perda G^{RU}, é mais razoável e aceitável do que o objetivo pretendido com o GCI, cuja fundamentação teórica não está devidamente explicitada;

b) o estimador definido pelo GCI coincide com o estimador de pós-estratificação nas tabelas onde $n_{rc} > 0 \ \forall$ célula rc com $N_{rc} > 0$;

c) o peso final calculado pelo GCI é resultado de um ajuste dos pesos iniciais da pós-estratificação simples de cada célula com amostra (cujo $n_{rc} > 0$) tal que, ao final, os pesos de <u>todas</u> as células da tabela são alterados para resolver a questão de algumas células não vazias sem amostra; neste aspecto, o GCI difere do procedimento adotado para solucionar este tipo de problema com o estimador de pós-estratificação simples, que procura agregar células rarefeitas com células vizinhas, de acordo com algum critério definido pelo pesquisador, normalmente escolhidas para manter homogeneidade interna na célula agrupada, o que altera os pesos iniciais apenas para as células que foram agregadas e, portanto, mantém para as demais células a consistência com as informações de universo; com o GCI, pode-se ter uma situação na qual, por exemplo, para a variável relação com o chefe do domicílio, se ocorrer $n_{rc} = 0$ e $N_{rc} \neq 0$ para alguma célula dos não chefes, então o total estimado de chefes será afetado pelo ajuste dos pesos, podendo ser inconsistente com o valor do número de chefes no universo; logo, a busca de consistência para cada célula da tabela pode custar a perda de consistência em alguma das marginais, devido à natureza do processo de ajuste empregado pelo GCI;

d) a magnitude do ajuste dos pesos em uma tabela de pós-estratificação qualquer é função da dimensão da tabela e da freqüência de células com valores nulos na amostra e com informação no universo, bem como do número de unidades da população existentes nessas células;

e) uma questão que não foi avaliada com respeito à definição do estimador GCI proposto por Costa(1987a) é o porque da parada na iteração 2; o método GCI, tal como definido, assegura consistência na marginal das linhas, mas não foi examinada por Costa em seus trabalhos a questão da convergência do método caso o processo iterativo de ajuste prosseguisse, nem a qualidade da aproximação da solução obtida após a segunda iteração;

f) o GCI gera um estimador viciado para estimar o total de uma determinada característica, embora o PIETOM também gere;

g) o GCI pode ser expresso em termos de um estimador de razão combinada a partir da iteração 1, sendo que os estimadores que definem o numerador e denominador da razão são viciados, o que complica substancialmente a obtenção da expressão para o cálculo do vício do estimador nas iterações 1 e 2; por outro lado, no caso do PIETOM, os estimadores que definem o numerador e denominador da razão são não viciados, uma das razões pelas quais as expressões do vício do estimador obtido pelo GCI e PIETOM são tão distintas;

h) para obtenção de um estimador razoável para a variância do estimador GCI do total nas iterações 0 e 1, para particulares amostras, seria necessário recorrer a agregações de linhas na tabela "final" de pós-estratificação usada sempre que ocorrerem contagens amostrais nulas ou de apenas 1 unidade em alguma célula dessa tabela; tal procedimento nos leva à questão do critério de reagrupamento levantada com a adoção do estimador de pós-estratificação, porém neste caso seriam necessários novos reagrupamentos de pós-estratos (ou células da tabela de pós-estratificação) para o cálculo do estimador da variância, diferentemente do estimador de pós-estratificação simples, cujo processo de reagrupamento é realizado uma só vez e serve para o cálculo do estimador do total e do estimador de variância correspondente; de certo modo, a idéia de um segundo reagrupamento dos pós-estratos é incoerente com a filosofia de criação do estimador GCI, que só é necessário quando ocorrem células não vazias sem amostra, e porque a filosofia em que esse método se baseia rejeita, a princípio, a estratégia de reagrupar pós-estratos para fazer estimação;

i) não foi possível obter um estimador para a variância do estimador GCI do total após a iteração 2, o que impede a avaliação da precisão desse estimador ao término do ciclo de iterações proposto e sua conseqüente comparação com a variância do PIETOM.

A análise de Bianchini(1991), cujos principais resultados se acham resumidos acima, conduziu à rejeição da idéia de adotar o método GCI para a expansão da amostra do Censo Demográfico de 1991. A conclusão pela não adoção desse método e o interesse em adotar uma solução ainda melhor que o PIETOM implicaram na abertura de novas frentes de estudo, e na pesquisa de novas alternativas para a metodologia do processo de expansão a serem consideradas para a tomada de decisão com vistas à expansão da amostra do Censo Demográfico de 1991. Na seção a seguir se relata os esforços realizados nesse sentido.

4. UM RESUMO DAS ALTERNATIVAS DE MÉTODOS DE ESTIMAÇÃO CONSIDERADAS

Após a análise do método GCI efetuada por Bianchini(1991), concluiu-se que não seria razoável o emprego desse método para a expansão da amostra do Censo Demográfico de 1991, e decidiu-se investigar, de maneira mais ampla, que outros métodos alternativos poderiam ser considerados para essa finalidade.

Num primeiro esforço, foram obtidas na literatura algumas referências tais como Oh e Scheuren(1987), que apresentam modificações do processo RREP de modo a melhorar o tratamento das células rarefeitas, mas mantendo a filosofia de estimação daquele processo. Como já foi mencionado anteriormente, a análise do método proposto por Oh e Scheuren(1987) não recomendou a adoção do mesmo, devido se tratar de método de implementação mais complexa em termos computacionais e devido não resolver de forma satisfatória o problema das células rarefeitas, a nosso ver.

Numa outra frente de trabalho, a partir de um seminário apresentado por Luiz Nery da Costa na Escola Nacional de Ciências Estatísticas sobre o método GCI, quando houve um questionamento por parte de Djalma Pessoa sobre como haviam sido escolhidas as variáveis e os cruzamentos para a definição da tabela de pós-estratificação do Censo Demográfico de 1980 (a mesma tabela que servira de base ao primeiro trabalho de Costa sobre a comparação do método GCI com o PIETOM), decidiu-se investigar a metodologia empregada para definir a pós-estratificação.

De fato, as informações disponíveis sobre a experiência do Censo Demográfico de 1980 dão conta de que a definição da tabela de pós-estratificação adotada foi feita de forma arbitrária, levando-se em conta as variáveis com dados disponíveis para 100% das pessoas ou domicílios (conforme o caso), procurando-se incluir na regra de formação dos pós-estratos pelo menos as mesmas variáveis já consideradas no Censo de 70, e procurando-se adicionar variáveis para as quais havia interesse em garantir consistência das estimativas amostrais com as contagens conhecidas na população.

Além dessas, decidiu-se incluir algumas variáveis especificamente com a intenção de aumentar o controle sobre algumas tendenciosidades detectadas na aplicação do mecanismo de seleção da amostra durante a coleta dos dados[9], caso específico da variável tamanho do domicílio (em número de moradores).

Apesar dessa arbitrariedade na escolha das variáveis que foram usadas para definir a tabela de pós-estratificação ser plenamente justificável por incorporar o conhecimento substantivo dos pesquisadores a respeito dos dados coletados no Censo, não se pode igualmente justificar a mesma arbitrariedade usada na definição de como essas variáveis foram usadas para formar a tabela de pós-estratificação, cuja especificação completa depende da indicação das variáveis que definem as linhas e as colunas, o que também define completamente quais as marginais e cruzamentos das variáveis auxiliares cuja consistência se vai privilegiar no processo de ajuste dos pesos.

Talvez essa forma arbitrária de definição das marginais que foram privilegiadas no processo de ajuste dos pesos tenha resultado do fato de não se dispor, à época do planejamento do Censo de 1980, de programas computacionais adequados e de fácil manejo para analisar a estrutura de dependência das variáveis auxiliares disponíveis, a fim de otimizar a escolha efetuada.

Acontece que esse não é mais o caso. Hoje em dia estão disponíveis programas de computador bastante poderosos e fáceis de manejar que permitem analisar a estrutura de dependência das variáveis auxiliares, visando facilitar a definição da tabela de pós-estratificação. Uma outra razão para considerar o uso de métodos automatizados para a definição da pós-estratificação é o fato de que a mesma precisa ser aplicada a um grande número de áreas de ponderação distintas (no Censo de 80 eram 4.219 ao todo).

---

[9]No Censo de 80, verificou-se que houve tendência de alguns recenseadores a pesquisar com o questionário da amostra uma proporção maior de domicílios com menor número de moradores, talvez motivados por um desequilíbrio do sistema de pagamento e dos preços fixados para cada tipo de questionário.

Não parece razoável supor que a mesma pós-estratificação inicial seja "boa" para todas as áreas de ponderação definidas a partir da base geográfica, devido à grande variabilidade apresentada por essas áreas em relação a seu tamanho e características de desenvolvimento econômico e social. Uma evidência clara disto foi a adoção, no Censo 80, de um procedimento "ad hoc" para reagrupamento de linhas e colunas da tabela de pós-estratificação, necessário para assegurar que as contagens nas marginais satisfizessem certas condições de tamanho. Tal procedimento de reagrupamento pode ser visto como uma tentativa de ajustar a pós-estratificação às características de cada área de ponderação, mediante um processo de contração (agregação) dos pós-estratos inicialmente definidos, numa espécie de "*backward elimination*"[10] dos cruzamentos definidos de início.

Acontece que outros métodos diferentes poderiam ser considerados para essa finalidade de ajustar a pós-estratificação às características de cada área de ponderação, e essa foi uma das linhas de investigação que se decidiu seguir na preparação da metodologia para a expansão da amostra do Censo Demográfico de 1991. Devido ao grande número de áreas de ponderação que se espera utilizar na expansão da amostra, decidiu-se buscar métodos para formar os pós-estratos que fossem de implementação computacional fácil ou que já estivessem implementados em programas de computador acessíveis, para viabilizar a automação da tarefa de formação dos pós-estratos para a expansão em cada área de ponderação.

Assim, a primeira alternativa considerada foi a metodologia denominada *CART - Classification And Regression Trees* - que se acha descrita com detalhes no livro de Breiman e outros(1984) e no artigo de Bloch e Segal(1989), e que se acha implementada no pacote (sistema) *CART*[11], adquirido recentemente pelo IBGE a nosso pedido.

---

[10]Isto é, um procedimento que parte de uma pós-estratificação mais detalhada para depois ir agregando os pós-estratos inicialmente formados, de modo que a tabela de pós-estratificação passe a satisfazer todas as condições de tamanho especificadas.

[11]Ver, a respeito, An Introduction to CART Methodology e Using the CART programs: version 1.3.

Ainda não estão disponíveis resultados da aplicação desse programa aos dados do Censo Demográfico Experimental de Limeira de 1988, com os quais está sendo testada essa metodologia para formação dos pós-estratos, mas esta é uma das linhas de trabalho que se vai seguir até a especificação completa da metodologia de expansão da amostra do Censo Demográfico de 1991.

Uma das principais razões da escolha do CART como alternativa para automação da definição dos pós-estratos está no fato de que essa metodologia e pacote permitem o tratamento simultâneo de variáveis categóricas (ou qualitativas) e numéricas (ou quantitativas) na definição dos pós-estratos, o que é uma das complicações do caso do Censo Demográfico, em que variáveis de ambos os tipos estão disponíveis e se tem interesse que sejam usadas na definição da pós-estratificação.

A idéia básica da metodologia CART é formar uma "árvore" cujos "galhos" formem uma partição da amostra (e da população), definindo assim os pós-estratos a serem usados em cada área de ponderação. Na formação da árvore, emprega-se a técnica de partições binárias a cada passo, e se adota uma regra de parada (que impede novas subdivisões) baseada na variabilidade restante dentro de cada "galho" e num número mínimo aceitável de unidades (pessoas ou domicílios) da amostra (ou da população) em cada "galho", número esse que pode ser arbitrado pelo pesquisador.

Para fazer a partição num dado passo, considera-se as unidades de um certo "galho" e procura-se escolher a variável auxiliar cuja partição binária mais contribui para a redução da variabilidade interna restante de uma variável resposta nos 2 galhos formados pela subdivisão do galho anterior. A aplicação sucessiva da regra de partição binária aos "galhos" que vão sendo formados, sempre condicionada à verificação da regra de parada, vai gerar a partição final da amostra (ou da população) nos "galhos" que, em seu conjunto, definem a "árvore", que traz implícita uma regra de classificação das unidades.

A idéia básica com a qual trabalhamos[12] é aplicar o CART aos dados da amostra de cada área de ponderação levando-se em conta o elenco mais completo possível de variáveis auxiliares disponíveis para formação dos pós-estratos, e gerar a partir da amostra uma regra de classificação que subdivida a amostra nos pós-estratos relevantes para a expansão. Em seguida, essa mesma regra de classificação seria aplicada aos dados da população para gerar a partição da população nos mesmos pós-estratos, o que permitiria efetuar as contagens populacionais nos pós-estratos, indispensável para o cálculo dos pesos do estimador de pós-estratificação simples, por exemplo.

Um exemplo típico do que seria uma árvore de classificação elaborada pelo CART com dados fictícios semelhantes aos que se coletou a 100% no Censo Demográfico de 91 é apresentado na Figura 1 a seguir.

Como essa metodologia (CART) se acha implantada num pacote de computador, que depende apenas da especificação de alguns parâmetros para ser aplicado a um certo conjunto de dados, imagina-se que seu emprego permitirá automatizar a definição da pós-estratificação dentro de cada área de ponderação, sendo necessário apenas definir, para o conjunto das áreas de ponderação, quais as variáveis auxiliares e qual a variável resposta que se pretende considerar na definição dos pós-estratos.

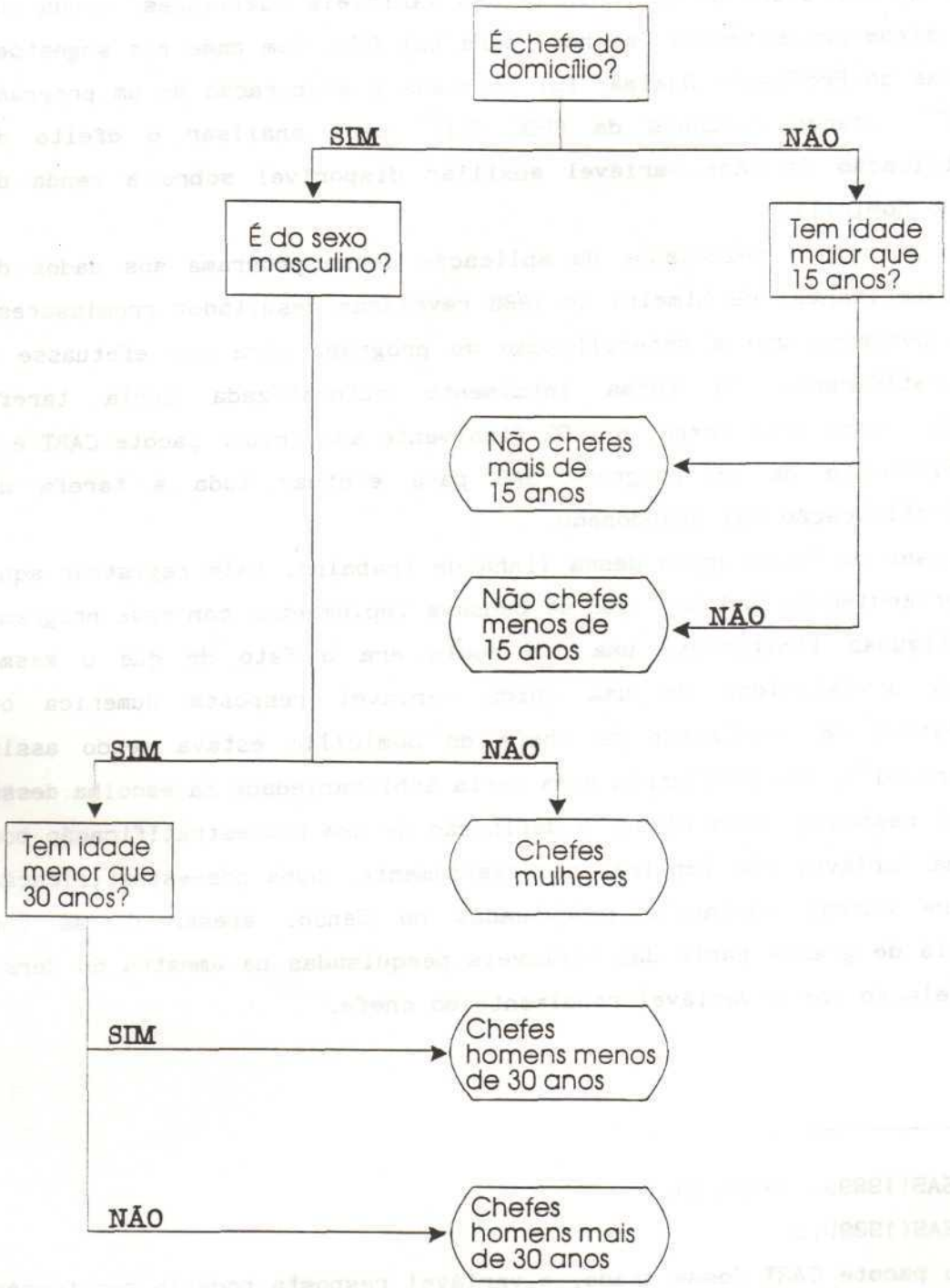
O uso desse pacote e metodologia vai permitir, também, solucionar o problema dos pós-estratos rarefeitos que persiste na metodologia do PIETOM ou do GCI, além de retirar, em grande parte, a arbitrariedade da definição dos pós-estratos. A única arbitrariedade que permaneceria seria a da escolha das variáveis auxiliares e da variável resposta a serem consideradas na pós-estratificação, mas mesmo essa escolha poderia ter sua arbitrariedade minimizada caso se adotasse uma opção conservadora do tipo usar todas as variáveis auxiliares disponíveis, como já sugerido. Quanto à variável resposta, não seria possível eliminar a arbitrariedade na sua escolha, mas pelo menos se poderia avaliar com precisão os efeitos da estratificação no sentido de redução da sua variabilidade. Uma escolha adequada permitiria alcançar a redução da variabilidade para variáveis relacionadas com a variável resposta.

---

[12]Ver o texto preliminar de Albieri(1992).

FIGURA 1

ÁRVORE DE CLASSIFICAÇÃO
EXEMPLO FICTÍCIO



Uma outra alternativa de método considerada para a definição da pós-estratificação foi sugerida aos autores pelo Professor Djalma Pessoa da ENCE, quando ainda não havia sido adquirido pelo IBGE o pacote CART. Segundo a sugestão do Prof. Djalma, mesmo na falta de um suporte computacional mais adequado valeria a pena adotar algum esquema de análise da estrutura de dependência das variáveis auxiliares, ainda que usando algum procedimento "ad hoc" para tal fim. Com base nas sugestões recebidas do Professor Djalma, foi iniciada a elaboração de um programa em SAS[13], usando recursos da PROC GLM[14] para analisar o efeito da estratificação de cada variável auxiliar disponível sobre a renda do chefe do domicílio.

Os primeiros resultados da aplicação desse programa aos dados do Censo Experimental de Limeira de 1988 revelaram resultados promissores, mas se percebeu que a generalização do programa para que efetuasse a pós-estratificação de forma totalmente automatizada seria tarefa complexa. Nesse meio tempo, o IBGE finalmente adquiriu o pacote CART e o desenvolvimento de um programa SAS para efetuar toda a tarefa de pós-estratificação foi abandonado.

Apesar da interrupção dessa linha de trabalho, vale registrar aqui que o procedimento "ad hoc" que se pensava implementar com esse programa tinha algumas limitações, uma das quais era o fato de que o mesmo obrigava à definição de uma única variável resposta numérica ou quantitativa (o rendimento do chefe do domicílio estava sendo assim considerado)[15], que implicaria numa certa arbitrariedade na escolha dessa variável resposta. Além disso, a definição de uma pós-estratificação boa para uma variável não implica, necessariamente, numa pós-estratificação boa para outras variáveis pesquisadas no Censo, apesar de se ter evidência de grande parte das variáveis pesquisadas na amostra do Censo terem relação com a variável rendimento do chefe.

---

[13] Veja SAS(1989a, 1990a,b).

[14] Veja SAS(1989b,c).

[15] Caso o pacote CART fosse usado, a variável resposta poderia ser também do tipo qualitativo ou categórico, o que daria maior liberdade de escolha ao pesquisador para a eleição da variável resposta a ser usada.

