12 Apuração dos dados

Uma vez encerrada a etapa de captura de dados, com sua armazenagem em meio magnético, teve início a etapa de apuração que, em síntese, destinava-se ao tratamento de todas as informações coletadas, preparando-as para a divulgação e disseminação.

Este capítulo descreve todas as etapas de trabalho, começando pelo processo de apuração dos Resultados Preliminares e da Sinopse Preliminar, que tiveram como fonte os dados das Cadernetas dos Setores, chegando à apuração dos Questionários Básicos e da Amostra que deram origem às divulgações dos Resultados do Conjunto Universo e Resultados da Amostra. Contém os procedimentos de: aplicação e verificação de códigos; de crítica, imputação e expansão dos dados, validação dos resultados e tabulação.

Cada uma destas etapas encerra um grau de complexidade elevado. Assim, constituem-se como fatores condicionantes à criação de mecanismos de controle e avaliação de cada uma delas, bem como um planejamento adequado que permita seu encadeamento de forma sincronizada, a fim de garantir o cumprimento de prazos e um padrão de qualidade adequado do processo de apuração.

12.1 Resultados Preliminares

Em dezembro de 2000, os primeiros resultados foram apresentados ao público com a publicação *Censo Demográfico 2000: resultados preliminares*. Divulgada após duas semanas

de encerrada a coleta dos dados, a publicação foi o primeiro retrato da população brasileira e teve, além de outros objetivos, fornecer subsídios ao Tribunal de Contas da União para o estabelecimento das cotas do Fundo de Participação dos Estados e do Fundo de Participação dos Municípios.

O Sistema de Indicadores Gerenciais da Coleta - SIGC foi a fonte de dados para a publicação. As informações correspondentes ao resumo do CD 1.06 – Caderneta do Setor foram disponibilizadas no sistema em diferentes datas, à medida que cada unidade da federação encerrava a sua coleta. Em função disso, os resultados tiveram como referência o dia 11 de dezembro de 2000, data da última atualização no sistema, tendo caráter preliminar, diferindo dos resultados definitivos divulgados posteriormente pelo IBGE.

A publicação *Censo Demográfico 2000: resultados preliminares* constou de uma parte impressa e outra em CD-ROM. O volume impresso continha os comentários sobre a evolução do crescimento da população, mapas da densidade populacional do Brasil, dos estados e de seus municípios, além de tabelas da população recenseada, para todos os municípios do País, segundo o sexo e a situação do domicílio. No CD-ROM, além das tabelas da publicação, constava também toda a Divisão Territorial do Brasil.

A publicação apresentou as informações para os 5 507 municípios criados e instalados em 1º de agosto de 2000 e, em anexo, para os 54 novos municípios, que foram instalados em 1º de janeiro de 2001.

12.2 Sinopse Preliminar

Dando continuidade à divulgação dos resultados preliminares do Censo 2000, foi apresentada, em maio de 2001, a *Sinopse preliminar do censo demográfico 2000*, dando seguimento à série histórica desta publicação, iniciada com o Censo de 1940.

Assim como na publicação dos resultados preliminares, divulgada em dezembro de 2000, os dados da sinopse preliminar tiveram como fonte o SIGC. Foram considerados para divulgação os dados atualizados no sistema até 31 de janeiro de 2001, ou seja, já levando em consideração algumas mudanças decorrentes da atividade de reabertura da coleta de setores em algumas unidades da federação (ver 9.6 Evolução da coleta). Assim, as informações da sinopse também diferiram dos dados definitivos do censo, divulgados mais tarde.

A publicação, além das notas metodológicas, apresentou uma ampla retrospectiva dos dados dos censos desde 1872, acompanhada de textos analíticos sobre a dinâmica da população brasileira e sobre os domicílios.

No que toca ao plano tabular, foram apresentados: os dados sobre população residente, segundo o sexo e situação do domicílio; e domicílios, segundo a espécie, para as grandes regiões, unidades da federação, regiões metropolitanas e todos os municípios do país. No anexo da publicação, encontram-se informações para os 54 novos municípios, instalados em 1º de janeiro de 2001.

No CD-ROM, que acompanhou a publicação impressa, foram reunidas todas as tabelas desse volume; entretanto, a divulgação só atingiu o nível de distrito.

12.3 Resultados do Conjunto Universo

Antes do início do trabalho de crítica e imputação dos dados, foi necessário fazer a formação do Conjunto Universo, a partir das informações comuns do Questionário Básico e do Questionário da Amostra.