Uma outra limitação era o fato de que as variáveis explicativas que o programa permitia considerar tinham que ser todas do tipo categórico (no caso de variáveis quantitativas como a idade, era necessário transformá-las em categóricas mediante prévio agrupamento dos valores em classes), não sendo permitido o uso de combinações de variáveis numéricas com categóricas, como permite o CART.

Além da questão da definição da pós-estratificação, as investigações prosseguiram também com a busca de novos procedimentos de estimação, e foram encontradas na literatura algumas indicações importantes. Primeiramente, foi encontrado um procedimento de estimação denominado *LINWEIGHT* por seus autores - Bethlehem e Keller(1984, 1985, 1987) - que está baseado num modelo linear que relaciona as variáveis de interesse (para as quais só se tem os dados da amostra) com as variáveis auxiliares (disponíveis a 100%). Esse procedimento foi inclusive implementado num programa de computador desenvolvido no *Central Bureau of Statistics* da Holanda, programa esse que recebeu o nome de *LINWEIGHT*[16]. Mais tarde, esse procedimento foi reimplementado em outro programa de computador mais geral, denominado *BASCULA*[17].

De fato, o procedimento de estimação LINWEIGHT é baseado num estimador geral de regressão[18], e permite incorporar, como caso particular, o estimador pós-estratificação simples, se as variáveis auxiliares utilizadas forem do tipo categórico e se forem usadas de modo que a interação de mais alto grau seja considerada (o cruzamento de todas as variáveis auxiliares definindo os pós-estratos). Mas um aspecto atraente desse processo de estimação é que permite tratar casos mais interessantes, como o uso de estratificação multivariada incompleta, tal como definida por Bethlehem e Keller(1984).

---

[16]Veja Bethlehem(1985).

[17]Ver Göttgens e outros(1991).

[18]Tal como considerado por Särndal, Swensson e Wretman(1988).

A literatura disponível sobre a metodologia LINWEIGHT dava conta de que o programa desenvolvido para sua implementação poderia ser obtido junto ao *Central Bureau of Statistics* da Holanda, e que esse programa serviria não só para calcular os estimadores e pesos respectivos dada uma pós-estratificação qualquer, como também poderia ser usado para ajudar a selecionar a pós-estratificação "ótima" a ser usada. Inclusive, essa pós-estratificação poderia ser multivariada incompleta. Mas quando o órgão que desenvolveu esse programa foi contatado para a cessão do mesmo ao IBGE, informou que o programa LINWEIGHT não poderia ser cedido porque estava sendo substituído por outro programa, denominado BASCULA, que eles estavam desenvolvendo.

Nessa oportunidade, foi obtida uma cópia do programa BASCULA, mas se verificou que esse programa apresentava algumas limitações que dificultavam considerá-lo uma alternativa para a expansão da amostra do Censo Demográfico de 1991. A primeira se devia ao fato de que o programa foi desenvolvido para uso em microcomputadores tipo IBM-PC, fazendo prever que dificilmente seria aplicável para expandir a amostra do Censo brasileiro por motivos de eficiência, já que o planejamento do sistema de processamento do Censo considera primordialmente o uso de computadores de porte muito maior, devido ao gigantesco volume de dados a serem processados[19]. A segunda limitação, mais séria, é devida ao fato de que a versão do programa que foi cedida ao IBGE não dispõe de recursos para calcular estimativas de precisão (ou do erro amostral), o que impediria de usar esse mesmo programa para executar toda a estimação necessária. A terceira limitação, comparativamente ao programa LINWEIGTH, se deve ao fato de que as facilidades para auxiliar na escolha da pós-estratificação que aquele programa dispunha não estão incorporadas na atual versão BASCULA, o que obrigaria a empregar algum outro programa para resolver essa questão antes da aplicação do BASCULA.

---

[19]São cerca de 35.000.000 de domicílios na população, 4.000.000 de domicílios na amostra, 145.000.000 de pessoas na população e 15.000.000 de pessoas na amostra.

Assim, muito embora se tenha conseguido uma cópia do programa BASCULA para teste e experimentação, logo se decidiu que o mesmo não seria considerado para a expansão da amostra do Censo Demográfico de 1991, devido a todas as limitações identificadas quanto à sua aplicação nesse caso. Entretanto, deve ser registrado que o programa tem algumas características interessantes, tais como dispor de recursos para aplicar vários estimadores diferentes (pós-estratificação simples, estimadores baseados no RREP e estimadores de regressão baseados na metodologia LINWEIGHT), ser de fácil manuseio e estar disponível para uso em computadores tipo IBM-PC, geralmente acessíveis. Por isso, considera-se que esse programa pode ser uma alternativa interessante a ser considerada para a estimação em pesquisas de menor porte, em que o volume de dados não seja tal que dificulte o uso de microcomputadores para o processamento.

Foi após toda essa busca que se tomou conhecimento, durante visita técnica à *Business Surveys Methods Division* do *Statistics Canada*[20], de uma outra metodologia alternativa para a expansão da amostra do Censo Demográfico de 1991. Trata-se da metodologia denominada *GLSEP - Generalized Least Squares Estimation Procedure* - desenvolvida por Bankier(1989,1990) para aplicação na expansão da amostra do Censo Demográfico canadense de 1991.

As características dessa metodologia a tornaram desde logo uma opção atraente a considerar no caso brasileiro, e seu estudo e análise motivaram sua eleição como principal alternativa para a expansão da amostra do Censo 91, além do que foram cedidas pelo *Statistics Canada* cópias dos programas desenvolvidos para implementação dessa metodologia[21]. O próximo capítulo é dedicado à exposição das características dessa metodologia e à apresentação detalhada da metodologia proposta para ser aplicada na expansão da amostra do Censo Demográfico de 1991.

---

[20]Veja Silva(1991).

[21]Tais programas foram elaborados por Michael Bankier, Stephen Rathwell, e Susana Rubin, e foram cedidos ao IBGE em janeiro de 1992.

5. A METODOLOGIA DE ESTIMAÇÃO PROPOSTA PARA O CENSO 91

5.1 Introdução

Conforme indicado anteriormente, este capítulo dedica-se à descrição da metodologia desenvolvida por Bankier(1989, 1990), denominada *Generalized Least Squares Estimation Procedure - GLSEP*, que está sendo aplicada para a expansão da amostra do Censo de População canadense de 1991 ora em andamento, e que se propõe adotar também para expandir a amostra do Censo Demográfico brasileiro de 1991. As razões principais para a recomendação de adoção dessa metodologia serão examinadas mais adiante neste capítulo, após a apresentação da descrição detalhada da mesma. A descrição aqui apresentada baseia-se diretamente em Bankier(1990) e em Rathwell(1991).

Será apresentada primeiramente uma descrição genérica da metodologia GLSEP para o caso univariado (a fim de introduzir as idéias básicas de forma facilmente comparável com as metodologias já examinadas antes, especialmente o RREP) e para o caso multivariado (que será a versão efetivamente proposta para aplicação), para depois fornecer detalhes que ajudem a compreender a proposta específica de aplicação feita para o caso do Censo Demográfico brasileiro de 1991. Ao final, são indicadas algumas condições necessárias para a aplicação dessa metodologia na expansão da amostra do Censo 91.

5.2 Descrição Geral da Metodologia GLSEP para o Caso Univariado

A primeira descrição da metodologia GLSEP será feita para o caso de uma função de perda univariada. Aqui, como no caso das metodologias de expansão já empregadas em censos anteriores, são admitidas algumas hipóteses quanto ao desenho amostral empregado na coleta que são similares às já indicadas na seção 2.1. Assim, em cada área de ponderação, admite-se que uma amostra aleatória sem reposição de tamanho fixo (m ou n para domicílios ou pessoas, conforme o caso) é selecionada da população (de M ou N domicílios ou pessoas, conforme a situação)

Admite-se também que cada unidade da população tem uma probabilidade não nula e conhecida de ser selecionada, e que a probabilidade de seleção de qualquer par de unidades da população também é não nula.

A descrição da metodologia GLSEP apresentada a seguir é feita considerando-se sua aplicação para expandir a amostra de domicílios do Censo, por razões que ficarão claras mais adiante. Apesar disso, vale notar que a descrição apresentada pode ser facilmente adaptada para o caso de sua aplicação na expansão da amostra de pessoas, se isso for necessário.

Admite-se também que as unidades da população (domicílios, doravante) são agrupadas para formar pós-estratos de acordo com uma tabela de contingência de duas entradas, com R linhas e C colunas, tal como na seção 2.3. Finalmente, admite-se que um peso não viciado $w_{rck}^{(0)}$ é atribuído ao k-ésimo domicílio da amostra pertencente à célula rc da tabela de pós-estratificação definida. Mais uma vez, coloca-se como objetivo do problema a determinação de um peso final ajustado, $w_{rck} = a_{rc} \cdot w_{rck}^{(0)}$ (ou dos fatores de ajuste a_{rc}) de tal forma que seja minimizada a função de perda ou de distância G^{GLSU} dada por

$$G^{GLSU} = \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{X}_{rc}^{(0)} \cdot \left(\frac{\hat{X}_{rc} - \hat{X}_{rc}^{(0)}}{\hat{X}_{rc}^{(0)}} \right)^2 =$$

$$= \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{X}_{rc}^{(0)} \cdot (a_{rc} - 1)^2$$

sujeito às restrições de consistência nas marginais da tabela

$\hat{X}_{r.} = X_{r.}$ para r=1,...,R;

$\hat{X}_{.c} = X_{.c}$ para c=1,...,C;

onde

G^{GLSU} é a função de perda correspondente à metodologia de mínimos quadrados generalizados (*Generalized Least Squares*) univariada;

$$X_{rc} = \sum_{k=1}^{M_{rc}} X_{rck}$$

M_{rc} é o número de domicílios da população na célula rc da tabela de pós-estratificação da área de ponderação considerada;

X_{rck} é o valor da variável auxiliar x para o k-ésimo domicílio da população na célula rc da tabela de pós-estratificação da área de ponderação considerada;

$$\hat{X}_{rc}^{(0)} = \sum_{k=1}^{m_{rc}} w_{rck}^{(0)} \cdot x_{rck}$$

$$\hat{X}_{rc} = \sum_{k=1}^{m_{rc}} w_{rck} \cdot x_{rck}$$

m_{rc} é o número de domicílios da amostra na célula rc da tabela de pós-estratificação da área de ponderação considerada;

x_{rck} é o valor da variável auxiliar x para o k-ésimo domicílio da amostra na célula rc da tabela de pós-estratificação da área de ponderação considerada.

Admite-se, na definição da função de perda G^{GLSU}, que a somação inclui apenas os pós-estratos (células da tabela de contingência) nos quais $\hat{X}_{rc}^{(0)} \neq 0$. Mais uma vez, para X_{rc}, $\hat{X}_{rc}$ e $\hat{X}_{rc}^{(0)}$ a somação sobre algum dos subíndices é indicada através da substituição desse subíndice por um ponto (.).

A principal diferença da função de perda G^{GLSU} para a que dá origem à metodologia RREP (G^{RU}) é que aqui as distâncias das estimativas dos totais da variável auxiliar x a nível de cada célula da tabela de pós-estratificação baseadas nos pesos iniciais $w_{rck}^{(0)}$ ($\hat{X}_{rc}^{(0)}$) para as estimativas finais baseadas nos pesos ajustados w_{rck} ($\hat{X}_{rc}$) são medidas com a distância do qui-quadrado. Por essa razão, o sistema de equações que precisa ser resolvido para determinar os pesos finais w_{rck} (ou os correspondentes coeficientes de ajuste a_{rc}) é linear, como está demonstrado no anexo A.2.1 de Bankier(1990). Por essa razão, é possível dispor de uma solução exata para o problema de minimização de G^{GLSU}, ao contrário do que ocorre com a metodologia RREP, que depende de um processo iterativo de solução. Apesar disso, deve ser notado que também é possível obter a solução da minimização de G^{GLSU} por um procedimento iterativo. Entretanto, a solução da minimização para o caso univariado não será apresentada aqui por não ser de interesse prático imediato, já que se propõe empregar na expansão da amostra do Censo Demográfico 91 a versão multivariada da metodologia GLSEP, que se passa imediatamente a descrever.

5.3 Descrição Geral da Metodologia GLSEP para o Caso Multivariado

Para apresentar essa descrição, admite-se as mesmas hipóteses da seção 5.2 quanto à amostra de domicílios selecionada. Quanto à pós-estratificação usada, porém, admite-se agora que a amostra é particionada em H pós-estratos ou classes de ajustamento de pesos (que podem tomar ou não a forma de uma tabela de pós-estratificação de duas entradas), com m_h domicílios na amostra da h-ésima classe de ajustamento de pesos[22].

Admite-se ainda que um peso não viciado $w_{hk}^{(0)}$ é atribuído ao k-ésimo domicílio da amostra pertencente à h-ésima classe de ajustamento de pesos. Coloca-se como objetivo do problema a determinação de um peso final ajustado, $w_{hk} = a_h \cdot w_{hk}^{(0)}$ (ou dos fatores de ajuste a_h) de tal forma que seja minimizada a função de perda ou de distância G^{GLSM} dada por

$$G^{GLSM} = \sum_{h=1}^{H} \hat{X}_{h\cdot}^{(0)} \cdot \left(\frac{\hat{X}_{h\cdot} - \hat{X}_{h\cdot}^{(0)}}{\hat{X}_{h\cdot}^{(0)}} \right)^2 =$$

$$= \sum_{h=1}^{H} \hat{X}_{h\cdot}^{(0)} \cdot (a_h - 1)^2$$

sujeito às restrições de consistência

$\hat{X}_{\cdot q} = X_q$ para $q=1,\ldots,Q$;

onde

G^{GLSM} é a função de perda correspondente à metodologia de mínimos quadrados generalizados (*Generalized Least Squares*) multivariada;

$$\hat{X}_{h\cdot}^{(0)} = \sum_{q=1}^{Q} \hat{X}_{hq}^{(0)}$$

Q é o número de variáveis auxiliares que definem as restrições de consistência consideradas na expansão;

---

[22] A partir dessa seção, o termo classe de ajustamento de pesos será usado em substituição ao termo pós-estrato para indicar a situação em que a definição dos "pós-estratos" depender de informações disponíveis somente na amostra, o que, a rigor, não é a pós-estratificação conhecida na literatura.

$\hat{X}_{hq}^{(0)} = \sum_{k=1}^{m_h} w_{hk}^{(0)} \cdot x_{hqk}$ é o total estimado da q-ésima variável auxiliar para a h-ésima classe de ajustamento de pesos considerando-se os pesos iniciais não viciados $w_{hk}^{(0)}$;

m_h é o número de domicílios da amostra na h-ésima classe de ajustamento de pesos da área de ponderação considerada;

x_{hqk} é o valor da q-ésima variável auxiliar para o k-ésimo domicílio da amostra da h-ésima classe de ajustamento de pesos da área de ponderação considerada;

$\hat{X}_{h.} = \sum_{q=1}^{Q} \hat{X}_{hq}$

$\hat{X}_{hq} = \sum_{k=1}^{m_h} w_{hk} \cdot x_{hqk}$ é o total estimado da q-ésima variável auxiliar para a h-ésima classe de ajustamento de pesos considerando-se os pesos finais ajustados w_{hk};

$a_h = \hat{X}_{h.} / \hat{X}_{h.}^{(0)}$ é o fator de ajustamento de pesos correspondente à h-ésima classe de ajustamento de pesos;

$\hat{X}_{.q} = \sum_{h=1}^{H} \hat{X}_{hq}$ é o total estimado da q-ésima variável auxiliar para a área de ponderação considerada;

$X_q = \sum_{k=1}^{M} X_{qk}$ é o total populacional (conhecido) da q-ésima variável auxiliar usada na expansão, para a área de ponderação considerada;

M é o número de domicílios da população da área de ponderação considerada, como já foi definido;

X_{qk} é o valor da q-ésima variável auxiliar para o k-ésimo domicílio da população da área de ponderação considerada.

Admite-se, na definição da função de perda G^{GLSM}, que a somação inclui apenas as classes de ajustamento de pesos nas quais $\hat{X}_{h.}^{(0)} \neq 0$.

A definição dessa função de perda multivariada é uma extensão natural do caso univariado quando se decide considerar mais de uma variável auxiliar para controlar a obtenção dos fatores de ajuste dos pesos não viciados iniciais.

Uma importante diferença entre a metodologia GLSEP multivariada e sua versão univariada ou o PIETOM (versão univariada do RREP) é que não precisa haver nenhuma relação entre as variáveis utilizadas para delimitar os pós-estratos ou classes de ajustamento de pesos e as variáveis auxiliares consideradas para definir as restrições de consistência a serem satisfeitas.

No caso do Censo, as variáveis auxiliares utilizadas com a metodologia PIETOM eram definidas como função da pós-estratificação adotada, já que representavam as contagens de pessoas ou de domicílios nas células da tabela de pós-estratificação. Agora, com a metodologia GLSEP multivariada, essa limitação pode ser eliminada, já que é possível utilizar variáveis auxiliares que não foram empregadas para formar as classes de ajustamento de pesos, ou melhor ainda, é possível formar classes de ajustamento de pesos inclusive usando variáveis cujos totais não são conhecidos na população. Um exemplo útil para ilustrar essa situação é a definição de cada domicílio da amostra como uma "classe de ajustamento de pesos", de modo que a metodologia GLSEP vá gerar um peso ajustado distinto para cada domicílio da amostra da área de ponderação considerada. Como se verá mais adiante, esse será o critério proposto para definir a "pós-estratificação" a ser usada para a expansão do Censo Demográfico 91.

A definição da função de perda multivariada G^{GLSM} pode ser reapresentada para tornar mais evidente a natureza multivariada do problema, como feito por Bankier(1990). Assim, observe-se que G^{GLSM} poderia ser reescrita como:

$$G^{GLSM} = \sum_{h=1}^{H} (a_h - 1) \cdot \mathring{X}_{h\cdot}^{(0)} \cdot (a_h - 1) =$$

$$= [(a_1 - 1) \ldots (a_H - 1)] \cdot \begin{bmatrix} \mathring{X}_{1\cdot}^{(0)} & 0 & \ldots & 0 \\ 0 & \mathring{X}_{2\cdot}^{(0)} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \mathring{X}_{H\cdot}^{(0)} \end{bmatrix} \cdot \begin{bmatrix} (a_1 - 1) \\ (a_2 - 1) \\ \vdots \\ (a_H - 1) \end{bmatrix}$$

$$= (\mathbf{a} - \mathbf{1}_H)' \cdot V \cdot (\mathbf{a} - \mathbf{1}_H)$$

onde

$\mathbf{a} = (a_1, a_2, \ldots, a_H)'$ é o vetor dos fatores de ajuste dos pesos para as H classes de ajustamento de pesos;

$\mathbf{1}_H = (1, 1, \ldots, 1)'$ é o vetor unitário de dimensão Hx1;

$$V = \begin{bmatrix} \hat{X}_{1.}^{(0)} & 0 & \dots & 0 \\ 0 & \hat{X}_{2.}^{(0)} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \hat{X}_{H.}^{(0)} \end{bmatrix}$$ é uma matriz simétrica positiva definida que não depende de a.

Bankier(1990) mostra que a função de perda G^{GLSM} assim definida é aproximadamente igual a sua análoga G^{RM} correspondente à versão multivariada da metodologia RREP, sempre que os fatores de ajustamento de pesos a_h não forem muito diferentes de 1. Este resultado está demonstrado no apêndice A.1.3 de Bankier(1990), e implica que os resultados do processo de ajustamento de pesos pelas metodologias RREP ou GLSEP multivariadas, consideradas as mesmas classes de ajustamento de pesos e as mesmas Q variáveis auxiliares, seriam semelhantes.

Bankier(1990) mostra também em seu apêndice A.2.1 que o sistema de equações que precisa ser resolvido para minimizar G^{GLSM} é linear, implicando que existe uma fórmula fechada para a solução exata do problema de minimização dessa função sujeito às restrições de consistência impostas. Bankier(1990) comenta ainda que, apesar da existência de uma solução exata, também é possível adotar um procedimento iterativo para a obtenção da solução, embora essa alternativa não vá ser necessária em nosso caso.

Com essas duas afirmações comprovadas por Bankier(1990), pode-se argumentar que a metodologia GLSEP é preferível à metodologia RREP, pois ambas vão ter resultados equivalentes e a implementação da metodologia GLSEP é muito mais simples devido à existência de uma solução exata para o problema de minimização.