A atividade de crítica teve início com a formação dos lotes de trabalho, tendo seguimento com a definição das funções de crítica, que analisaram, em separado, as inconsistências das informações relativas aos domicílios e às pessoas. Para a análise dos dados de pessoa, foi necessário implementar as etapas da crítica intra e entre registros, onde, respectivamente, a investigação e correção dos erros levava em consideração variáveis para uma ou entre pessoa(s) moradora(s) de um mesmo domicílio.

12.3.1 Formação do Conjunto Universo

A formação do Conjunto Universo do Censo Demográfico 2000 consistiu na reunião dos domicílios e das pessoas investigados no Questionário Básico – CD 1.01 – e no Questionário da Amostra – CD 1.02 – associando a cada um(a) deles(as) o conjunto de informações comuns aos dois questionários, ou seja, aquelas coletadas para 100% da população.

As variáveis comuns aos dois questionários são:

a) variáveis de domicílio

- Espécie ou Espécie do domicílio – V0201;
- Tipo ou Tipo de domicílio – V0202;
- Condição de ocupação do domicílio – V0203 no CD 1.01 e V0205 no CD 1.02;
- Condição de ocupação do terreno do domicílio – V0204 no CD 1.01 e V0206 no CD 1.02;
- Forma de abastecimento de água – V0205 no CD 1.01 e V0207 no CD 1.02;
- Canalização da água – V0206 no CD 1.01 e V0208 no CD 1.02;
- Número de banheiros – V0207 no CD 1.01 e V0209 no CD 1.02;
- Existência de sanitário – V0208 no CD 1.01 e V0210 no CD 1.02;
- Tipo de escoadouro – V0209 no CD 1.01 e V0211 no CD 1.02;
- Destino do lixo – V0210 no CD 1.01 e V0212 no CD 1.02;

b) variáveis de pessoa

- Sexo - V0401;
- Relação com a pessoa responsável pelo domicílio - V0402;
- Mês e ano de nascimento – V0403 no CD 1.01 e V0405 no CD 1.02;
- Idade em 31 de Julho de 2000 – V0404 no CD 1.01 e V0406 no CD 1.02;

- Idade presumida – V0405 no CD1.01 e V0407 no CD1.02;

- Sabe ler e escrever – V0406 no CD1.01 e V0428 no CD1.02;

- Curso mais elevado que freqüentou no qual concluiu pelo menos uma série[1]

V0407 no CD 1.01 ou
V0432 no CD 1.02, para a pessoa que não estava freqüentando escola, mas já havia freqüentado; e
V0430 no CD 1.02, para a pessoa que estava freqüentando escola a partir da segunda série; ou o grau do curso imediatamente anterior ao declarado no quesito 4.30, para a pessoa que estava freqüentando escola na primeira série.

- Última série concluída com aprovação[1] – V0408 no CD 1.01 ou

V0433 no CD 1.02, para a pessoa que não estava freqüentando escola, mas já havia freqüentado; e
V0431 no CD 1.02, para a pessoa que estava freqüentando escola a partir da segunda série; ou a última série correspondente ao grau do curso imediatamente anterior ao declarado no quesito 4.30, para a pessoa que estava freqüentando escola na primeira série.

- Rendimento bruto do mês de Julho de 2000[1] – valor declarado no quesito 4.09 no CD 1.01, e o somatório dos registrados nos seguintes quesitos do CD 1.02:

Rendimento no trabalho principal – 4.51;

Rendimento nos demais trabalhos – 4.52;

Proventos de aposentadoria ou pensão – 4.57;

Proventos de aluguel – 4.58;

Proventos de pensão alimentícia, mesada ou doação recebida de não-morador – 4.59;

Proventos de renda mínima, bolsa escola, etc. – 4.60; e

Proventos de outros rendimentos – 4.61.

12.3.2 Crítica e Imputação dos Dados

O desenvolvimento da crítica e imputação das informações do Conjunto Universo tiveram início com a definição dos lotes de apuração que, em última análise, constituíram as bases de dados a serem submetidas aos aplicativos de depuração das inconsistências.

Para o trabalho de detecção e correção das incompatibilidades dos dados do Conjunto Universo foi utilizado o sistema DIA - Detección e Imputación Automática de errores para datos cualitativos, que se acha descrito, de maneira breve, no anexo de CD-ROM desse capítulo[2]. Como o DIA não admite a

[1] No Conjunto Universo, essa variável apresenta valor somente para a Pessoa Responsável pelo Domicílio ou Individual em Domicílio Coletivo

[2] Caso o leitor não conheça o sistema DIA, recomenda-se consultar o anexo, antes de dar seqüência à leitura do capítulo.

utilização de funções de crítica que envolvam variáveis de registros distintos, foi necessário implementar uma estratégia que permitisse superar essa limitação do sistema.

Foram constituídos lotes de trabalho e, em cada um, a seqüência de execução dos aplicativos através do sistema DIA foi a seguinte: Características da Pessoa, que tratava as críticas entre registros; e, para as críticas intra registros, Características do Domicílio, Características da Pessoa Responsável pelo Domicílio ou Individual em Domicílio Coletivo, e Características das Demais Pessoas.

12.3.2.1 Formação dos lotes

Assim que os dados do SIGC estavam disponíbilizados, estabeleceram-se critérios para a formação dos lotes de apuração, com vistas à execução da detecção e correção automática dos erros através do sistema DIA. Esses critérios foram os mesmos utilizados no Censo Demográfico 1991.

Um fator importante para a definição dos critérios de formação dos lotes é a proximidade geográfica, o que significa ter os questionários de uma mesma região geográfica em um mesmo lote, o que constitui-se num fator de homogeneidade de características. Essa homogeneidade é importante em função da metodologia de imputação utilizada, que se baseia na distribuição dos dados dos registros "bons" observados no lote, entendendo-se como tal aqueles que não apresentam qualquer erro, segundo as regras de crítica definidas.

Os critérios para a formação dos lotes basearam-se nos seguintes pontos:

a) obtenção do menor número possível de lotes, para minimizar o número de relatórios a serem analisados, após cada aplicação do sistema DIA;

b) obtenção de um tamanho mínimo a fim de viabilizar o processo de correção, tendo em vista a utilização das distribuições de registros "bons", como base da imputação; e

c) geração dos lotes levando em conta a situação do domicílio (urbana e rural), bem como a divisão geográfica do país, contemplando as partições de cada unidade da federação, através da ordenação dos respectivos setores segundo a mesorregião, microrregião, município, distrito e subdistrito.

A quantidade de domicílios particulares ocupados – DPO – fornecida pelo SIGC - foi o ponto de partida para o processo de formação dos lotes. A escolha de um tamanho máximo de 90 000 domicílios para cada lote, foi feita levando em conta um acréscimo ao número estabelecido para o Censo de 1991, que foi de 70 000.

A quantidade de lotes a serem formados, segundo a situação do domicílio, em cada unidade da federação, foi o resultado da divisão do número de domicílios particulares ocupados por 90 000, arredondado para o inteiro seguinte.

O tamanho, aproximado, de cada lote em cada unidade da federação, por situação do domicílio, foi obtido pela divisão do correspondente DPO pelo número de lotes encontrado. De posse desse tamanho aproximado, cada lote foi formado, fazendo-se os cortes na relação ordenada de setores citada no item "c".

Baseando-se nesses critérios, foram gerados 526 lotes de apuração, sendo 429 urbanos e 97 rurais, que vão apresentados na tabela seguinte, acompanhados do número de domicílios particulares ocupados, informados no SIGC, por unidade da federação e situação do domicílio.

Tabela 12.1 - Número de lotes e quantidade de domicílios particulares ocupados no SIGC, do Conjunto Universo, por situação do domicílio, segundo as Unidades da Federação

Unidades da Federação	Situação do domicílio			
	Urbana		Rural	
	Domicílios particulares ocupados	Número de lotes	Domicílios particulares ocupados	Número de lotes
Brasil	**37 455 153**	**429**	**7 567 163**	**97**
Rondônia	229 944	3	121 533	2
Acre	91 093	2	39 619	1
Amazonas	455 707	6	123 199	2
Roraima	59 368	1	16 135	1
Pará	913 540	11	411 420	5
Amapá	89 387	1	10 100	1
Tocantins	212 501	3	70 667	1
Maranhão	757 926	9	484 191	6
Piauí	430 566	5	233 103	3
Ceará	1 294 941	15	468 621	6
Rio Grande do Norte	505 278	6	168 565	2
Paraíba	624 314	7	227 840	3
Pernambuco	1 558 239	18	420 682	5
Alagoas	463 455	6	192 227	3
Sergipe	320 653	4	118 685	2
Bahia	2 218 482	25	976 917	11
Minas Gerais	3 977 365	45	805 740	9
Espírito Santo	685 193	8	160 242	2
Rio de Janeiro	4 107 268	46	157 028	2
São Paulo	9 756 179	109	639 303	8
Paraná	2 216 678	25	464 969	6
Santa Catarina	1 205 879	14	298 814	4
Rio Grande do Sul	2 518 408	28	534 051	6
Mato Grosso do Sul	480 028	6	89 406	1
Mato Grosso	525 143	6	134 112	2
Goiás	1 231 918	14	177 167	2
Distrito Federal	525 700	6	22 827	1

Fonte: IBGE, Censo Demográfico 2000, Sistema de Indicadores Gerenciais da Coleta.