Uma outra questão de interesse é que a formulação matricial apresentada para a função de perda G^{GLSM} pode ser ainda mais generalizada, se a matriz V ali introduzida for considerada apenas como uma matriz diagonal HxH, positiva definida, independente de a.

Um exemplo de definição alternativa para a matriz V é o seguinte:

$$V_p = \begin{bmatrix} (\hat{X}_{1.}^{(0)})^{2p} & 0 & \dots & 0 \\ 0 & (\hat{X}_{2.}^{(0)})^{2p} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & (\hat{X}_{H.}^{(0)})^{2p} \end{bmatrix}$$ onde $0 \leq p \leq 1$.

Essa definição alternativa da matriz $\mathbb{V}$ se reduziria ao caso anterior se $p = 1/2$. Com essa matriz $\mathbb{V}_p$, seria possível privilegiar a manutenção dos fatores de ajuste de pesos a_h mais próximos de 1 para as classes de ajustamento de pesos com valores grandes de $\mathbb{X}_{h.}^{(0)}$ que na situação padrão ($p = 1/2$), mediante a escolha de um valor de $p > 1/2$. Se a situação oposta for de interesse, isto é, se for preferível atenuar a participação das classes de ajustamento de pesos com valores grandes de $\mathbb{X}_{h.}^{(0)}$ na determinação dos fatores de ajuste, então um valor de $p < 1/2$ deveria ser escolhido. Outras alternativas para $\mathbb{V}$ podem ser consideradas, mas para o momento basta saber que essa matriz pode ser definida de forma mais geral, para acomodar situações particulares de interesse prático. Mais tarde se voltará a examinar essa questão.

Bankier(1990) mostra em seu apêndice A.2.1 que a solução exata da minimização da função G^{GLSM} sujeito às restrições de consistência impostas é dada por:

$$\mathbf{a} = \mathbf{1}_H + \mathbb{V}^{-1} \cdot \hat{\mathbb{X}}^{(0)} \cdot \lambda$$

onde

$$\hat{\mathbb{X}}^{(0)} = \begin{bmatrix} \mathbb{X}_{11}^{(0)} & \mathbb{X}_{12}^{(0)} & \cdots & \mathbb{X}_{1Q}^{(0)} \\ \mathbb{X}_{21}^{(0)} & \mathbb{X}_{22}^{(0)} & \cdots & \mathbb{X}_{2Q}^{(0)} \\ \vdots & \vdots & & \vdots \\ \mathbb{X}_{H1}^{(0)} & \mathbb{X}_{H2}^{(0)} & \cdots & \mathbb{X}_{HQ}^{(0)} \end{bmatrix} = [\mathbb{X}_{hq}^{(0)}]$$

é a matriz de estimativas de total das Q variáveis auxiliares para cada uma das H classes de ajustamento de pesos definidas, calculadas utilizando os pesos não viciados iniciais $w_{hk}^{(0)}$;

$$\lambda = (\hat{\mathbb{X}}^{(0)\prime} \cdot \mathbb{V}^{-1} \cdot \hat{\mathbb{X}}^{(0)})^{-1} \cdot (\mathbb{X}'.\mathbf{1}_M - \hat{\mathbb{X}}^{(0)\prime}.\mathbf{1}_H)$$

$\mathbb{X} = [X_{kq}]$ é uma matriz MxQ com os valores das variáveis auxiliares para os M domicílios da população da área de ponderação considerada.

É imediato notar das equações acima que λ é um vetor Qx1. Bankier(1990, seção 7.4.1) demonstra que $\hat{\mathbb{X}}^{(0)\prime} \cdot \mathbb{V}^{-1} \cdot \hat{\mathbb{X}}^{(0)}$ é inversível e portanto, que há uma solução única **a** para a minimização de G^{GLSM}, sempre que $\hat{\mathbb{X}}^{(0)}$ for de posto (*rank*) Q, isto é, se as colunas de $\hat{\mathbb{X}}^{(0)}$ correspondentes às diferentes restrições de consistência impostas para o ajuste dos pesos forem linearmente independentes.

Será admitido, doravante, que essa situação ocorre, isto é, será admitido que o posto da matriz $\hat{X}^{(0)}$ é Q. Esta é uma hipótese razoável pois, caso não fosse verificada, significaria que existiriam colunas linearmente dependentes na matriz $\hat{X}^{(0)}$, o que implicaria na existência de restrições de consistência redundantes (e portanto, desnecessárias e descartáveis).

Na prática, caso sejam detectadas restrições de consistência linearmente dependentes entre as restrições impostas, estas podem ser eliminadas sem prejuízo algum ao processo de estimação ou à manutenção da consistência para os totais estimados dessas variáveis com base na amostra. Bankier(1990) mostra também que, caso ocorram algums fatores de ajuste iguais, as classes de ajustamento de pesos correspondentes podem ser agregadas (*collapsed*), juntamente com as linhas e colunas correspondentes da matriz V, sem qualquer alteração nos resultados dos estimadores, o que implicaria em maior economia computacional. Bankier estuda essa questão com detalhes, inclusive indicando uma maneira computacionalmente viável para determinar, a priori, que pós-estratos poderiam ser agregados. Apesar disso, essa questão não será detalhada aqui, por não ter relevância na aplicação da metodologia GLSEP que se propõe fazer no Censo 91.

Uma das características mais atraentes reveladas pelos estimadores resultantes da aplicação da metodologia GLSEP no caso multivariado advém do fato de que o estimador de total resultante de sua aplicação é na verdade um estimador de regressão múltipla, como mostrado por Bankier(1990, seção 6.2).

Assim, denote-se por

$$\hat{Y}^{(0)} = \begin{bmatrix} \hat{Y}_1^{(0)} \\ \hat{Y}_2^{(0)} \\ \vdots \\ \hat{Y}_H^{(0)} \end{bmatrix} = [\hat{Y}_h^{(0)}]$$

o vetor de estimativas de total da variável de interesse y para as H classes de ajustamento de pesos definidas, calculadas usando os pesos não viciados iniciais $w_{hk}^{(0)}$;

$$\hat{Y}_h^{(0)} = \sum_{k=1}^{m_h} w_{hk}^{(0)} \cdot y_{hk}$$

o total estimado da variável de interesse y para a h-ésima classe de ajustamento de pesos calculado usando os pesos iniciais não viciados $w_{hk}^{(0)}$;

y_{hk} o valor da variável de interesse y para o k-ésimo domicílio da amostra da h-ésima classe de ajustamento de pesos da área de ponderação considerada;

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_H \end{bmatrix} = [\hat{Y}_h]$$

o vetor de estimativas de total da variável de interesse y para as H classes de ajustamento de pesos definidas, calculadas usando os pesos finais ajustados w_{hk};

$$\hat{Y}_h = \sum_{k=1}^{m_h} w_{hk} \cdot y_{hk} = \sum_{k=1}^{m_h} a_h \cdot w_{hk}^{(0)} \cdot y_{hk} = a_h \cdot \hat{Y}_h^{(0)}$$

o total estimado da variável de interesse y para a h-ésima classe de ajustamento de pesos, calculado usando os pesos finais ajustados w_{hk}.

Usando essa notação pode-se verificar, aplicando a solução da minimização da função de perda G^{GLSM} apresentada anteriormente, que o estimador do total Y da variável de interesse y para o conjunto da população da área de ponderação considerada é da forma:

$$\hat{Y}^{GLSM} = \sum_{h=1}^{H} \hat{Y}_h = \hat{Y}' . 1_H = \hat{Y}^{(0)'} . a =$$

$$= \hat{Y}^{(0)'} . 1_H + \hat{Y}^{(0)'} . V^{-1} . \hat{X}^{(0)} . (\hat{X}^{(0)'} . V^{-1} . \hat{X}^{(0)})^{-1} . (X' . 1_M - \hat{X}^{(0)'} . 1_H)$$

Portanto, $\hat{Y}^{GLSM}$ é da forma

$$\hat{Y}^{GLSM} = \hat{Y}^{(0)'} . 1_H + \hat{\beta}' . (X' . 1_M - \hat{X}^{(0)'} . 1_H)$$

onde

$\hat{\beta} = (\hat{X}^{(0)'} . V^{-1} . \hat{X}^{(0)})^{-1} . \hat{X}^{(0)'} . V^{-1} . \hat{Y}^{(0)}$ é um vetor Qx1.

A partir das equações apresentadas para $\hat{Y}^{GLSM}$ e para $\hat{\beta}$ se pode ver que $\hat{Y}^{GLSM}$ é de fato um estimador de regressão múltipla. O estimador $\hat{\beta}$ foi escolhido de modo que fica minimizada a função de perda modificada dada por:

$$G^{GLSMM} = (\hat{Y}^{(0)} - \hat{X}^{(0)}.\beta)' \cdot V^{-1} \cdot (\hat{Y}^{(0)} - \hat{X}^{(0)}.\beta)$$

onde β é o "parâmetro" do modelo a ser estimado minimizando a função G^{GLSMM}. Bankier(1990, p. 11) mostra que a solução para β que minimiza a função G^{GLSMM} é dada pela equação já apresentada acima para $\hat{\beta}$.

Ainda com respeito a essa maneira de encarar o estimador obtido com a metodologia GLSEP multivariada, Bankier(1990) relata que Särndal, Swensson e Wretman(1989) propõem usar para a matriz V que entra na função de perda G^{GLSMM} a seguinte forma particular:

$V = diag(v)$

onde

$v = \hat{X}^{(0)}.\gamma$ é um vetor Hx1;

$diag(v)$ é uma matriz diagonal com os elementos do vetor v na diagonal principal;

γ é qualquer vetor Qx1 tal que nenhum elemento do vetor v seja nulo.

Segundo Bankier(1990), a utilização de matrizes V dessa forma tem por conseqüência reduzir a expressão do estimador $\hat{Y}^{GLSM}$ à seguinte forma mais simples:

$$\hat{Y}^{GLSM} = \hat{\beta}'.X'.1_M .$$

Além disso, Bankier(1990) alega que o uso de matrizes V da forma proposta permite obter maior "proteção" contra a influência indevida de resíduos grandes do ajuste que aparecem na expressão do estimador $\hat{Y}^{GLSM}$ através da parcela $(\hat{Y}^{(0)'}.1_H - \hat{\beta}'.\hat{X}^{(0)'}.1_H)$, parcela que desaparece com a utilização de matriz V da forma proposta.

É imediato notar que se $\gamma = 1_Q$ então a forma da matriz V proposta por Särndal, Swensson e Wretman(1989) se reduz à matriz cuja diagonal é formada por $X_{h.}^{(0)}$, para $h=1,\ldots,H$, e que foi inicialmente usada para introduzir a versão multivariada da metodologia GLSEP. Como se admite que somente classes de ajustamento de pesos com valores não nulos de $X_{h.}^{(0)}$ serão admitidas, verifica-se que essa é uma das formas aceitáveis para a matriz V.

Mais tarde, entretanto, será mostrada a necessidade de introduzir alguns ajustes nessa matriz de modo que os pesos finais ajustados sejam sempre maiores ou iguais que 1, o que é uma importante necessidade prática.

Bankier(1990) mostra que $E[\hat{Y}^{GLSM}] \cong Y$, isto é, que o estimador $\hat{Y}^{GLSM}$ é aproximadamente não viciado, e que o erro quadrático médio (EQM) desse estimador[23] pode ser obtido aproximadamente por:

$$EQM[\hat{Y}^{GLSM}] \cong VAR[\hat{Y}^{GLSM}] \cong VAR\left[\hat{Y}^{(0)} - \sum_{q=1}^{Q} \beta_q \cdot \hat{X}_{\cdot q}^{(0)} \right] =$$

$$= VAR\left[\sum_{h=1}^{H} \sum_{k=1}^{m_h} w_{hk}^{(0)} \cdot u_{hk} \right] = VAR(\hat{U}^{(0)})$$

onde

$\hat{Y}^{(0)} = \sum_{h=1}^{H} \hat{Y}_h^{(0)} = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_{hk}^{(0)} \cdot y_{hk}$

é o estimador do total Y da variável de interesse y calculado com os pesos não viciados iniciais $w_{hk}^{(0)}$;

$\beta_q = E(\hat{\beta}_q)$ para $q=1,\ldots,Q$;

$\hat{X}_{\cdot q}^{(0)} = \sum_{h=1}^{H} \hat{X}_{hq}^{(0)} = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_{hk}^{(0)} \cdot x_{hqk}$

é o estimador do total X_q da q-ésima variável auxiliar calculado com os pesos não viciados iniciais $w_{hk}^{(0)}$ para $q=1,\ldots,Q$;

$u_{hk} = y_{hk} - \sum_{q=1}^{Q} \beta_q \cdot x_{hqk}$

é uma espécie de "resíduo" da estimação de y_{hk} por $\sum_{q=1}^{Q} \beta_q \cdot x_{hqk}$, para $k=1,\ldots,m_h$ e $h=1,\ldots,H$;

---

[23] Vale notar que $\hat{Y}^{GLSM}$ é o estimador de um total a nível da área de ponderação como um todo, e que as expressões aqui apresentadas são válidas para estimativas a esse nível. Para estimar totais de interesse referentes a áreas menores que as áreas de ponderação e suas respectivas medidas de precisão é necessário empregar estimadores especiais, os quais não serão aqui considerados por estar a solução desse problema fora do escôpo deste trabalho.

$\hat{U}^{(0)} = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_{hk}^{(0)} \cdot u_{hk}$ é um estimador do "total dos resíduos".

Assim, uma aproximação para o erro quadrático médio de $\hat{Y}^{GLSM}$ pode ser determinada mediante o cálculo numérico dos valores dos "resíduos" u_{hk} e sua substituição na fórmula da variância da iteração 0, isto é, na fórmula da variância do estimador não viciado de total que se utilizaria caso os pesos não viciados $w_{hk}^{(0)}$ fossem usados para calcular as estimativas.

Admitindo-se que uma amostra aleatória simples sem reposição (não estratificada) de m domicílios fosse selecionada em cada área de ponderação (hipótese equivalente à que foi empregada para a avaliação da precisão dos estimadores resultantes da metodologia PIETOM no Censo de 80), e que os pesos iniciais não viciados fossem dados por

$w_{hk}^{(0)} = M/m$ para $k=1,\ldots,m_h$ e $h=1,\ldots,H$,

o que equivale a calcular pesos iguais para todos os domicílios da amostra da área de ponderação (portanto, sem considerar efeitos de pós-estratificação para o cálculo dos pesos iniciais não viciados $w_{hk}^{(0)}$) e que a pós-estratificação foi usada apenas para a determinação dos pesos finais ajustados segundo a metodologia GLSEP, é possível calcular a variância de $\hat{U}^{(0)}$.

Essa variância é dada pela expressão da variância do estimador simples (não viciado) de total na amostragem aleatória simples sem reposição tal como encontrada, por exemplo, em Cochran(1977, p. 24):

$$VAR(\hat{U}^{(0)}) = M^2 \cdot \left[\frac{1}{m} - \frac{1}{M}\right] \cdot \frac{1}{M-1} \cdot \sum_{h=1}^{H} \sum_{k=1}^{M_h} \left[u_{hk} - \frac{U}{M}\right]^2$$

onde

M_h é o número de domicílios da população no h-ésimo pós-estrato, para $h=1,\ldots,H$;

$U = \sum_{h=1}^{H} \sum_{k=1}^{M_h} u_{hk}$.

Daí, é imediata a construção de um estimador para o erro quadrático médio de $\hat{Y}^{GLSM}$, bastando para isso substituir os valores de β_q na expressão de u_{hk} pelos correspondentes estimadores $\hat{\beta}_q$, obtendo-se assim os resíduos amostrais $\hat{u}_{hk}$, e logo em seguida substituir esses valores na expressão para o estimador da variância do estimador do total que, sob as mesmas hipóteses de desenho acima, seria dada por:

$$\text{var}(\hat{U}^{(0)}) = M^2 \cdot \left(\frac{1}{m} - \frac{1}{M}\right) \cdot \frac{1}{m-1} \cdot \sum_{h=1}^{H} \sum_{k=1}^{m_h} \left(\hat{u}_{hk} - \frac{u}{m}\right)^2$$

onde

$$\hat{u}_{hk} = y_{hk} - \sum_{q=1}^{Q} \hat{\beta}_q \cdot x_{hqk} \qquad \text{para } k=1,\ldots,m_h \text{ e } h=1,\ldots,H;$$

$$u = \sum_{h=1}^{H} \sum_{k=1}^{m_h} \hat{u}_{hk} .$$

Bankier(1990) mostrou assim que o estimador $\hat{Y}^{GLSM}$ é aproximadamente não viciado e que,sob as mesmas condições consideradas para a avaliação da precisão dos estimadores resultantes da metodologia PIETOM no Censo Demográfico de 1980, estão disponíveis espressões simples para aproximar e estimar o erro quadrático médio desse estimador.

Com a apresentação dessas propriedades do estimador $\hat{Y}^{GLSM}$, pode-se considerar encerrada a descrição genérica da metodologia GLSEP para o caso multivariado. Na seção seguinte apresenta-se os detalhes de como essa metodologia será empregada para definir a metodologia de expansão da amostra do Censo Demográfico de 1991.

Vale a pena comentar, ainda, que essa metodologia é considerada como defensável ou mais aceitável que a denominada GCI sugerida por Costa(1987a), pois todas as propriedades do estimador são conhecidas e o processo de estimação é guiado por princípios estatísticos de minimização de distâncias geralmente aceitos, o que não se dá com o estimador GCI, cuja base empírica não permitiu identificar em que princípios estatísticos de estimação tal estimador se baseia.

5.4 Detalhamento da Metodologia de Estimação Proposta para o Censo 91

Para expandir a amostra do Censo Demográfico de 1991 propõe-se adotar a metodologia descrita a seguir, baseada na aplicação da metodologia GLSEP multivariada em dois estágios, tal como proposta por Bankier(1990) e descrita resumidamente por Rathwell(1991). Para facilidade de referência, propõe-se chamar a metodologia proposta, particularizada para o caso do Censo Demográfico de 1991, de MQG2 - Mínimos Quadrados Generalizados em 2 estágios. A justificativa para aplicar o processo de ajustamento de pesos duas vezes (em dois estágios) se deve à tentativa de manter a consistência das estimativas de totais também a nível dos setores censitários, questão que não era considerada na montagem das metodologias de expansão adotadas nos Censos passados, as quais se preocuparam exclusivamente em garantir consistência das estimativas para totais a nível da área de ponderação como um todo.

A metodologia MQG2 foi concebida de maneira a produzir pesos que, quando usados para calcular as estimativas amostrais para os totais de algumas características básicas (tais como idade e sexo, por exemplo), fornecerão estimativas que se igualam aos valores conhecidos dos totais dessa características obtidos a partir da pesquisa censitária (pesquisa de 100% da população, efetuada usando o questionário básico, modelo CD-1.01). O resultado principal da aplicação dessa metodologia (e do programa de computador desenvolvido para sua implementação[24]) é um peso final ajustado para cada domicílio da amostra pesquisada (cada questionário de amostra modelo CD-1.02).

O programa de computador desenvolvido para implementação da metodologia MQG2 resulta de uma adaptação realizada por Pedro Luis do Nascimento Silva e Sonia Albieri de um programa cedido pelo *Statistics Canada*, o qual foi elaborado por Michael Bankier, Stephen Rathwell e Susana Rubin para aplicação na expansão da amostra do censo canadense de 1991.

Na expansão da amostra do Censo Demográfico de 1980 foram empregados dois processos de estimação distintos e separados, um para as características de pessoas e outro para as de domicílios, embora ambos baseados na mesma metodologia de estimação: o PIETOM. Essa opção tem uma limitação conceitual devida ao fato de que para algumas características, os dois pesos resultantes do PIETOM poderiam ser utilizados, e produziriam resultados diferentes para a mesma quantidade de interesse (que poderiam ainda ser diferentes do valor conhecido para 100% da população). Por essa razão, e por mais algumas outras que ainda se vai enumerar adiante, decidiu-se empregar apenas um peso para a expansão da amostra do Censo Demográfico de 1991.

Assim, a metodologia MQG2 será usada para calcular apenas um peso para cada domicílio da amostra, e a estimação das características de pessoas será feita utilizando-se, para cada pessoa, o peso correspondente ao domicílio onde ela foi recenseada. Os pesos assim calculados serão os únicos disponíveis e serão usados para estimar os totais ou quaisquer outras quantidades de interesse referentes às características pesquisadas por amostragem no Censo.

A exemplo do que ocorreu nos Censos anteriores, a metodologia MQG2 será aplicada separadamente (independentemente) para cada uma das áreas de ponderação que vierem a ser definidas para controlar a expansão da amostra do Censo Demográfico de 1991. Também como nos Censos anteriores, as áreas de ponderação serão formadas por grupos de setores censitários contíguos, e mutuamente exclusivos, de modo que cada setor pertença a uma única área de ponderação e todo setor pertença a alguma área de ponderação.

Das vezes anteriores, os critérios usados para definir as áreas de ponderação sempre incorporaram limites mínimos para o tamanho da população que uma área deveria conter para que pudesse vir a se constituir numa área de ponderação (esse limite foi de 5.000 pessoas no Censo de 80 para áreas de ponderação formadas como subdivisões de municípios).

Para o Censo 91, será também necessário impor um limite superior para a população de uma área de ponderação, pois do contrário a utilização do programa de computador desenvolvido para implementar a metodologia MQG2 ficaria inviabilizada. Esse limite máximo ainda não foi definido, mas poderá ser um número da ordem de até 50.000 pessoas. Nos testes já feitos com o programa, áreas de ponderação com até cerca de 40 setores censitários ou com população da ordem de 40.000 pessoas foram processadas sem qualquer problema de eficiência computacional.

Pode parecer que essa restrição ao tamanho máximo das áreas de ponderação seria uma limitação ou desvantagem da metodologia MQG2 em comparação com as dos censos anteriores, mas tal limitação não é significativa e pode até mesmo representar uma vantagem, se vista de novo ângulo. Acontece que, recentemente, tem aumentado de forma significativa a demanda por informações e estimativas referentes a áreas cada vez menores, cujo atendimento depende da pronta disponibilidade dessas estimativas. O atendimento a essa demanda fica facilitado se o processo regular de expansão da amostra do Censo for efetuado para um maior número de áreas de ponderação de menor tamanho cada uma, pois para obter estimativas referentes a áreas maiores será necessário tão somente agregar (somar) as estimativas das áreas de ponderação que compõem a área maior.

Para as demandas por dados mais desagregados, essa maior pulverização das áreas de ponderação vai facilitar sobremaneira o atendimento, desde que a definição das áreas de ponderação obedeça um critério de partição razoável (levando em conta as divisões administrativas, por exemplo, como já foi feito nos censos anteriores com o uso do município como regra básica de partição das áreas de ponderação). Uma discussão mais detalhada da questão do critério para formação de áreas de ponderação está fora do escopo deste trabalho, mas trata-se de uma das componentes fundamentais do processo de expansão da amostra do Censo. Por essa razão, na seção 7 será apresentada uma proposta de linha de trabalho e investigação que conduza à definição de um critério apropriado para a formação das áreas de ponderação do Censo.

Assim, em cada uma das áreas de ponderação definidas, será aplicada a metodologia MQG2 de ajuste dos pesos que se vai descrever mais adiante na forma de um algoritmo. Antes dessa descrição, entretanto, vale notar que a metodologia MQG2 recebe o nome de Mínimos Quadrados Generalizados em 2 estágios devido ao fato de que o processo de ajuste dos pesos será feito em dois estágios distintos: no primeiro estágio do ajuste, os pesos iniciais não viciados serão ajustados de modo a buscar a consistência entre as estimativas de totais de características de domicílios e de pessoas a nível de cada setor censitário componente da área de ponderação; no segundo estágio, os pesos ajustados resultantes do primeiro ajuste serão novamente ajustados para buscar a consistência entre as estimativas de totais das mesmas características de pessoas e de domicílios com os respectivos totais conhecidos a nível da área de ponderação como um todo, acrescentando-se apenas a busca de consistência com as contagens de domicílios e pessoas existentes em cada setor da área de ponderação.

Embora o segundo estágio do ajuste destrua parcialmente a consistência atingida para as estimativas de totais a nível de setor censitário, as estimativas finais a nível de setor censitário ainda deverão permanecer mais próximas dos respectivos totais populacionais conhecidos que se esse primeiro estágio de ajuste não tivesse sido efetuado.

Apesar do nome indicar que a metodologia MQG2 depende do ajuste dos pesos em 2 estágios, há uma série de operações que precisam ser executadas antes que o cálculo dos pesos ajustados seja efetuado. Além disso, o cálculo dos pesos em cada estágio pode ser subdividido em vários passos. Por essa razão, optou-se por descrever a metodologia na forma de um algoritmo, para ser aplicado separadamente a cada uma das áreas de ponderação definidas. Na apresentação desse algoritmo, considera-se uma área de ponderação típica e descrevem-se os estágios e passos do cálculo para essa área de ponderação.

5.4.1 Estágio Preparatório

Nesta seção são descritos os passos preparatórios ao cálculo dos pesos ajustados propriamente ditos.

Passo 1 - Ordenação dos setores

Um passo preliminar fundamental para a aplicação da metodologia de estimação proposta consiste na ordenação dos setores censitários da área de ponderação considerada em ordem crescente (do menor para o maior) segundo o tamanho, onde o tamanho de cada setor é definido como o número de domicílios particulares permanentes existente no setor (total da população, não da amostra).

Passo 2 - Agregação dos setores

Caso existam na área de ponderação setores com número de domicílios muito pequeno na população, tais setores são agregados (*"collapsed"*) para efeito do cálculo dos pesos, de modo a evitar que possam ocorrer pesos extremos devido à presença desses setores "rarefeitos". A agregação é feita sempre com o setor de tamanho imediatamente superior na relação ordenada dos setores (e não levando em conta qualquer idéia de contigüidade ou proximidade física).

O limite mínimo de tamanho para que os setores não sejam agregados é definido de modo a evitar que, nesses setores, o tamanho da amostra a ser considerada na expansão fique muito pequeno. Na aplicação experimental que foi feita com dados do Censo Demográfico Experimental de Limeira de 1988 (e que está descrita no capítulo 6 a seguir) setores com menos de 40 domicílios na população foram agregados para o cálculo dos pesos.

Sem perda de generalidade, pode-se considerar aqui a notação usada na seção 2.1 para indicar os setores da área de ponderação ($t=1,\ldots,T$) como aplicável ao conjunto dos setores resultante após a ordenação e agregação dos setores rarefeitos efetuadas nos passos 1 e 2, sendo T redefinido como o número total de "setores censitários" da área de ponderação após a agregação. Para facilidade de referência e para evitar confusão com o termo setor censitário, de uso bastante difundido, os "setores censitários" resultantes da agregação serão denominados doravante setores da expansão.

Passo 3 - Definição das classes de ajustamento de pesos

O próximo passo preparatório compreende a definição dos "pós-estratos" ou classes de ajustamento de pesos. A proposta para o Censo 91 é fazer de cada domicílio da amostra na área de ponderação uma "classe de ajustamento de pesos", classe essa com apenas 1 unidade. Assim, o número H de classes de ajustamento de pesos será igual a m, o número total de domicílios da amostra da área de ponderação.

Com essa definição para as classes de ajustamento de pesos, pode-se rever a notação definida na seção 5.3 que previa o uso de dois subíndices (h e k) para indicar as classes de ajustamento de pesos e os domicílios da amostra dentro dessas classes. Agora, como cada classe terá apenas um domicílio e como os setores da expansão serão usados durante o cálculo dos pesos, os subíndices t e i serão adotados (como na seção 2.1) para indicar o setor e o domicílio da amostra dentro do setor, respectivamente. Para a ligação das fórmulas apresentadas no restante desse capítulo com as apresentadas na seção 5.3 será necessário, portanto, adaptar a notação de maneira apropriada, o que a nosso ver não representa problema.

O uso do domicílio da amostra como "classe de ajustamento de pesos" significa que, em comparação com as metodologias de expansão usadas nos Censos passados, se estará utilizando uma "pós-estratificação" muito mais detalhada, sem ter o ônus de aplicar métodos sofisticados ou arbitrários para formar os "pós-estratos".

Além da simplicidade dessa definição das classes de ajustamento de pesos, como já se viu anteriormente que quando os fatores de ajuste forem iguais em duas classes de ajustamento distintas essas classes poderiam ser grupadas, se dispõe de um critério prático para grupar, a posteriori (caso desejável e necessário) os domicílios com pesos finais ajustados iguais (ou muito próximos), critério esse sintonizado com o processo de estimação empregado, o que nem sempre foi garantido pelas metodologias de estimação adotadas no passado.

Um outro aspecto da questão que merece destaque é o fato de que, com essa definição das classes de ajustamento de pesos, as restrições de consistência impostas na estimação é que passam a exercer o papel fundamental na tentativa de garantia de consistência das estimativas amostrais com os totais conhecidos da população e não mais as contagens populacionais de cada "pós-estrato", que no caso não são definidas.

Portanto, a importância que antes se dava à construção de pós-estratos para controlar a expansão, deve agora ser deslocada para a definição das restrições de consistência que devem ser consideradas durante o processo de ajuste dos pesos, o que equivale a concentrar a atenção na definição das variáveis auxiliares a serem usadas para apoiar a expansão.

Passo 4 - Cálculo dos pesos iniciais não viciados

O próximo passo da metodologia compreende o cálculo dos pesos iniciais $w_{ti}^{(0)}$ para cada domicílio da amostra da área de ponderação, $i=1,\ldots,m_t$ e $t=1,\ldots,T$.

No Censo de 80, os pesos iniciais usados pelo PIETOM foram calculados como o inverso da fração amostral observada (de domicílios - M/m - ou de pessoas - N/n, conforme a aplicação desse processo para a expansão de domicílios ou pessoas, respectivamente) para o conjunto da área de ponderação considerada.

Para o Censo 91, propõe-se calcular os pesos não viciados iniciais $w_{ti}^{(0)}$ como sendo o inverso da fração amostral (de domicílios) observada em cada setor censitário da área de ponderação (M_t/m_t, para domicílios do setor t), isto é:

$$w_{ti}^{(0)} = \frac{M_t}{m_t} \quad \text{para } i=1,\ldots,m_t \text{ e } t=1,\ldots,T.$$

Com essa nova forma de calcular os pesos iniciais, se estará respeitando mais o desenho efetivamente adotado na seleção da amostra do Censo, além de privilegiar mais a busca de consistência das estimativas de totais a nível de setores censitários, um dos objetivos eleitos para nortear a metodologia de expansão da amostra recomendada para o Censo 91.

Passo 5 - Definição do conjunto de restrições de consistência

O próximo passo da metodologia compreende a definição do conjunto de restrições de consistência que as estimativas amostrais devem satisfazer, isto é, do conjunto de totais populacionais conhecidos de cujos valores as estimativas amostrais correspondentes devem se aproximar (ou aos quais devem se igualar).

Na expansão da amostra do Censo Demográfico de 1991 propõe-se usar restrições baseadas em totais conhecidos de características de domicílios e também de características de pessoas, apesar do fato de que os pesos serão atribuídos aos domicílios durante o processo de ajuste. Os totais conhecidos que servirão para definir as restrições são referentes ao conjunto da área de ponderação (restrições a nível da área de ponderação) e também a cada um dos setores da expansão (restrições a nível de setores), com os quais serão definidos basicamente dois conjuntos de restrições:

a) um conjunto de restrições baseadas em totais das características auxiliares de domicílios e de pessoas para as quais se deseja assegurar consistência tanto a nível de setores da expansão como a nível da área de ponderação como um todo, e que terá um número fixo inicial de restrições idêntico em todas as áreas de ponderação[25]; essas restrições serão denominadas de restrições do grupo 1 e serão usadas nos dois estágios do processo de ajuste dos pesos - no primeiro estágio serão as únicas restrições de consistência impostas a nível de setores da expansão, e no segundo estágio, juntamente com as restrições do grupo 2 definido a seguir, formarão o conjunto das restrições de consistência impostas a nível da área de ponderação como um todo;

---

[25]Na aplicação experimental agora sendo executada com dados do Censo Experimental de Limeira de 1988 estão sendo consideradas 34 restrições nesse grupo, como se verá mais adiante.

b) um conjunto de restrições baseadas nas contagens conhecidas de domicílios e pessoas de cada setor da expansão, que terá um número inicial de restrições variável em função do número de setores da expansão existente em cada área de ponderação; as restrições desse conjunto formarão o chamado grupo 2 de restrições, e serão usadas somente no sengundo estágio do processo de ajuste dos pesos; na aplicação experimental agora sendo executada com dados do Censo Experimental de Limeira de 1988 estão sendo consideradas 2 restrições para cada setor da expansão, uma para garantir consistência com o número total de domicílios e outra com o número total de pessoas existentes em cada setor da expansão.

Para a aplicação experimental da metodologia MQG2 que vem sendo realizada com dados do Censo Experimental de Limeira de 1988, a definição das restrições de consistência que vêm sendo consideradas foi feita de forma arbitrária, mas para a aplicação da metodologia aos dados do Censo Demográfico de 1991 será necessário ainda realizar um trabalho de investigação para definição mais apropriada dos conjuntos de restrições de consistência que devem ser satisfeitas.

Uma possibilidade nessa direção seria o emprego do pacote CART que foi adquirido pelo IBGE. Acredita-se que sua utilização para definir o que seria uma pós-estratificação razoável para controlar a expansão da amostra do Censo poderia gerar uma definição adequada para o conjunto das restrições de consistência a serem garantidas com a aplicação da metodologia MQG2.

Para fins de tornar mais clara essa definição das restrições e sua ligação com a descrição genérica da metodologia GLSEP apresentada na seção 5.3, são apresentadas a seguir as restrições usadas na aplicação experimental da metodologia MQG2 que vem sendo desenvolvida com base nos dados do Censo Experimental de Limeira de 1988.

Restrições a nível de setor da expansão

São listadas a seguir as restrições definidas com base em características de pessoas moradoras em domicílios particulares permanentes:

X_{t1}= Número total de pessoas moradoras em domicílios particulares permanentes da população do setor t;

X_{t2}= Número de pessoas de 5 anos ou mais de idade do setor t;

X_{t3}= Número de pessoas do sexo masculino do setor t;

X_{t4}= Número de pessoas de 10 anos ou mais de idade do sexo feminino do setor t;

X_{t5}= Número de pessoas de 5 a 9 anos de idade do setor t;

X_{t6}= Número de pessoas de 10 a 14 anos de idade do setor t;

X_{t7}= Número de pessoas de 15 a 19 anos de idade do setor t;

X_{t8}= Número de pessoas de 20 a 24 anos de idade do setor t;

X_{t9}= Número de pessoas de 25 a 29 anos de idade do setor t;

X_{t10}= Número de pessoas de 30 a 34 anos de idade do setor t;

X_{t11}= Número de pessoas de 35 a 39 anos de idade do setor t;

X_{t12}= Número de pessoas de 40 a 44 anos de idade do setor t;

X_{t13}= Número de pessoas de 45 a 49 anos de idade do setor t;

X_{t14}= Número de pessoas de 50 a 59 anos de idade do setor t;

X_{t15}= Número de pessoas de 60 anos ou mais de idade ou com idade ignorada do setor t;

X_{t16}= Número de pessoas residentes na área urbana do setor t[26];

X_{t17}= Número de pessoas do sexo masculino que são chefes de domicílio do setor t;

X_{t18}= Número de chefes de domicílio com menos de 20 anos de idade do setor t;

X_{t19}= Número de chefes de domicílio com idade de 20 a 29 anos do setor t;

X_{t20}= Número de chefes de domicílio com idade de 30 a 39 anos do setor t;

X_{t21}= Número de chefes de domicílio com idade de 40 a 49 anos do setor t;

X_{t22}= Número de chefes de domicílio de 50 anos ou mais de idade ou de idade ignorada do setor t.

São listadas a seguir as restrições definidas com base em características de domicílios particulares permanentes:

X_{t23}= Número total de domicílios particulares permanentes do setor t;

X_{t24}= Número de domicílios com 1 pessoa moradora do setor t;

---

[26]Obviamente, esta restrição só faz sentido para setores urbanos, sendo descartada automaticamente para setores rurais.

X_{t25}= Número de domicílios com 2 pessoas moradoras do setor t;

X_{t26}= Número de domicílios com 3 pessoas moradoras do setor t;

X_{t27}= Número de domicílios com 4 pessoas moradoras do setor t;

X_{t28}= Número de domicílios com 5 pessoas moradoras do setor t;

X_{t29}= Número de domicílios com 6 ou mais pessoas moradoras do setor t;

X_{t30}= Número de domicílios da área urbana do setor t;

X_{t31}= Número de domicílios sem nenhum banheiro do setor t;

X_{t32}= Número de domicílios com 1 banheiro do setor t;

X_{t33}= Número de domicílios com 2 banheiros do setor t;

X_{t34}= Número de domicílios com 3 ou mais banheiros ou com número ignorado de banheiros do setor t.

As restrições de 1 a 34 formam o grupo 1 de restrições definido acima, e são consideradas durante o primeiro estágio do processo de ajuste dos pesos efetuado separadamente para cada setor da expansão da área de ponderação, isto é para t=1,...,T. Com essa definição das restrições de consistência, o número de restrições consideradas é Q_1=34.

Restrições a nível da área de ponderação (AP)

Conforme já foi indicado, as restrições que serão aplicadas no segundo estágio do processo de ajuste dos pesos efetuado a nível da área de ponderação são as mesmas do grupo 1 agora consideradas a nível da área de ponderação acrescidas das restrições do grupo 2.

São listadas a seguir as restrições derivadas das restrições do grupo 1, agora calculadas a nível da área de ponderação como um todo:

X_1= Número total de pessoas moradoras em domicílios particulares permanentes da população da área de ponderação;

X_2= Número de pessoas de 5 anos ou mais de idade da área de ponderação;

X_3= Número de pessoas do sexo masculino da área de ponderação;

X_4= Número de pessoas de 10 anos ou mais de idade do sexo feminino da área de ponderação;

X_5= Número de pessoas de 5 a 9 anos de idade da área de ponderação;

X_6= Número de pessoas de 10 a 14 anos de idade da área de ponderação;

X_7= Número de pessoas de 15 a 19 anos de idade da área de ponderação;

X_8= Número de pessoas de 20 a 24 anos de idade da área de ponderação;

X_9= Número de pessoas de 25 a 29 anos de idade da área de ponderação;

X_{10}= Número de pessoas de 30 a 34 anos de idade da área de ponderação;

X_{11}= Número de pessoas de 35 a 39 anos de idade da área de ponderação;

X_{12}= Número de pessoas de 40 a 44 anos de idade da área de ponderação;

X_{13}= Número de pessoas de 45 a 49 anos de idade da área de ponderação;

X_{14}= Número de pessoas de 50 a 59 anos de idade da área de ponderação;

X_{15}= Número de pessoas de 60 anos ou mais de idade ou com idade ignorada da área de ponderação;

X_{16}= Número de pessoas residentes na área urbana da área de ponderação;

X_{17}= Número de pessoas do sexo masculino que são chefes de domicílio da área de ponderação;

X_{18}= Número de chefes de domicílio com menos de 20 anos de idade da área de ponderação;

X_{19}= Número de chefes de domicílio com idade de 20 a 29 anos da área de ponderação;

X_{20}= Número de chefes de domicílio com idade de 30 a 39 anos da área de ponderação;

X_{21}= Número de chefes de domicílio com idade de 40 a 49 anos da área de ponderação;

X_{22}= Número de chefes de domicílio de 50 anos ou mais de idade ou de idade ignorada da área de ponderação;

X_{23}= Número total de domicílios particulares permanentes da área de ponderação;

X_{24}= Número de domicílios com 1 pessoa moradora da área de ponderação;

X_{25}= Número de domicílios com 2 pessoas moradoras da área de ponderação;

X_{26}= Número de domicílios com 3 pessoas moradoras da área de ponderação;

X_{27}= Número de domicílios com 4 pessoas moradoras da área de ponderação;

X_{28}= Número de domicílios com 5 pessoas moradoras da área de ponderação;

X_{29}= Número de domicílios com 6 ou mais pessoas moradoras da área de ponderação;

X_{30}= Número de domicílios da área urbana da área de ponderação;

X_{31}= Número de domicílios sem nenhum banheiro da área de ponderação;

X_{32}= Número de domicílios com 1 banheiro da área de ponderação;

X_{33}= Número de domicílios com 2 banheiros da área de ponderação;

X_{34}= Número de domicílios com 3 ou mais banheiros ou com número ignorado de banheiros da área de ponderação.

Além das restrições do grupo 1 redefinidas a nível da área de ponderação, são consideradas no segundo estágio do processo de ajuste de pesos restrições de consistência baseadas nas contagens de domicílios e de pessoas de cada um dos T setores da expansão existentes na área de ponderação. Essas restrições são listadas a seguir, e são em número de 2.T para uma área de ponderação com T setores da expansão.

X_{35}= Número total de pessoas moradoras em domicílios particulares permanentes no setor 1 da área de ponderação;

⋮

X_{34+T}= Número total de pessoas moradoras em domicílios particulares permanentes no setor T da área de ponderação;

X_{34+T+1}= Número total de domicílios particulares permanentes no setor 1, da área de ponderação;

⋮

$X_{34+2.T}$= Número total de domicílios particulares permanentes no setor T, da área de ponderação.

Com essa definição de restrições, o número Q_2 de restrições de consistência consideradas para o segundo estágio do processo de ajuste de pesos é dado por: $Q_2 = Q_1 + 2.T$.

Assim, uma área de ponderação com T=10 setores da expansão teria 54 restrições de consistência para satisfazer no segundo estágio do processo de ajuste de pesos, supondo essa definição para os grupos 1 e 2 de restrições.

Passo 6 - Eliminação de restrições a nível da área de ponderação

Uma vez definidas as restrições de consistência que devem ser consideradas para o processo de ajustamento de pesos em cada área de ponderação, há que efetuar um tratamento prévio destinado a eliminar, excluir ou descartar as restrições raras, as restrições redundantes ou quase redundantes, e as que podem implicar na obtenção de pesos extremos, consideradas as Q_1 restrições do grupo 1 a nível da área de ponderação.

A idéia de eliminar as restrições definidas a nível da área de ponderação em primeiro lugar, apesar de que o primeiro estágio do ajuste dos pesos será efetuado a nível de cada setor da expansão separadamente, tem por base a idéia de economia computacional, pois as restrições eliminadas a nível da área de ponderação certamente seriam também eliminadas caso fossem consideradas a nível de setor. Essa eliminação é feita por partes, na ordem descrita a seguir.

Passo 6.1 - Eliminação de restrições raras a nível da área de ponderação

Primeiramente são eliminadas as restrições que são denominadas raras. Cada restrição de consistência, para ser estatisticamente razoável, deve representar um número suficientemente grande de domicílios, pois do contrário o procedimento de estimação adotado pode ficar instável, no sentido de que pequenas mudanças na amostra resultem em grandes variações nos pesos ajustados.

Por outro lado, só é possível exigir consistência (igualdade) entre uma estimativa amostral e um total populacional se houver pelo menos um domicílio na amostra ao qual a restrição seja aplicável.

Por essas razões, após a definição inicial das restrições, são eliminadas todas as que se revelem raras, isto é, que representem um número de domicílios, na população, menor ou igual que um certo limite inferior de tamanho (atualmente fixado em 40) ou que não representem nenhum domicílio na amostra, no sentido de não existir nenhuma unidade amostral com as características que definem a restrição na área de ponderação em questão.

Para restrições referentes a características de pessoas, o tamanho da restrição é medido em número de domicílios nos quais tenha pelo menos uma pessoa à qual a restrição se aplique.

Isto significa que qualquer restrição que represente menos de 40 domicílios na população da área de ponderação considerada ou para a qual não seja encontrado nenhum domicílio na amostra ao qual a restrição se aplique será desconsiderada no processo de ajuste de pesos (nos dois estágios).

Como o limite mínimo de tamanho fixado para definir as restrições raras é o mesmo usado para definir os setores rarefeitos que deveriam ser agregados (passo 1), é imediato notar que essa etapa de eliminação das restrições raras só precisa ser aplicada ao conjunto de restrições a nível da área de ponderação baseadas no grupo 1 de restrições definidas.

Passo 6.2 - Eliminação de restrições redundantes a nível da área de ponderação

Após a eliminação das restrições raras, são executadas duas operações destinadas a eliminar a restrições redundantes. São definidas como restrições redundantes as restrições que gerarem colunas linearmente dependentes na matriz $\hat{X}^{(0)}{}' . V^{-1} . \hat{X}^{(0)}$, onde:

$\hat{X}^{(0)}$ é a matriz $m \times Q_2$ de estimativas de total das Q_2 variáveis auxiliares que definem as Q_2 restrições de consistência consideradas a nível da área de ponderação, calculadas utilizando os pesos não viciados iniciais $w_{ti}^{(0)}$; cada linha dessa matriz é formada por um vetor $1 \times Q_2$, onde a q-ésima coluna deste vetor é dada por $\hat{X}_{tiq}^{(0)} = w_{ti}^{(0)} \cdot x_{tiq}$, sendo x_{tiq} o valor da q-ésima restrição[27] para o domicílio i do setor t, $q=1,\ldots,Q_2$, $i=1,\ldots,m_t$ e $t=1,\ldots,T$;

V é uma matriz diagonal $m \times m$, com os valores de $\hat{X}_{ti}^{(0)} = \sum_{q=1}^{Q_2} \hat{X}_{tiq}^{(0)}$ na diagonal principal.

---

[27] Vale registrar que x_{tqi} é igual a 1 ou 0 caso se trate de uma restrição baseada em características de domicílios (a restrição se aplica ou não se aplica), ou é igual ao número de pessoas daquele domicílio aos quais a restrição se aplica, no caso de restrições baseadas em características de pessoas (por exemplo, $x_{tqi} = 2$ pode significar que há 2 pessoas no domicílio i do setor t aos quais a q-ésima restrição se aplica).

O cálculo dos pesos ajustados depende, se adotada a solução exata do problema de minimização da função G^{GLSMM}, da inversão da matriz $\hat{X}^{(0)}{}' V^{-1} \hat{X}^{(0)}$. Entretanto, se essa matriz tiver colunas linearmente dependentes, seu posto será menor que o número de colunas, e portanto a matriz será não inversível, tornando inviável a obtenção da solução do problema da maneira proposta.

Entretanto, Bankier(1990) comprova que se houver colunas linearmente dependentes nessa matriz, isto implica que as colunas correspondentes na matriz $\hat{X}^{(0)}$ são linearmente dependentes, o que implica que algumas restrições são redundantes, no sentido de que podem ser descartadas ou eliminadas sem que, mais tarde, as estimativas resultantes da expansão da amostra com base nos pesos ajustados deixem de ser consistentes para os respectivos totais conhecidos.

Para dar um exemplo concreto de situação em que poderiam ocorrer restrições de consistência linearmente dependentes, imagine-se que é de interesse garantir consistência entre as estimativas do número total de pessoas da área de ponderação (X_p) e o número total de pessoas da área de ponderação por sexo (X_m=masculino e X_f=feminino). É evidente que os totais conhecidos satisfazem $X_p = X_m + X_f$, e que as estimativas correspondentes a partir da amostra satisfazem $\hat{X}_p = \hat{X}_m + \hat{X}_f$. Assim, bastaria garantir, por exemplo, que $\hat{X}_p = X_p$ e $\hat{X}_m = X_m$, para que a igualdade entre $\hat{X}_f = X_f$ também ficasse assegurada.

Na situação exemplificada, se as restrições X_p, X_f e X_m fossem consideradas simultaneamente, a matriz $\hat{X}^{(0)}{}' V^{-1} \hat{X}^{(0)}$ teria colunas linearmente dependentes e seria não inversível, a menos que uma das colunas correspondentes a uma dessas três restrições (X_p, X_m ou X_f) fosse eliminada. No conjunto de restrições definidas para a aplicação experimental, algumas restrições que sabidamente gerariam conjuntos linearmente dependentes já deixaram de ser consideradas, como se pode observar do fato de que apenas o total de pessoas e o total de pessoas do sexo masculino foram definidas como restrições de consistência (e o número de pessoas do sexo feminino não foi).

Em outros casos, apesar de se saber a priori que alguns conjuntos de restrições gerariam colunas linearmente dependentes, as mesmas foram deixadas no conjunto inicial, ficando para essa etapa de eliminação a escolha (automática) das restrições que deveriam ser eliminadas para acabar com a condição de dependência linear entre as restrições (devido ao fato de que, a priori, não se saberia qual delas excluir com maior proveito).

Assim, para garantir a existência da inversa da matriz $\hat{X}^{(0)}{}' . V^{-1} . \hat{X}^{(0)}$ serão sempre eliminadas as menores restrições entre as que forem detectadas como linearmente dependentes, onde por menores restrições se entende as que se aplicarem ao menor número de unidades da população (domicílios ou pessoas) da área de ponderação.

Após a eliminação das restrições redundantes, as matrizes $\hat{X}^{(0)}$ e V são redefinidas excluindo-se as colunas correspondentes às restrições eliminadas. As matrizes assim atualizadas é que são empregadas na etapa de eliminação seguinte.

Passo 6.3 - Eliminação de restrições quase redundantes a nível da área de ponderação

Após a eliminação das restrições redundantes (linearmente dependentes) a nível da área de ponderação, são também eliminadas restrições que são quase redundantes, devido corresponderem a colunas da matriz $\hat{X}^{(0)}{}' . V^{-1} . \hat{X}^{(0)}$ que são quase linearmente dependentes, isto é, a colunas dessa matriz que não são exatamente iguais a combinações lineares de outras colunas, mas que são quase iguais ou podem ser muito bem aproximadas por combinações lineares desse tipo.

Esse passo é necessário devido ao fato de que, apesar de não haver problemas para o cálculo em computador da inversa da matriz $\hat{X}^{(0)}{}' . V^{-1} . \hat{X}^{(0)}$, as estimativas resultantes da aplicação da metodologia MQG2 com tal matriz apresentariam instabilidade indesejável.

A identificação das restrições quase linearmente dependentes é feita com base no número de condição da matriz $\hat{X}^{(0)}{}' . V^{-1} . \hat{X}^{(0)}$ (razão entre seu maior e menor autovalores). A ocorrência de um valor muito alto para o número de condição indica que essa matriz é quase singular ou inversível e revela a presença de restrições quase linearmente dependentes.

Assim, após a eliminação das restrições raras e redundantes, sempre que necessário, aplica-se um processo de eliminação das restrições que se revelem as maiores responsáveis por valores elevados do número de condicão da matriz $\hat{X}^{(0)}{}' V^{-1} \hat{X}^{(0)}$. As restrições eliminadas por esse motivo são denominadas quase linearmente dependentes ou quase redundantes. Acredita-se que a eliminação de restrições desse tipo não prejudique de forma significativa a manutenção da consistência para as estimativas desejadas, devido ao fato de tratar-se de restrições que poderiam ser bem aproximadas por combinações lineares de outras restrições que foram mantidas.

5.4.2 Primeiro estágio de ajuste dos pesos

Nesta seção são descritos os passos necessários para o cálculo dos pesos ajustados de modo a garantir a consistência das estimativas amostrais com os totais conhecidos a nível de setores da expansão. Vale notar que os passos descritos nesta seção são repetidos separadamente (ou independentemente) para cada um dos setores da expansão.

Passo 7 - Eliminação das restrições raras e redundantes a nível de cada setor da expansão

Neste passo, as restrições que não tenham sido eliminadas a nível da área de ponderação são consideradas para cada setor da expansão, eliminando-se as restrições raras e as restrições que se revelem redundantes a nível de cada setor da expansão. Os critérios de eliminação de restrições nesta etapa são idênticos aos aplicados para eliminação de restrições a nível da área de ponderação.

Algumas restrições que não tenham sido eliminadas como raras ou linearmente dependentes ao nível da área de ponderação poderão ser eliminadas nesse momento, pois podem ser raras ou linearmente dependentes a nível de algum setor particular.

Por razões de economia computacional, não é aplicado a nível de setor o algoritmo para eliminação de restrições quase linearmente dependentes ou quase redundantes.

Passo 8 - Cálculo dos pesos ajustados a nível de setor da expansão

Antes de calcular os pesos ajustados para cada setor, as restrições que permanecem são ordenadas segundo seu tamanho (número de unidades aos quais se aplicam) da menor para a maior, e o conjunto de restrições de consistência de cada setor é subdividido em dois: um com as restrições de números ímpares e outro com as de números pares. Cada um desses subconjuntos de restrições é usado para calcular um peso ajustado correspondente a cada domicílio em cada setor da área de ponderação.

Para apresentar o cálculo dos pesos em termos da notação que vem sendo usada, será necessário introduzir alguma notação adicional. Assim, denote-se por:

C_{t1} o conjunto de restrições de números de ordem ímpar do setor t;

C_{t2} o conjunto de restrições de números de ordem par do setor t;

Q_{t1} o número de restrições em C_{t1};

Q_{t2} o número de restrições em C_{t2}.

As funções de distância que se vai minimizar, para dar origem aos pesos ajustados a nível de setor, em cada subconjunto de restrições, são dadas por:

$$G^r = \sum_{i=1}^{m_t} x_{ti\cdot}^{(0)} \cdot \left(\frac{x_{ti\cdot}^{(r)} - x_{ti\cdot}^{(0)}}{x_{ti\cdot}^{(0)}} \right)^2$$

sujeitas a

$$x_{t\cdot q}^{(r)} = X_{t\cdot q} \qquad \text{para } q \in C_{tr},\ r=1,2,\ t=1,\ldots,T;$$

onde

G^r é a função de distância para o r-ésimo grupo de restrições ($r=1,2$) do setor t, $t=1,\ldots,T$;

$$x_{ti\cdot}^{(0)} = \sum_{q=1}^{Q_{tr}} w_{ti}^{(0)} \cdot x_{tiq}$$

$$x_{ti\cdot}^{(r)} = \sum_{q=1}^{Q_{tr}} w_{ti}^{(r)} \cdot x_{tiq}$$

$w_{ti}^{(r)} = a_{ti}^{(r)} \cdot w_{ti}^{(0)}$ é o peso ajustado para o grupo de restrições C_{tr}, correspondente ao i-ésimo domicílio da amostra do setor t;

$\hat{X}^{(r)}_{t\cdot q} = \sum_{i=1}^{m_t} w^{(r)}_{ti} \cdot x_{tiq}$ é o estimador do total da q-ésima variável auxiliar do grupo r para o setor t;

$X_{t\cdot q} = \sum_{I=1}^{M_t} X_{tiq}$ é o total conhecido da q-ésima variável auxiliar do grupo r para o setor t.

A solução da minimização das funções G^r (para r=1,2) é obtida, conforme já mencionado, calculando-se os vetores de coeficientes de ajuste $a^{(r)}_t$ dados por:

$$a^{(r)}_t = 1_{m_t} + (V_{tr})^{-1} \cdot \hat{X}^{(0)}_{tr} \cdot \left[\hat{X}^{(0)}_{tr}, \cdot (V_{tr})^{-1} \cdot \hat{X}^{(0)}_{tr}\right]^{-1} \cdot ((X_t)' \cdot 1_{M_t} - \hat{X}^{(0)}_{tr}, \cdot 1_{m_t})$$

onde

$a^{(r)}_t = (a^{(r)}_{t1}, a^{(r)}_{t2}, \ldots, a^{(r)}_{tm_t})'$ é o vetor $m_t x 1$ de coeficiente de ajuste de pesos para o grupo de restrições r do setor t;

$1_{m_t} = (1, 1, \ldots, 1)'$ é o vetor unitário de dimensão $m_t x 1$;

$V_{tr} = \text{diag}(v_{tr})$ é uma matriz $m_t x m_t$;

$v_{tr} = (\hat{X}^{(0)}_{t1\cdot}, \hat{X}^{(0)}_{t2\cdot}, \ldots, \hat{X}^{(0)}_{tm_t\cdot})'$;

$\hat{X}^{(0)}_{tr} = [\hat{X}^{(0)}_{tiq}]$ é matriz $m_t x Q_{tr}$;

$\hat{X}^{(0)}_{tiq} = w^{(0)}_{ti} \cdot x_{tiq}$;

$X_t = [X_{t\cdot q}]$ é matriz $M_t x Q_{tr}$;

$1_{M_t} = (1, 1, \ldots, 1)'$ é o vetor unitário de dimensão $M_t x 1$.

De acordo com a metodologia geral GLSEP, os pesos ajustados deveriam satisfazer apenas a condição de pertencerem ao intervalo $(0, +\infty)$. Entretanto, do ponto de vista prático, é indesejável a obtenção de pesos ajustados com valores extremos.

Algumas vezes, os pesos resultantes de algum desses dois conjuntos de restrições podem assumir valores considerados extremos[28] (ser menores que 1 ou maiores que 25). Sempre que isso ocorre, as restrições causadoras da obtenção dos pesos extremos são identificadas e eliminadas do respectivo grupo de restrições, no setor correspondente. Em seguida à eliminação dessas restrições, os pesos ajustados para esse grupo de restrições são recalculados. Esse processo é repetido até que todos os pesos ajustados fiquem dentro do intervalo [1,25] definido como intervalo de aceitação dos pesos ajustados.

Os pesos ajustados para cada um dos grupos de restrições são, então, calculados por:

$$w_{ti}^{(r)} = w_{ti}^{(0)} \cdot a_{ti}^{(r)} \quad \text{para } r=1,2.$$

O peso ajustado ao final do primeiro estágio de ajuste é finalmente obtido calculando-se a média aritmética entre os dois pesos assim calculados, isto é:

$$w_{ti}^{(A)} = \frac{w_{ti}^{(1)} + w_{ti}^{(2)}}{2}$$

onde

$w_{ti}^{(A)}$ é o peso ajustado para o i-ésimo domicílio do setor t ao final do primeiro estágio do processo de ajuste.

Essa forma de calcular os pesos ajustados no primeiro estágio de ajuste teve por finalidades garantir a eficiência computacional, pois assim a dimensão das matrizes a inverter foi dividida por dois e o processo de busca e eliminação das restrições causadoras de pesos extremos também pode ser feito em dois conjuntos de restrições de metade do tamanho do conjunto original cada um.

---

[28] Os limites usados para considerar pesos como extremos (muito pequenos ou muito grandes) foram 1 e 25, seguindo sugestão da experiência canadense, mas tais valores podem ser facilmente modificados nos programas usados para implementar a metodologia MQG2, e estudos posteriores serão necessários para definir os valores a serem efetivamente usados na expansão da amostra do Censo 91.

Bankier(1990, p. 25) comenta, entretanto, que as estimativas que seriam obtidas caso os pesos ajustados $w_{ti}^{(A)}$ fossem usados para expandir a amostra de cada setor não seriam exatamente iguais aos respectivos totais conhecidos da população. Isto significa que se poderia ter

$$|\hat{X}_{t\cdot q}^{(A)} - X_{t\cdot q}| > 0 \quad \text{para algum } q,$$

onde

$\hat{X}_{t\cdot q}^{(A)} = \sum_{i=1}^{m_t} w_{ti}^{(A)} \cdot x_{tiq}$ é o estimador do total da q-ésima restrição para o setor t calculado usando os pesos ajustados do primeiro estágio $w_{ti}^{(A)}$;

$X_{t\cdot q} = \sum_{I=1}^{M_t} X_{tiq}$ é o total conhecido da q-ésima restrição para o setor t.

Mas Bankier(1990) relata também que, num exemplo numérico que executou com essa metodologia, as diferenças entre as estimativas calculadas com os pesos ajustados foram, na maior parte das vezes, menores que as que se teria obtido caso as estimativas fossem calculadas usando os pesos não viciados iniciais $w_{ti}^{(0)}$. Isto é, Bankier(1990) alega que, na maior parte das vezes, se observou que

$$|\hat{X}_{t\cdot q}^{(A)} - X_{t\cdot q}| \leq |\hat{X}_{t\cdot q}^{(0)} - X_{t\cdot q}|$$

comprovando na prática (para o exemplo numérico que ele executou) que as estimativas baseadas nos pesos ajustados (ao final do primeiro estágio) são mais "consistentes" que as estimativas baseadas nos pesos não ajustados.

5.4.3 Segundo estágio de ajuste dos pesos

Nesta seção são descritos os passos necessários para o cálculo dos pesos ajustados de modo a garantir a consistência das estimativas amostrais com os totais conhecidos a nível da área de ponderação como um todo.

Passo 9 - Cálculo dos pesos ajustados a nível da área de ponderação

Finalmente, após o cálculo dos pesos ajustados do primeiro estágio (para garantir consistência a nível de setores da expansão), os pesos ajustados $w_{ti}^{(A)}$ são usados como entrada para o segundo estágio do ajuste de pesos, que visa garantir consistência das estimativas com os totais conhecidos a nível da área de ponderação.

Nesse passo, todas as restrições não eliminadas no passo 6 por serem raras, redundantes ou quase redundantes a nível da área de ponderação são usadas durante o processo de ajuste dos pesos, acrescidas das restrições do grupo 2 que visam manter a consistência das contagens de domicílios e pessoas a nível de cada setor da expansão. Chamando de:

C_B o conjunto de restrições não eliminadas no passo 6 que serão consideradas para o segundo estágio do ajuste dos pesos;

Q_B o número de restrições em C_B.

A função de distância que se vai minimizar, para dar origem aos pesos ajustados a nível da área de ponderação, consideradas as restrições do conjunto C_B, é dada por:

$$G^B = \sum_{t=1}^{T} \sum_{i=1}^{m_t} x_{ti\cdot}^{(A)} \cdot \left(\frac{x_{ti\cdot} - x_{ti\cdot}^{(A)}}{x_{ti\cdot}^{(A)}} \right)^2$$

sujeitas a

$$x_{\cdot\cdot q} = X_q \qquad \text{para } q \in C_B,$$

onde

G^B é a função de distância para o segundo estágio de ajustamento dos pesos;

$$x_{tiq}^{(A)} = w_{ti}^{(A)} \cdot x_{tiq} \qquad \text{para } q \in C_B,\ i=1,\ldots,m_t \text{ e } t=1,\ldots,T;$$

$$x_{ti\cdot}^{(A)} = \sum_{q=1}^{Q_B} x_{tiq}^{(A)} \qquad \text{para } i=1,\ldots,m_t \text{ e } t=1,\ldots,T;$$

$$x_{tiq} = w_{ti} \cdot x_{tiq} \qquad \text{para } q \in C_B,\ i=1,\ldots,m_t \text{ e } t=1,\ldots,T;$$

$$w_{ti} = a_{ti} \cdot w_{ti}^{(A)} \qquad \text{para } i=1,\ldots,m_t \text{ e } t=1,\ldots,T;$$

$$x_{ti\cdot} = \sum_{q=1}^{Q_B} x_{tiq} \qquad \text{para } i=1,\ldots,m_t \text{ e } t=1,\ldots,T;$$

$$\hat{X}_{..q} = \sum_{t=1}^{T} \sum_{i=1}^{m_t} \hat{X}_{tiq} \quad \text{para } q \in C_B;$$

$$X_q = \sum_{t=1}^{T} \sum_{I=1}^{M_t} X_{tIq}$$ é o total conhecido da q-ésima restrição a nivel da área de ponderação;

A solução da minimização da função G^B é obtida, conforme já mencionado, calculando-se o vetor de coeficientes de ajuste a dado por:

$$\mathbf{a} = 1_m + \mathbb{V}_A^{-1}.\hat{\mathbb{X}}^{(A)}.\left[\hat{\mathbb{X}}^{(A)\prime}.\mathbb{V}_A^{-1}.\hat{\mathbb{X}}^{(A)}\right]^{-1}.(\mathbb{X}'.1_M - \hat{\mathbb{X}}^{(A)\prime}.1_m)$$

onde

$\mathbf{a} = (a_{11}, a_{12}, \ldots, a_{1m_1}, \ldots, a_{T1}, a_{T2}, \ldots, a_{Tm_T})'$
é o vetor mx1 de coeficientes de ajuste de pesos para o segundo estágio de ajuste dos pesos;

$1_m = (1,1,\ldots,1)'$ é o vetor unitário de dimensão mx1;

$\mathbb{V}_A = \text{diag}(\omega_A)$ é uma matriz mxm;

$\omega_A = (\hat{X}_{11\cdot}^{(A)}, \hat{X}_{12\cdot}^{(A)}, \ldots, \hat{X}_{1m_1\cdot}^{(A)}, \ldots, \hat{X}_{T1\cdot}^{(A)}, \hat{X}_{T2\cdot}^{(A)}, \ldots, \hat{X}_{Tm_T\cdot}^{(A)})'$;

$\hat{\mathbb{X}}^{(A)} = [\hat{X}_{tiq}^{(A)}]$ é matriz mxQ_B;

$\mathbb{X} = [X_{tIq}]$ é matriz MxQ_B;

$1_M = (1,1,\ldots,1)'$ é o vetor unitário de dimensão Mx1.

De acordo com a metodologia geral GLSEP, os pesos ajustados w_{ti} deveriam satisfazer apenas a condição de pertencerem ao intervalo $(0,+\infty)$. Entretanto, do ponto de vista prático, é indesejável a obtenção de pesos ajustados com valores extremos. Mais uma vez podem ocorrer pesos extremos (fora dos limites especificados como aceitáveis - os limites usados são os mesmos que no primeiro estágio de ajuste). Sempre que isso ocorrer, as restrições identificadas como causadoras da situação são eliminadas e os pesos ajustados recalculados, de modo que ao final do segundo estágio de ajuste nenhum peso ficará fora dos limites especificados.

Os pesos w_{ti} resultantes do segundo estágio do ajuste é que serão usados para a expansão da amostra do Censo Demográfico 91, caso a metodologia proposta venha a ser adotada. Para facilitar a expansão, esses pesos poderiam ser gravados nos registros de dados de cada domicílio ou pessoa provenientes do questionário da amostra (CD-1.02) do Censo 91.

Os pesos finais ajustados poderão ser usados para calcular a estimativa de total e a correspondente estimativa de erro quadrático médio para uma variável de interesse y a nível da área de ponderação usando:

$$\hat{Y} = \sum_{t=1}^{T} \sum_{i=1}^{m_t} w_{ti} \cdot y_{ti}$$

$$eqm(\hat{Y}) = \sum_{t=1}^{T} M_t^2 \cdot \left[\frac{1}{m_t} - \frac{1}{M_t}\right] \cdot s_t^2$$

com

$$s_t^2 = \frac{1}{m_t - 1} \cdot \sum_{i=1}^{m_t} (z_{ti} - \bar{z}_t)^2 \ ;$$

$$\bar{z}_t = \frac{1}{m_t} \cdot \sum_{i=1}^{m_t} z_{ti} \ .$$

Nas fórmulas acima, os valores de z_{ti} são os resíduos da estimação de y_{ti} com base na regressão de y contra as Q_B variáveis auxiliares que definem as restrições de consistência consideradas a nível da área de ponderação, os quais podem ser obtidos calculando-se o vetor mx1 de resíduos dado pela expressão:

$$\mathbf{z} = B^{(A)} \cdot B \cdot \mathbf{y}$$

onde

$\mathbf{z} = (z_{11}, \ldots, z_{1m_1}, \ \ldots \ , z_{T1}, \ldots, z_{Tm_T})'$ é o vetor mx1 de resíduos;

$B^{(A)}$ é uma matriz mxm diagonal por blocos, com as matrizes $B_t^{(A)}$ formando os blocos da diagonal, $t=1, \ldots, T$;

$B_t^{(A)} = I_{m_t} - P_t^{(A)} \cdot diag(W_t^{(0)})$ é uma matriz $m_t x m_t$, para cada $t=1, \ldots, T$;

I_{m_t} é a matriz identidade de dimensão $m_t x m_t$, para cada $t=1, \ldots, T$;

$diag(W_t^{(0)})$ é uma matriz diagonal $m_t x m_t$, com os pesos iniciais não viciados $w_{ti}^{(0)}$ na diagonal, para cada $t=1, \ldots, T$;

$$P_t^{(A)} = \frac{1}{2} \mathbf{x}_t \sum_{r=1}^{2} \left[X_{tr}^{(0)\prime} \cdot (V_{tr})^{-1} \cdot X_{tr}^{(0)} \right]^{-1} \cdot X_{tr}^{(0)\prime} \cdot (V_{tr})^{-1} \ ;$$

$\mathbf{x}_t = [x_{tiq}]$ é uma matriz $m_t x Q$;

$\hat{X}_{tr}^{(0)} = [\hat{x}_{tiq}^{(0)}]$ é matriz $m_t x Q_{tr}$ que contém as colunas correspondentes às restrições em C_{tr}, para $r=1,2$ e $t=1, \ldots, T$;

$V_{tr} = \text{diag}(v_{tr})$ é uma matriz $m_t \times m_t$ tal como já definida na seção 5.4.2;

$v_{tr} = (X^{(0)}_{t1\cdot}, X^{(0)}_{t2\cdot}, \ldots, X^{(0)}_{tm_t\cdot})'$;

$B = I_m - P \cdot \text{diag}(W^{(A)})$ é uma matriz mxm;

I_m é a matriz identidade de dimensão mxm;

$P = \mathbf{x} \cdot \left[\hat{X}^{(A)\prime} \cdot (V_A)^{-1} \cdot \hat{X}^{(A)}\right]^{-1} \cdot \hat{X}^{(A)\prime} \cdot (V_A)^{-1}$ é uma matriz mxm;

$\mathbf{x} = [x_{tiq}]$ é uma matriz $m \times Q_B$ com os subíndices t e i definindo as linhas;

$\hat{X}^{(A)} = [\hat{X}^{(A)}_{tiq}]$ é matriz $m \times Q_B$ já definida;

$V_A = \text{diag}(v_A)$ é uma matriz mxm também já definida;

$v_A = (X^{(A)}_{11\cdot}, X^{(A)}_{12\cdot}, \ldots, X^{(A)}_{1m_1\cdot}, \ldots, X^{(A)}_{T1\cdot}, X^{(A)}_{T2\cdot}, \ldots, X^{(A)}_{Tm_T\cdot})'$;

$\mathbf{y} = (y_{11}, \ldots, y_{1m_1}, \ldots, y_{T1}, \ldots, y_{Tm_T})'$ é o vetor mx1 de valores da variável de interesse y para os m domicílios da amostra na área de ponderação.

Embora as fórmulas para o cálculo dos resíduos usados para estimar o erro médio quadrático do estimador do total de uma característica de interesse y qualquer possam parecer complexas, tais fórmulas dependem apenas de duas matrizes $B^{(A)}$ e B, que serão únicas para qualquer variável de interesse y. Portanto, essas matrizes podem ser calculadas uma única vez (durante o processo de ajustamento de pesos) e armazenadas, de modo que o erro quadrático médio possa mais tarde ser estimado para qualquer característica de interesse y sem maiores problemas. De fato, essa foi a estratégia utilizada no desenvolvimento dos programas de computador que implementam a metodologia MQG2, o que torna bastante simples e barato o processo de estimar totais e sua respectiva precisão, uma vez efetuado o processo de ajuste de pesos para uma área de ponderação qualquer.

Vale notar também que as matrizes $B^{(A)}$ e B representam, respectivamente, os resíduos das regressões referentes ao ajuste do primeiro e do segundo estágio. Mais ainda, como a fórmula proposta para estimar o erro médio quadrático de $\hat{Y}$ é baseada na soma de parcelas de variância calculadas dentro de cada setor da expansão, essa fórmula não deixa de incorporar o efeito da estratificação da amostra por setores, como fazia o estimador da precisão usado no Censo 80 com a metodologia PIETOM.

O único efeito do desenho amostral não incorporado é o efeito da seleção sistemática dos domicílios dentro de cada setor, que é desprezado quando se admite que a amostra de domicílios dentro de cada setor é aleatória simples, mas essa é uma hipótese largamente utilizada na prática. Com isso, encerra-se a descrição do algoritmo de ajuste de pesos.

5.5 Razões que Recomendam a Adocão da Metodologia MQG2

Nesta seção, são examinadas as razões principais a recomendar a adoção da metodologia MQG2 para a expansão da amostra do Censo 91. Mas antes que essas razões sejam expostas, vale a pena indicar alguns requisitos que se deseja ver satisfeitos pela metodologia de expansão a ser adotada no próximo Censo.

O primeiro deles, bastante natural, é o interesse em utilizar métodos de estimação capazes de produzir estimativas mais precisas (com menor erro quadrático médio) que as que seriam produzidas caso a metodologia usada no Censo 80 fosse novamente utilizada, isto é, o desejo natural de aumentar a eficiência do processo de estimação.

O segundo requisito pode ensejar alguma controvérsia, mas em se tratando do Censo, uma pesquisa cujos resultados serão a base para estudos, pesquisas e análises durante toda a década que se segue, não pode ser desprezado por representar uma visão estratégica. Trata-se da idéia de passar a garantir consistência aproximada das estimativas a nível de setores da expansão, e não só a nível da área de ponderação, como antes. Mais ainda, de passar a utilizar maior número de áreas de ponderação de menor tamanho para controlar a expansão da amostra.

Ambas as idéias estão sintonizadas com o aumento da demanda por informações com maior detalhamento geográfico (estimativas para pequenas áreas) que se vem observando e que se espera observar ainda mais para os anos que virão. Caso a metodologia de expansão a ser usada no Censo não contemple esse requisito, será muito mais difícil e caro produzir as estimativas para pequenas áreas que vão ser demandadas.

O terceiro requisito tem origem no desenvolvimento metodológico e dos recursos de informática verificado durante a última década. Trata-se da idéia de adotar uma solução computacionalmente eficiente, que aumente o grau de automação do processo de expansão da amostra, exigindo um mínimo de intervenção manual durante a operação.

O quarto e último requisito é a idéia de que os métodos de expansão adotados respeitem ao máximo possível o desenho amostral adotado durante a coleta dos dados, sem deixar de usar as informações auxiliares disponíveis para o conjunto da população.

Após a indicação desses requisitos, fica mais fácil entender as razões que levam à recomendação de adoção da metodologia MQG2 para a expansão da amostra do Censo 91.

Em primeiro lugar, o uso do domicílio como classe de ajustamento de pesos durante o processo de ajustamento de pesos vai atender a três dos requisitos acima:

a) por eliminar a necessidade de processo sofisticado ou arbitrário de construção de pós-estratos, simplifica sobremaneira essa etapa do processo;

b) o maior detalhamento da "pós-estratificação" assim conseguido promove maior eficiência (menor erro quadrático médio) das estimativas resultantes, segundo Bankier(1990);

c) respeita-se mais o desenho amostral adotado na coleta, por eliminar as hipóteses efetuadas na expansão de características de pessoas durante o Censo 80.

Além disso, efetuar um único processo de ajuste de pesos (e não dois como se fez no Censo 80) reduzirá o custo da expansão e eliminará os problemas de inconsistências que poderiam resultar devido à existência de dois conjuntos de pesos para expansão (o de pessoas e o de domicílios, antes disponíveis).

A mudança no critério de formação de "pós-estratos" para controlar a expansão atribui maior peso à definição das restrições de consistência, que antes eram uma conseqüência da pós-estratificação usada. Com a metodologia MQG2, explicita-se o problema de definição das restrições de consistência, dando maior controle sobre esse aspecto do processo de ajuste dos pesos.

Como a definição das restrições de consistência independe da "pós-estratificação", é possível passar a considerar distribuições marginais das variáveis auxiliares não mais limitadas à visão restrita dada pelas marginais das linhas ou das colunas de uma tabela de contingência de duas entradas.

Além disso, as regras de eliminação de restrições de consistência raras, redundantes, quase redundantes ou causadoras de pesos extremos que são propostas asseguram automação fácil do processo, fornecem um mecanismo apropriado para adaptar os requisitos de consistência impostos às peculiaridades de cada área de ponderação de forma totalmente integrada com o processo de estimação, e não prejudicam a obtenção da consistência desejada, por obedecerem a critérios estatisticamente razoáveis.

Outro aspecto a recomendar a metodologia MQG2 é o cálculo dos pesos iniciais não viciados a nível de cada setor da expansão, e não mais a nível da área de ponderação como um todo, implicando maior respeito ao desenho amostral usado na coleta e maior consistência das estimativas a nível de setor.

Por todas essas razões, e mais ainda, devido se ter relato do sucesso da experiência canadense com a aplicação dessa metodologia para a expansão da amostra do Censo 91 daquele país e à disponibilidade de programas de computador que implementam a metodologia MQG2 tal como descrita neste capítulo, programas esses já adaptados para funcionarem no IBGE com dados do Censo Demográfico Experimental de Limeira e cuja adaptação para serem usados no Censo Demográfico 91 não é difícil nem custosa, é que se recomenda fortemente a adoção dessa metodologia para ser usada na expansão da amostra do Censo 91.

Vale destacar, entretanto, que a adoção dessa metodologia depende de algumas condições para que possa ser bem sucedida.

Primeiramente, é necessário que o processo de apuração dos dados do Censo 91 assegure a ligação perfeita entre os registros de dados em meio magnético de cada questionário de amostra (modelo CD-1.02) com os correspondentes registros de dados do questionário básico respectivo (modelo CD-1.01) que foram "criados" por cópia durante a coleta para agilizarem a apuração dos resultados correspondentes às variáveis investigadas a 100% no Censo. A experiência com os dados do Censo de Limeira mostrou que essa ligação não foi sempre possível, por não ter sido contemplada como um dos objetivos do processo de apuração dos dados daquele Censo Experimental.

Em segundo lugar, mas não menos importante, está a manutenção dos programas necessários à implementação da metodologia MQG2 em computador, que depende do pacote SAS (módulos básico e IML[29], e linguagem MACRO[30]). Caso esses programas não estejam disponíveis, não será possível usar os programas já cedidos pelo *Statistics Canada* e adaptados para uso no IBGE. Nessa linha, será necessário também assegurar o armazenamento em disco dos dados necessários para expandir a amostra de cada área de ponderação, para garantir a eficiência computacional do processo.

Em terceiro lugar, é necessário aceitar a idéia de usar áreas de ponderação de tamanho máximo limitado para controlar a expansão da amostra. Essa é uma limitação atual do programa que implementa o processo de ajuste de pesos da metodologia MQG2. Até o presente momento, esse programa foi testado com sucesso para áreas de ponderação de tamanho igual a 40 setores e cerca de 40.000 pessoas. Esses limites não são definitivos, mas é preciso aceitar a idéia de que vai haver limites máximos para o tamanho das áreas de ponderação que se poderá usar durante a expansão.

---

[29] IML - Interactive Matrix Language - ver SAS(1989d).

[30] Ver SAS(1990c).

Finalmente, mas também não menos importante, será necessário assegurar os meios e as condições de acesso de técnicos do IBGE a um processo de intercâmbio com os técnicos do *Statistics Canada* que trabalharam no desenvolvimento e na aplicação da metodologia usada para expansão da amostra do Censo canadense de 1991, não só para obter deles maiores informações sobre os problemas e sucessos da experiência, como também para sanar eventuais dúvidas que ainda possam surgir decorrentes da implementação da metodologia MQG2 para expandir a amostra do Censo brasileiro.

Isto termina este capítulo, em que a descrição da metodologia proposta foi apresentada e discutida. Na seção seguinte, uma aplicação experimental dessa metodologia MQG2 aos dados do Censo Demográfico Experimental de Limeira de 1988 é relatada e analisada.

6. UMA APLICAÇÃO EXPERIMENTAL DA METODOLOGIA MQG2

Nesta seção é descrita uma aplicação experimental da metodologia MQG2 aos dados coletados no Censo Demográfico Experimental de Limeira de 1988, através da qual se buscou avaliar a viabilidade e conveniência da utilização dessa metodologia para a expansão da amostra do Censo Demográfico de 1991. Vale registrar que os resultados descritos nesse trabalho são os obtidos até o mês de maio de 1991, momento em que os testes foram interrompidos pela greve dos servidores que paralisou o computador de grande porte.

Como ponto de partida para determinar um limite para o tamanho das áreas de ponderação a serem consideradas para aplicação da metodologia MQG2, foram utilizadas 3 áreas de ponderação de tamanhos diferentes definidas um tanto arbitrariamente como subconjuntos dos setores do município de Limeira, quais sejam:

- a área de ponderação 1 (AP1) constituída pelos 10 primeiros setores rurais do município de Limeira;

- a área de ponderação 2 (AP2) constituída pelos 20 setores rurais do município de Limeira;

- a área de ponderação 3 (AP3) constituída pelos 40 primeiros setores urbanos do município de Limeira.

As dimensões de cada área de ponderação, em termos do número de pessoas, domicílios particulares permanentes na população e o número de domicílios particulares permanentes na amostra, são apresentadas na tabela 1 a seguir.

TABELA 1 - Número de pessoas residentes em domicílios particulares permanentes, número de domicílios particulares permanentes e número de setores censitários na população, e número de domicílios particulares permanentes na amostra, segundo as áreas de ponderação

Áreas de ponderação	Dimensões na População			Domicílios na amostra
	Pessoas	Domicílios	Setores	
AP1	5.125	1.251	10	145
AP2	10.195	2.441	20	285
AP3	38.650	10.674	40	1.014

Foram utilizadas para definir as restrições de consistência as informações de cada pessoa para as seguintes variáveis auxiliares investigadas a 100% no Censo 91: sexo, idade, relação de parentesco com o chefe do domicílio, situação do domicílio, número de banheiros do domicílio de moradia e tamanho do domicílio em número de pessoas moradoras.

A partir dessas variáveis auxiliares foram definidas as 34 restrições do grupo 1 tal como apresentadas na seção 5.4.1, sendo 22 referentes a características de pessoas e 12 referentes a características de domicílios.

Além dessas 34 restrições do grupo 1, para serem consideradas a nível de setor e a nível da área de ponderação como um todo, foram definidas para o segundo estágio do processo de ajuste dos pesos (a nível da área de ponderação) 2 restrições para cada setor: uma é a contagem do total de pessoas residentes no setor e outra é a contagem de domicílios particulares permanentes do setor.

Para simplicidade de referência nas listagens de saída da aplicação e nesse texto, as restrições foram codificadas da seguinte maneira: no caso de restrições do grupo 1, a primeira letra (R) representa a palavra Restrição; a segunda letra varia de acordo com o tipo de restrição, sendo P se a restrição for referente a características de pessoas e D se a restrição se referir a características de domicílios; os dois últimos dígitos correspondem ao número de ordem da restrição no grupo. Por outro lado, no caso de restrições do grupo 2 que representam contagens a nível de setor censitário, acrescenta-se uma terceira letra (S), seguida pelo número de ordem do setor da expansão (isto é, após agregação dos setores rarefeitos).

As restrições que foram utilizadas na aplicação experimental acompanhadas dos respectivos códigos são descritas a seguir.

Restrições do grupo 1 - pessoas

RP01. Número total de pessoas;
RP02. Número de pessoas de 5 anos ou mais de idade;
RP03. Número de pessoas do sexo masculino;
RP04. Número de pessoas de 10 anos ou mais de idade do sexo feminino;
RP05. Número de pessoas de 5 a 9 anos de idade;
RP06. Número de pessoas de 10 a 14 anos de idade;
RP07. Número de pessoas de 15 a 19 anos de idade;

RP08. Número de pessoas de 20 a 24 anos de idade;

RP09. Número de pessoas de 25 a 29 anos de idade;

RP10. Número de pessoas de 30 a 34 anos de idade;

RP11. Número de pessoas de 35 a 39 anos de idade;

RP12. Número de pessoas de 40 a 44 anos de idade;

RP13. Número de pessoas de 45 a 49 anos de idade;

RP14. Número de pessoas de 50 a 59 anos de idade;

RP15. Número de pessoas de 60 anos ou mais de idade ou com idade ignorada;

RP16. Número de pessoas residentes na área urbana;

RP17. Número de pessoas do sexo masculino que são chefes de domicílio;

RP18. Número de chefes de domicílio com menos de 20 anos de idade;

RP19. Número de chefes de domicílio com idade de 20 a 29 anos;

RP20. Número de chefes de domicílio com idade de 30 a 39 anos;

RP21. Número de chefes de domicílio com idade de 40 a 49 anos;

RP22. Número de chefes de domicílio de 50 anos ou mais de idade ou de idade ignorada.

Restrições do grupo 1 - domicílios

RD01. Número total de domicílios particulares permanentes;

RD02. Número de domicílios com 1 pessoa moradora;

RD03. Número de domicílios com 2 pessoas moradoras;

RD04. Número de domicílios com 3 pessoas moradoras;

RD05. Número de domicílios com 4 pessoas moradoras;

RD06. Número de domicílios com 5 pessoas moradoras;

RD07. Número de domicílios com 6 ou mais pessoas moradoras;

RD08. Número de domicílios da área urbana;

RD09. Número de domicílios sem nenhum banheiro;

RD10. Número de domicílios com 1 banheiro;

RD11. Número de domicílios com 2 banheiros;

RD12. Número de domicílios com 3 ou mais banheiros ou com número ignorado de banheiros.

Restrições do grupo 2 - pessoas e domicílios

RPS$_t$. Número total de pessoas moradoras em domicílios particulares permanentes no setor t, $t=1,2,\ldots,T$, sendo T o número de setores da expansão da área de ponderação em questão;

RDS$_t$. Número total de domicílios particulares permanentes no setor t, $t=1,2,\ldots,T$, sendo T o número de setores da expansão da área de ponderação em questão.

Cabe notar que as restrições de consistência consideradas para essa aplicação foram definidas, na maioria das vezes, por características simples, evitando-se cruzamentos complexos das variáveis auxiliares disponíveis.

Como já foi mencionado, devido à natureza das restrições de consistência consideradas, o número inicial de restrições de consistência para o segundo estágio do processo de ajuste dos pesos em cada área de ponderação vai variar em decorrência da variação do número de setores de cada área. Esse número total de restrições numa dada área de ponderação pode ser decomposto numa parte fixa (formada pelas 34 restrições do grupo 1) mais uma parte variável (que terá exatamente 2.T restrições nesta aplicação experimental, onde T é o número total de setores da expansão da área de ponderação). A tabela 2 apresenta o número de restrições por grupo segundo as áreas de ponderação definidas para a aplicação experimental da metodologia MQG2.

TABELA 2 - Número total inicial de restrições por grupo segundo as áreas de ponderação

Áreas de ponderação	Número de restrições		
	Total	Grupo 1	Grupo 2
AP1	54	34	20
AP2	74	34	40
AP3	114	34	80

Para cada uma das 3 áreas de ponderação do estudo a aplicação da metodologia MQG2 implicou num processo de análise e eliminação de restrições quando ocorresse uma das seguintes situações (tal como definidas na seção 5.4):

a) restrição rara;

b) restrição redundante ou linearmente dependente;

c) restrição quase redundante ou quase linearmente dependente;

d) restrição extrema - restrição eliminada por resultar na obtenção de pesos ajustados com valores extremos (inferiores a 1 ou superiores a 25).

Dessa forma, uma restrição inicialmente definida pode ter sido excluída por um dos seguintes motivos: ser rara, linearmente dependente, quase linearmente dependente ou extrema. O quadro 1 a seguir apresenta uma relação das restrições que foram eliminadas na aplicação experimental correspondente à AP1 indicando, para cada uma, o motivo da eliminação.

Portanto, das 54 restrições definidas para a AP1 foram excluídas ou eliminadas 24 restrições, sendo 16 a nível da área de ponderação (9 referentes a dados de pessoas e 7 a dados de domicílios) e 8 a nível de setor (6 referentes a dados de pessoas e 2 referentes a dados de domicílios). Analisando os motivos das exclusões, verifica-se que 3 restrições foram excluídas por serem raras, 4 por serem linearmente dependentes, 10 por serem quase linearmente dependente e 7 por causarem a obtenção de pesos extremos.

QUADRO 1 - Restrições eliminadas da AP1 por motivo da eliminação

Restrições	Motivo da eliminação
RP07	extrema
RP12	quase linearmente dependente
RP13	linearmente dependente
RP14	extrema
RP16	rara
RP17	quase linearmente dependente
RP18	rara
RP19	linearmente dependente
RP21	quase linearmente dependente
RD02	linearmente dependente
RD05	extrema
RD06	quase linearmente dependente
RD07	extrema
RD08	rara
RD09	quase linearmente dependente
RD12	linearmente dependente
RPS2	quase linearmente dependente
RPS3	quase linearmente dependente
RPS4	extrema
RPS6	quase linearmente dependente
RPS9	extrema
RPS10	extrema
RDS2	quase linearmente dependente
RDS4	quase linearmente dependente

De acordo com a metodologia MQG2, um único peso ou fator de expansão é associado a cada domicílio da amostra para ser usado na estimação de características de pessoas, famílias e domicílios. Além da vantagem da economia computacional, o uso de um único peso reflete o fato de que as unidades efetivamente selecionadas para a amostra do Censo foram os domicílios, o que confere maior aderência do processo de estimação com o desenho amostral efetivamente usado na coleta, se comparada essa metodologia com as anteriormente usadas para a expansão da amostra dos Censos Demográficos brasileiros.

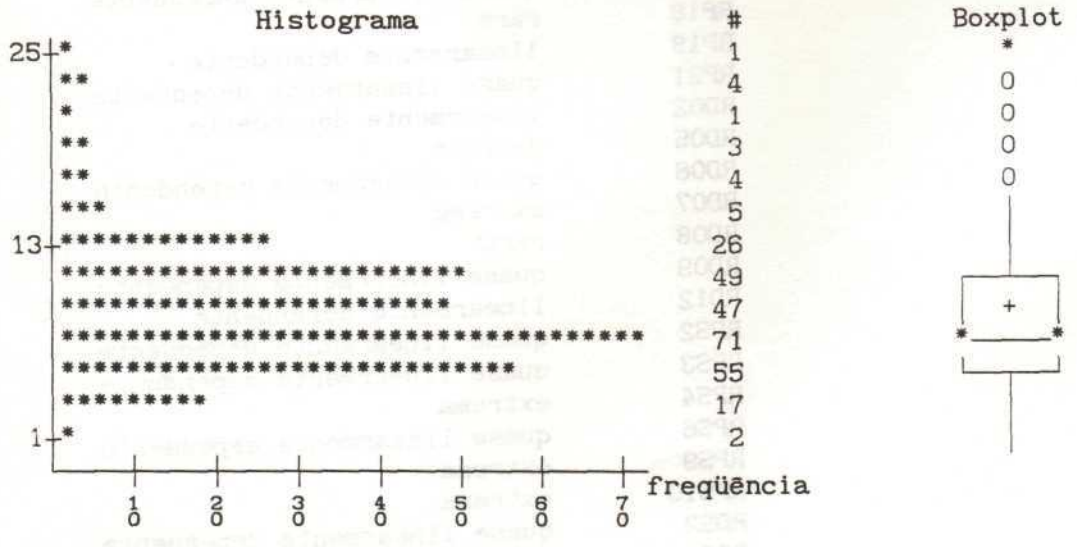
Para cada área de ponderação considerada na aplicação experimental foram calculados pesos ajustados para cada domicílio usando a metodologia MQG2. A título de exemplo, a distribuição dos pesos ajustados da área de ponderação 2 (AP2) é apresentada na figura 2 a seguir.

Figura 2

ANÁLISE DA DISTRIBUIÇÃO DOS PESOS AJUSTADOS
ÁREA DE PONDERAÇÃO 2 - COM 20 SETORES
PESO AJUSTADO FINAL

Diagrama dos 5 Números

	7,99	Mediana
5,97	10,39	Quartis
1,67	24,20	Extremos



* pode representar até 2 valores

Da análise das informações da figura 2 se verifica que os pesos ajustados estão centrados em torno de 8 (valor aproximado da mediana e da média), e não em torno de 10, valor teórico esperado (inverso da fração amostral teórica aplicada na seleção da amostra). Isto reflete dois efeitos: o primeiro devido ao fato de que na área de ponderação 2 a fração efetiva global de amostragem foi de 11,68% (285 domicílios na amostra para 2.441 domicílios na população), ligeiramente superior ao valor previsto (o peso "médio" seria dado pelo inverso dessa fração amostral efetiva e seria de 8,56); o segundo reflete o fato de que a distribuição dos pesos ajustados é ligeiramente assimétrica à direita.

Observa-se também que os valores mínimo e máximo observados para os pesos ajustados se situam dentro dos limites impostos (1 e 25, nessa aplicação experimental), como não podia deixar de ser.

Como já foi dito antes, os pesos ajustados calculados são fracionários. Para obtenção de estimativas a partir da amostra foram usados os pesos ajustados tal como calculados, com todas as casas decimais, efetuando-se o arredondamento para o inteiro mais próximo do resultado da estimativa obtida.

Com a finalidade de avaliar os resultados da aplicação da metodologia MQG2 foram calculadas as estimativas amostrais para os totais das restrições do grupo 1 e do grupo 2 a nível da área de ponderação como um todo, usando os pesos ajustados obtidos. Foram também calculadas as diferenças entre essas estimativas e os respectivos valores dos totais conhecidos da população para cada uma dessas restrições.

As tabelas 3 e 4 apresentam os valores conhecidos da população e as diferenças observadas em relação às estimativas amostrais correspondentes às restrições do grupo 1 para cada uma das 3 áreas de ponderação do experimento, segundo as restrições de pessoas e domicílios, respectivamente.

As diferenças não nulas encontradas nas tabelas 3 e 4 são decorrentes da eliminação das restrições correspondentes. Este fato pode ser observado para a AP1 comparando a lista das restrições eliminadas a nível da área de ponderação apresentada no quadro 1 com as diferenças não nulas encontradas nas tabelas 3 e 4 para essa mesma área de ponderação.

Quando apenas uma restrição de um conjunto de restrições linearmente dependentes tiver sido eliminada, a consistência da estimativa para essa restrição não é perdida. Porém, se após a exclusão dessa restrição, alguma(s) outra(s) restrição(ões) do mesmo grupo for(em) excluída(s), inconsistências podem ser geradas para as restrições excluídas.

TABELA 3 - Total de pessoas da população (POP) e diferença entre a estimativa e o valor da população (DIFP), por área de ponderação, segundo as restrições de pessoas

Restrições de pessoas	AP1		AP2		AP3	
	POP	DIFP	POP	DIFP	POP	DIFP
RP01	5.129	0	10.195	0	38.650	0
RP02	4.587	0	9.101	0	35.659	0
RP03	2.635	0	5.232	0	18.576	-336
RP04	1.954	0	3.873	0	16.966	0
RP05	571	0	1.146	0	3.136	0
RP06	525	0	1.016	0	2.918	0
RP07	463	-48	971	0	3.270	0
RP08	473	0	976	0	3.593	0
RP09	444	0	864	0	3.461	0
RP10	408	0	774	0	2.925	0
RP11	334	0	641	0	2.432	326
RP12	284	47	567	0	2.353	-294
RP13	266	-69	505	-28	2.084	-32
RP14	418	70	801	28	4.190	0
RP15	401	0	840	0	5.297	0
RP16	0	0	0	0	38.650	0
RP17	1.199	31	2.313	39	8.820	0
RP18	13	-13	28	14	57	-19
RP19	215	26	421	26	1.190	-178
RP20	352	0	639	0	2.122	193
RP21	254	-18	509	-32	2.164	0
RP22	422	0	858	0	5.174	0

TABELA 4 - Total de domicílios da população (DOM) e diferença entre a estimativa e o valor da população (DIFD), por área de ponderação, segundo as restrições de domicílios

Restrições de domicílios	AP1		AP2		AP3	
	DOM	DIF	DOM	DIF	DOM	DIF
RD01	1.251	0	2.441	0	10.674	0
RD02	54	-18	111	-31	763	-5
RD03	194	0	364	32	2.178	0
RD04	240	0	481	0	2.363	0
RD05	319	70	571	0	2.568	0
RD06	201	-22	415	1	1.622	0
RD07	243	-30	499	0	1.180	5
RD08	0	0	0	0	10.674	0
RD09	72	0	119	-41	150	-12
RD10	1.007	0	2.066	44	6.977	0
RD11	125	0	201	0	2.287	0
RD12	47	0	52	-3	1.260	12

A análise das diferenças encontradas para as restrições do grupo 1 revela que se consegue garantir a consistência para a maioria das restrições impostas. Quando não se consegue consistência exata, se consegue uma aproximação satisfatória do valor do total conhecido, exceto para totais referentes a características bastante raras (as maiores diferenças relativas se verificam exatamente para estimativas de totais das características mais raras na população).

Uma análise conclusiva, entretanto, só poderá ser feita quando for comparada a consistência das estimativas baseadas na metodologia MQG2 com as baseadas em pesos calculados pelo processo PIETOM, usado no Censo 80, o que entretanto não foi possível fazer até a conclusão deste relatório.

As tabelas 5 e 6 apresentam os valores conhecidos da população e as diferenças das estimativas amostrais para esses valores, referentes às contagens de pessoas e de domicílios de cada setor censitário (restrições do grupo 2), para cada uma das três áreas de ponderação consideradas na aplicação experimental.

TABELA 5 - Total de pessoas da população (POP) e diferença entre a estimativa e o valor da população (DIFP), por área de ponderação, segundo as restrições de pessoas do grupo 2 a nível de setor

Restrições de pessoas	AP1		AP2		AP3	
	POP	DIFP	POP	DIFP	POP	DIFP
RPS1	369	-40	198	5	271	-46
RPS2	416	69	236	-13	587	12
RPS3	441	14	239	-14	610	-21
RPS4	456	-81	369	20	616	-20
RPS5	454	0	416	37	711	-27
RPS6	548	36	419	-36	716	-97
RPS7	595	0	441	10	739	0
RPS8	538	0	454	-15	743	0
RPS9	653	-63	456	-100	772	0
RPS10	659	65	459	0	778	-59
RPS11			538	-1	826	33
RPS12			548	0	834	45
RPS13			595	-24	913	16
RPS14			598	0	938	124
RPS15			653	-80	939	11
RPS16			658	48	942	0
RPS17			659	34	960	-27
RPS18			698	0	961	-13
RPS19			733	198	963	-32
RPS20			828	20	972	10
RPS21					994	-133
RPS22					999	-41
RPS23					1.015	-24
RPS24					1.033	109
RPS25					1.036	76
RPS26					1.072	16
RPS27					1.080	-15
RPS28					1.087	-3
RPS29					1.087	-40
RPS30					1.091	-90
RPS31					1.092	-26
RPS32					1.163	22
RPS33					1.174	22
RPS34					1.178	27
RPS35					1.187	-30
RPS36					1.249	-64
RPS37					1.254	-3
RPS38					1.262	51
RPS39					1.361	65
RPS40					1.445	35

TABELA 6 - Total de domicílios da população (POP) e diferença entre a estimativa e o valor da população (DIFD), por área de ponderação, segundo as restrições de domicílios do grupo 2 a nível de setor

Restrições de domicílios	AP1		AP2		AP3	
	DOM	DIFD	DOM	DIFD	DOM	DIFD
RDS1	75	-12	46	2	87	1
RDS2	105	11	60	9	184	11
RDS3	107	0	68	9	190	4
RDS4	107	2	75	0	192	-9
RDS5	117	0	104	-2	208	1
RDS6	125	0	105	0	211	6
RDS7	132	0	107	0	221	9
RDS8	137	0	107	0	222	-6
RDS9	172	0	117	0	226	0
RDS10	174	0	119	0	229	0
RDS11			125	0	245	0
RDS12			132	0	250	0
RDS13			134	0	256	0
RDS14			137	0	258	0
RDS15			140	0	264	0
RDS16			168	0	266	0
RDS17			172	0	269	0
RDS18			172	0	273	0
RDS19			174	0	273	0
RDS20			179	0	275	0
RDS21					279	0
RDS22					282	0
RDS23					282	0
RDS24					283	0
RDS25					284	0
RDS26					285	0
RDS27					287	0
RDS28					290	0
RDS29					290	0
RDS30					291	0
RDS31					292	0
RDS32					295	0
RDS33					298	0
RDS34					306	0
RDS35					312	-16
RDS36					319	0
RDS37					333	0
RDS38					349	0
RDS39					335	0
RDS40					383	0

A análise das diferenças nas tabelas 5 e 6 também revela comportamento um tanto diferente para a consistência das estimativas de contagens de domicílios e de pessoas. No caso, as contagens de domicílios a nível dos setores são consistentes em número muito maior de casos, o que decorre diretamente do fato de que os pesos iniciais não viciados são calculados como inverso da fração amostral de domicílios em cada setor (da expansão). Para as contagens de pessoas, o número de vezes que a consistência exata é verificada é pequeno, mas as estimativas podem ser consideradas boas aproximações das contagens conhecidas, com algumas poucas exceções. Apesar disso, acredita-se que essa consistência ainda deve ser maior que a que seria obtida caso os pesos fossem ajustados pelo processo PIETOM, que não privilegia a busca de consistência de estimativas a nível de setor.

A interrupção do funcionamento do computador de grande porte do IBGE em maio por causa da greve dos funcionários não permitiu concluir a aplicação experimental que se vinha realizando. Está faltando calcular as estimativas com base em pesos ajustados derivados do processo PIETOM para uma análise comparativa da questão da consistência, bem como está faltando calcular as estimativas do erro quadrático médio das estimativas derivadas da metodologia MQG2 para comparar essas medidas de precisão com as do processo PIETOM. Tais atividades serão retomadas tão logo se restabeleça o funcionamento normal da instituição.

Apesar de estar incompleta a aplicação experimental, decidiu-se divulgar este relatório da situação, como forma de subsidiar debates visando à tomada de decisão com respeito à metodologia que será adotada para expandir a amostra do Censo Demográfico de 1991. Vale registrar um comentário importante que toda a operação de adaptação do programa para executar a aplicação da metodologia MQG2 aqui relatada foram necessários os esforços de apenas 3 técnicos, em regime de dedicação parcial, durante 3 meses, não contado o prazo de elaboração deste relatório, o que mostra ser essa uma alternativa de custo relativamente baixo para implementação, devido até mesmo ao fato de que o principal esforço de adaptação dos programas para o caso brasileiro já ter sido feito.

7. PENDÊNCIAS E LINHAS DE TRABALHO PARA COMPLETAR A ESPECIFICAÇÃO DA METODOLOGIA A SER USADA NA EXPANSÃO DA AMOSTRA DO CENSO 91

Esta seção é dedicada a indicar as principais pendências e linhas de trabalho que devem ser seguidas para viabilizar a especificação completa e detalhada da metodologia a ser empregada para expandir a amostra do Censo Demográfico de 1991. Representa a visão do estágio dos trabalhos que se tem do ponto de vista da Divisão de Metodologia, e deve ser encarada como uma contribuição ao planejamento da atividade de obtenção dos resultados do Censo, na forma de subsídio técnico para a discussão e tomada de decisão.

Antes de qualquer providência técnica, urge retomar o funcionamento do Grupo de Trabalho sobre Uso de Amostragem no Censo Demográfico, para que seja ele o fórum privilegiado do debate técnico acerca da metodologia a ser adotada na expansão da amostra. Esse grupo está desde dezembro de 1991 sem se reunir. Uma primeira tarefa que tal grupo poderia empreender seria uma discussão organizada a partir deste relatório que ora se apresenta, visando posicionar-se sobre a proposta de metodologia aqui contida, e sobre a rejeição da proposta da metodologia GCI formulada por Costa(1991) e que vinha sendo considerada para implementação pela Coordenação Técnica do Censo Demográfico na Diretoria de Pesquisas e na Diretoria de Informática.

Uma vez encaminhada pelo Grupo de Trabalho essa questão, caso a recomendação de uso da metodologia MQG2 seja aceita, falta realizar uma série de tarefas que estão indicadas de forma resumida a seguir.

1) conclusão da aplicação experimental aos dados do Censo Experimental de Limeira de 1988, inclusive com comparação dos resultados com os que seriam obtidos caso a metodologia PIETOM fosse usada; elaboração de relatório complementar sobre essa aplicação;

2) ampliação do experimento via aplicação dessa metodologia a dados de algumas áreas de ponderação, já com base nos resultados disponíveis do Censo 91, para o que seria necessário adiantar a apuração dos dados dessas áreas de ponderação selecionadas para estudo;

3) discussão de critérios alternativos aos existentes para formação das áreas de ponderação a serem usadas na expansão, levando em conta os limites máximos para o tamanho das áreas de ponderação, a idéia de atendimento da demanda por informações mais desagregadas e, se possível, o emprego de critérios automáticos de formação das áreas de ponderação que viabilizem realizar essa tarefa com baixo custo e pouca intervenção manual;

4) estudo para definir as restrições de consistência que devem ser consideradas, considerando as informações disponíveis a 100%, a experiência passada e a disponibilidade de novos métodos e pacotes computacionais que podem auxiliar nessa tarefa, tais como o sistema CART;

5) estudo para definir os valores de vários dos parâmetros de controle necessários à especificação completa da metodologia MQG2, tais como o tamanho mínimo para que setores não sejam considerados rarefeitos, o limite mínimo usado para definir as restrições raras, os limites mínimo e máximo definidos para aceitação dos pesos ajustados, etc.;

6) conclusão da adaptação dos programas que implementam a metodologia MQG2 para funcionamento no IBGE com os dados do Censo Demográfico 91;

7) preparação de especificações detalhadas para a apuração dos dados do questionário da amostra, de forma a satisfazer o requisito de permitir a ligação perfeita dos registros de dados de cada questionário da amostra com o correspondente questionário básico, sem o que a metodologia MQG2 (ou qualquer outra metodologia do mesmo tipo, inclusive o PIETOM) não pode ser aplicada de forma confiável;

8) avaliação da forma de cálculo de precisão proposta na metodologia MQG2 e preparação de proposta sobre a forma de divulgação das medidas de precisão calculadas;

9) documentação detalhada dos programas de computador existentes, elaboração de manuais de operação adequados para utilização desses programas durante a execução da expansão da amostra, e divulgação ampla da nova metodologia proposta, não só entre técnicos do IBGE mas também entre usuários dos resultados amostrais do Censo;

10) preparação de relato das experiências realizadas para comunicação ao *Statistics Canada* sobre os resultados alcançados, problemas detectados e respectivas soluções, sucessos verificados, etc., de modo a dar *feedback* sobre a aplicação da metodologia MQG2 aqui no Brasil, única maneira de garantir a continuidade da cooperação que eles prestaram ao ceder a documentação, os programas e informações sobre a metodologia que hoje se tem disponível no IBGE;

11) preparar uma equipe integrada por técnicos das várias unidades envolvidas com o Censo para tratar da implementação da metodologia proposta, preferencialmente fazendo com que técnicos Diretoria de Informática participem da solução desde o início e trabalhem em estreita colaboração com os técnicos da Diretoria de Pesquisas.

Acredita-se que a implementação das medidas sugeridas permitirá fazer a expansão da amostra do Censo Demográfico 91 com qualidade superior à que se teve no passado, sem problemas de prazo (pois ainda há tempo suficiente para realizar os estudos e preparar as definições ainda necessárias sem comprometer os prazos de apuração do Censo) e com custo adicional bastante modesto, em vista dos investimentos já efetuados em pesquisa metodológica e na obtenção e adaptação de programas de computador, que se mostrou são aplicáveis quase que imediatamente para a implementação da metodologia proposta.

8. REFERÊNCIAS

ALBIERI, S. *A formação de pós-estratos através do uso de árvores de classificação: uma solução para o problema de pós-estratos vazios?*. Rio de Janeiro: IBGE, 1991a. 5 p.

ALBIERI, S. *Censo demográfico brasileiro de 1991 - Métodos de estimação em estudo.* Rio de Janeiro: IBGE, 1991b. 6 p.

ALBIERI, S. *Avaliação da definição das variáveis auxiliares do processo de expansão da amostra do censo demográfico.* Rio de Janeiro: IBGE, 1992.

AN INTRODUCTION to CART methodology. [California]: California Statistical Software, 1985, 19 p.

ARORA, H.R. e BRACKSTONE, G.J. *An investigation of the properties of raking ratio estimators with simple random sampling.* Ottawa: Statistics Canada, 1977a.

ARORA, H.R. e BRACKSTONE, G.J. *Weighting procedures of raking ratio estimation.* Ottawa: Statistics Canada, 1977b.

BANKIER, M.D. *Generalized least squares estimation under poststratification.* Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 730-735. 1989.

BANKIER, M.D. *Two step generalized least squares estimation.* Ottawa: Statistics Canada, 1990. 66 p.

BETHLEHEM, J.G. *LINWEIGHT user manual.* Voorburg: Netherlands Central Bureau of Statistics, 1985. 17 p. (BPA n. 3912-85-M1).

BETHLEHEM, J.G. e KELLER, W.J. *A general method for weighting sample surveys.* Voorburg: Netherlands Central Bureau of Statistics, 1984. 20 p. (BPA n. 1401-84-M1 - first draft).

BETHLEHEM, J.G. e KELLER, W.J. *A general method for weighting sample surveys.* Voorburg: Netherlands Central Bureau of Statistics, 1985. 23 p. (BPA n. 1401-84-M1 - third version).

BETHLEHEM, J.G. e KELLER, W.J. Linear weighting of sample survey data. *Journal of Official Statistics*, v. 3, n. 2, pp. 141-153. 1987.

BIANCHINI, Z.M. *Expansão da amostra do censo demográfico de 1991: análise do método grupo de controle iterativo.* Rio de Janeiro: IBGE, 1991. 45 p.

BLOCH, D.A. e SEGAL, M.R. Empirical comparison of approaches to forming strata: using classification trees to adjust for covariates. *Journal of the American Statistical Association*, v. 84, n. 408, pp. 897-905. 1989.

BRACKSTONE, G.J. e RAO, J.N.K. An investigation of raking ratio estimators. *Sankhya*, série C, v. 41, pp. 97-114. 1979.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A. e STONE, C.J. *Classification and regression trees.* Wadsworth Inc. 1984.

COCHRAN, W.G. *Sampling techniques - third edition.* New York: John Wiley & Sons, 1977. 428 p.

COSTA, L.N. *Processo de expansão da amostra do censo demográfico 1990 - estudos experimentais.* Rio de Janeiro: IBGE, 1987a. 51 p.

COSTA, L.N. Aplicação da amostragem na coleta dos censos demográficos no Brasil. *Revista Brasileira de Estatística*, v. 48, n. 189/190, pp. 35-64. 1987b.

COSTA, L.N. *X recenseamento geral do Brasil: censo demográfico experimental de Limeira - estudos experimentais sobre os processos de ponderação da amostra (projeto).* Rio de Janeiro: IBGE, 1988. 19 p.

COSTA, L.N. *X recenseamento geral do Brasil: censo demográfico experimental de Limeira - estudos experimentais sobre os processos de ponderação da amostra (projeto) - versão 2.* Rio de Janeiro: IBGE, 1989. 19 p.

COSTA, L.N. *X recenseamento geral do Brasil: expansão da amostra do censo demográfico - estudos experimentais aplicados no censo de Limeira - considerações adicionais.* Rio de Janeiro: IBGE, 1991. 19 p.

COSTA, L.N. e LIMA, M.I.F. *X recenseamento geral do Brasil: expansão da amostra do censo demográfico - estudos experimentais aplicados no censo de Limeira.* Rio de Janeiro: IBGE, 1991. 92 p.

GÖTTGENS, R.; ZELLEN, B.; ODEKERKEN, M. e HOFMAN, L. *BASCULA version 1.0 - a weighting package under MS-DOS - user manual.* Voorburg: Netherlands Central Bureau of Statistics, 1991. 121 p. (BPA n. 11176-91-M3).

METODOLOGIA do censo demográfico de 1980. Rio de Janeiro: IBGE, 1983. 477 p.

OH, H.L. e SCHEUREN, F. Modified raking ratio estimation. *Survey Methodology*, v. 13, n. 2, pp. 209-219. 1987.

PINHEIRO, J.C.R.C. e LIMA, J.M. *Avaliação dos efeitos de redução da fração de amostragem no Censo Demográfico.* Rio de Janeiro: IBGE, 1988. 48 p. (Relatórios Técnicos da ENCE, 04/88).

PINHEIRO, J.C.R.C. e LIMA, J.M. Avaliação dos efeitos da redução da fração de amostragem no Censo Demográfico. *Revista Brasileira de Estatística,* Rio de Janeiro, v.51, n.195, pp. 53-84. 1990.

RATHWELL, S. *Requirements for the 1991 census weighting system.* Ottawa: Statistics Canada, 1991. 67 p.

SÄRNDAL, C.E.; SWENSSON, B. e WRETMAN, J.H. *The weighted residual technique for estimating the variance of the general regression estimator.* Stockholm: Statistics Sweden, 1988. 23 p. (Report n. 10).

SAS language and procedures: usage, version 6, first edition. Cary, NC: SAS Institute Inc., 1989a. 638 p.

SAS/STAT user's guide, version 6, fourth edition, v. 1. Cary, NC: SAS Institute Inc., 1989b. 943 p.

SAS/STAT user's guide, version 6, fourth edition, v. 2. Cary, NC: SAS Institute Inc., 1989c. 846 p.

SAS/IML software: usage and reference, version 6, first edition. Cary, NC: SAS Institute Inc., 1989d. 501 p.

SAS language: reference, version 6, first edition. Cary, NC: SAS Institute Inc., 1990a. 1042 p.

SAS procedures guide, version 6, third edition. Cary, NC: SAS Institute Inc., 1990b. 705 p.

SAS guide to MACRO processing, version 6, second edition. Cary, NC: SAS Institute Inc., 1990c. 319 p.

SILVA, L.F. e BIANCHINI, Z.M. *A redução da amostra e a utilização de duas frações amostrais no censo demográfico de 1990.* Rio de Janeiro: IBGE, 1990. 49 p. (Textos para Discussão, nº 33).

SILVA, P.L.N. *Relatório da visita técnica ao Statistics Canada e ao Netherlands Central Bureau of Statistics.* Rio de Janeiro: IBGE, 1991. 44 p.

SILVA, P.L.N. e ALBIERI, S. *Análise da adequação das hipóteses do processo de expansão da amostra de domicílios do Censo Demográfico.* Rio de Janeiro: IBGE, 1991. 18 p.

USING the CART programs: version 1.3. [California]: California Statistical Software, 1991, 38 p.

Textos para Discussão já publicados

- ✓ *Pesquisas Contínuas da Indústria - Vol. 1*, nº 1, janeiro 1988
- ✓ *Pesquisas Agropecuárias Contínuas: Metodologia - Vol.I*, nº 2, 1988
- ✓ *Uma Filosofia de Trabalho: As experiências com o SNIPC e com o SINAPI - Vol. I*, nº 3, março 1988
- ✓ *O Sigilo das Informações Estatísticas: Idéias para reflexão - Vol. I*, nº 4, abril 1988
- ✓ *Projeções da População Residente e do Número de Domicílios Particulares Ocupados: 1985-2020 - Vol. I*, nº 5, maio 1988
- ✓ *Classificacão de Atividades e Produtos, Matérias-Primas e Serviços Industriais: Indústria Extrativa Mineral e de Transformação - Vol. 1*, nº 6, agosto 1988
- ✓ *A Mortalidade Infantil no Brasil nos Anos 80 - Vol. I*, nº 7, setembro 1988
- ✓ *Ensaio sobre o Produto Real da Agropecuária - Vol. I*, nº 9, setembro 1988
- ✓ *Principais Características das Pesquisas Econômicas, Sociais e Demográficas - Vol. I*, número especial, outubro 1988
- ✓ *Novo Sistema de Contas Nacionais, Ano Base 1980 - Resultados Provisórios - Vol. I*, nº 10, dezembro 1988
- ✓ *Pesquisa de Orçamentos Familiares - Metodologia para Obtenção das Informações de Campo* - nº 11, janeiro 1989
- ✓ *De Camponesa a Bóia-fria: Transformações do trabalho feminino* - nº 12, fevereiro 1989
- ✓ *Pesquisas Especiais do Departamento de Agropecuária - Metodologia e Resultados* - nº 13, fevereiro 1989
- ✓ *Brasil - Matriz de Insumo-Produto - 1980* - nº 14, maio 1989
- ✓ *As Informações sobre Fecundidade, Mortalidade e Anticoncepção nas PNADs* - nº 15, maio 1989
- ✓ *As Estatísticas Agropecuárias e a III Conferência Nacional de Estatística* - nº 16, junho 1989
- ✓ *Brasil - Sistema de Contas Nacionais Consolidadas* - nº 17, agosto 1989
- ✓ *Brasil - Produto Interno Bruto Real Trimestral - Metodologia* - nº 18, agosto 1989
- ✓ *Estatísticas e Indicadores Sociais para a Década de 90* - nº 19, setembro 1989
- ✓ *Uma Análise do Cotidiano da Pesquisa no DEREN (As Estatísticas do Trabalho)* - nº 20, outubro 1989
- ✓ *Coordenação Estatística Nacional - Reflexões sobre o caso Brasileiro* - nº 21, novembro 1989

- ✓ *Pesquisa Industrial Anual 1982/84 - Análise dos Resultados* - nº 22, novembro 1989
- ✓ *O Departamento de Comércio e Serviços e a III Conferência Nacional de Estatística* - nº 23, dezembro 1989
- ✓ *Um projeto de Integração para as Estatísticas Industriais* - nº 24, dezembro 1989
- ✓ *Cadastro de Informantes de Pesquisas Econômicas* - nº 25, janeiro 1990
- ✓ *Ensaios sobre a Produção de Estatística* - nº 26, janeiro 1990
- ✓ *O Espaço das Pequenas Unidades Produtivas: Uma tentativa de delimitação* - nº 27, fevereiro 1990
- ✓ *Uma Nova Metodologia para Correção Automática no Censo Demográfico Brasileiro: Experimentação e primeiros resultados* - nº 28, fevereiro 1990
- ✓ *Notas Técnicas sobre o Planejamento de Testes e Pesquisas Experimentais* - nº 29, março 1990
- ✓ *Estatísticas, Estudos e Análises Demográficas - Uma visão do Departamento de População* - nº 30, abril 1990
- ✓ *Crítica de Equações de Fechamento de Empresas no Censo Econômico de 1985* - nº 31, maio 1990
- ✓ *Efeito de Conglomeração da Malha Setorial do Censo Demográfico de 1980* - nº 32, maio 1990
- ✓ *A Redução da Amostra e a Utilização de Duas Frações Amostrais no Censo Demográfico de 1990* - nº 33, junho 1990
- ✓ *Estudos e Pesquisas de Avaliação de Censos Demográficos - 1970 a 1990* - nº 34, julho 1990
- ✓ *A Influência da Migração no Mercado de Trabalho das Capitais do Centro-Oeste - 1980* - nº 35, agosto 1990
- ✓ *Pesquisas de Conjuntura: Discussão sobre Variáveis a Investigar* - nº 36, setembro 1990
- ✓ *Um Modelo para Estimar o Nível e o Padrão da Fecundidade por Idade com Base em Parturições Observadas* - nº 37, outubro 1990
- ✓ *A Estrutura Operacional de Uma Pesquisa por Amostra* - nº 38, novembro 1990
- ✓ *Produção Agrícola, Agroindustrial e de Máquinas e Insumos Agrícolas no Anos 80: Novas Evidências Estatísticas* - nº 39, dezembro 1990
- ✓ *A Inflação Medida pelo Índice de Precos ao Consumidor* - nº 40, janeiro 1991
- ✓ *A Participação Política Eleitoral no Brasil - 1988, Análise Preliminar* - nº 41, fevereiro 1991
- ✓ *Ensaios sobre Estatísticas do Setor Produtivo* - nº 42, março 1991

- ✓ *A Produção Integrada de Estatística e as Contas Nacionais: Agenda para Formulação de um Novo Plano Geral de Informações Estatísticas e Geográficas* - nº 43, março 1991
- ✓ *Matriz de Fluxos Migratórios Intermunicipais - Brasil - 1980* - nº 44, abril 1991
- ✓ *Fluxos Migratórios Intrametropolitanos - Brasil - 1970-1980* - nº 45, abril 1991
- ✓ *A Revisão da PNAD, A Questão Conceitual e Relatório das Contribuições* - nº- 46, maio 1991
- ✓ *A Dimensão Ambiental no Sistema de Contas Nacionais* - nº 47, maio 1991
- ✓ *Estrutura das Contas Nacionais Brasileiras* - nº 48, junho 1991
- ✓ *Mercado do Couro e Resultados da Pesquisa Anual do Couro* - nº 49, junho 1991
- ✓ *As Estatísticas e o Meio Ambiente* - nº 50, julho 1991
- ✓ *Novo Sistema de Contas Nacionais Séries Correntes : 1981-85 Metodologia, Resultados Provisórios e Avaliação do Projeto* - nº 51, julho 1991 (2 Volumes : Volume 1-Metodologia, Resultados Provisórios e Avaliacão do Projeto; Volume 2-Tabelas)
- ✓ *O Censo Industrial de 1985 -- Balanço da Experiência* - nº 52, agosto 1991
- ✓ *Análise da Inflação Medida Pelo INPC 1989* - nº 53, agosto 1991
- ✓ *Revisão da PNAD : A Questão Amostral : Módulo II do Anteprojeto* nº 54, setembro 1991
- ✓ *A Força de Trabalho e os Setores de Atividade - Uma Análise da Região Metropolitana de São Paulo* - 1986-1990 - nº 55, outubro 1991
- ✓ *Revisão da PNAD : Apuração das Informações : Módulo III do Anteprojeto* - nº 56, novembro 1991
- ✓ *Novos Usos para Pesquisa Industrial Mensal : A Evolução dos Salários Industriais, O Desempenho da Pecuária* - nº 57, novembro 1991
- ✓ *Revisão da PNAD : A Disseminação das Informações Módulo IV do Anteprojeto* - nº 58, dezembro 1991
- ✓ *Estatísticas Agropecuárias : Sugestões para o Novo Plano Geral de Informações* - nº 59, dezembro 1991
- ✓ *Análise Conjuntural e Pesquisa Industrial* - nº 60. janeiro 1992
- ✓ *Exploração dos Dados da Pesquisa Industrial Mensal de Dados Gerais* - nº 61, fevereiro 1992