12.3.2.2 Tratamento das omissões da variável "espécie do domicílio"

Uma forma de resolver essa questão, é fazer a imputação dos valores em branco dessa variável – V0201 – levando em conta o preenchimento ou não da seqüência dos demais campos do bloco Características do Domicílio. No entanto, preferiu-se não adotar essa estratégia, pela existência de erros de preenchimento que trariam prejuízo à imputação.

Assim, durante a crítica intra-registros, quando da execução do Aplicativo Características do Domicílio, essa variável foi tratada como fixa no sistema DIA, o que recomendava um tratamento prévio que eliminasse as omissões de informação. Resolveu-se, então, considerar o preenchimento do quesito 1.09 – Número na Folha de Domicílio Coletivo, pertencente ao bloco Identificação. A solu-

ção implicou na criação da variável auxiliar V1090 “existência de domicílio coletivo”que classificava o domicílio em particular ou coletivo, respectivamente, quando a V0109 assumisse o valor zero ou outro qualquer.

Assim, a detecção de omissão e a conseqüente imputação determinística para a variável V0201 foi feita através de procedimento específico, implementado durante o processo de formação dos lotes de trabalho a serem submetidos aos aplicativos do DIA, da seguinte forma:

- V0201 era igual a 1 – Particular permanente –, quando a V1090 indicasse a não-existência de domicílio coletivo; e .

- V0201 era igual a 3 – Coletivo –, quando a V1090 indicasse o contrário.

12.3.2.3 Tratamento das omissões da variável "sexo"

A princípio, não estava previsto fazer o tratamento prévio das situações de omissão da variável “sexo” V0401. Assim, tentou-se tratar essa situação realizando sua depuração de acordo com as estratégias definidas para o aplicativo Características da Pessoa – na crítica entre registros – ou seja, em conjunto com as demais inconsistências. Entretanto, a análise de alguns dos resultados da imputação mostrou a inconveniência da utilização desse procedimento.

A solução encontrada foi executar um aplicativo DIA, apenas para a correção das omissões na variável “sexo”. A estratégia elaborada para esse aplicativo foi fazer a imputação através de distribuição conjunta, baseada na variável auxiliar criada V4702 “grupo qüinqüenal de idade” e na variável V0402 “relação com a pessoa responsável pelo domicílio” através do método proporcional.

É importante ressaltar que a variável “sexo” poderia, durante a execução do aplicativo seguinte, Características da Pessoa, sofrer nova alteração, caso a categoria que lhe fora atribuída ficasse inconsistente perante o conjunto estabelecido para as funções de crítica entre registros.

12.3.2.4 Crítica entre registros

A crítica dos dados do Conjunto Universo que levava em conta as regras de crítica entre registros, foi realizada pelo Aplicativo Características da Pessoa, sendo apenas objeto da imputação as variáveis V0401 e V0402.

Inicialmente, para que os lotes pudessem ser submetidos ao sistema de crítica foi necessária a execução de um programa de ordenação lógica das pessoas em cada domicílio. Os critérios para essa ordenação foram definidos levando-se em conta os procedimentos estabelecidos no Manual do Recenseador para a elaboração da lista de moradores e a idade das pessoas; no documento *Esquema de ordenação lógica das pessoas no questionário básico - censo 2000* (2001), podem ser consultados mais detalhes desse trabalho.

Para que a crítica entre registros pudesse ser executada através do DIA, foi necessária uma estratégia especial de criação de um novo arquivo, onde as informações das pessoas moradoras de um mesmo domicílio foram rearrumadas, de forma a comporem um único registro, considerando-se todos os domicílios com até quarenta moradores.

A variável V0402 foi imputada através da distribuição condicional, que levava em conta a variável auxiliar V0702 “grupo de idade da pessoa” através do método proporcional.

Embora durante a execução da crítica entre registros o total de pessoas do domicílio estivesse correto, pois a informação já havia sido tratada pela crítica quantitativa nos Centros de Captura de Dados – CCDs, os totais por sexo poderiam sofrer alterações em razão das imputações realizadas. Em razão disso, após a execução do DIA, os totais por sexo tiveram que ser recalculados.

A existência de domicílios com mais de quarenta moradores pôde ser constatada nas seguintes Unidades da Federação: Rio Grande do Sul; Mato Grosso; Pará; Pernambuco; Minas Gerais; São Paulo; Sergipe e Ceará. Para cada uma dessas unidades, encontrou-se apenas um único domicílio cujo número de moradores era, respectivamente: 49; 62; 64; 42; 54; 54; 43 e 59. Dessas situações, em quatro unidades – RS, MT, PA e MG – detectou-se inconsistências na crítica entre registros; esses casos foram corrigidos manualmente.

12.3.2.5 Crítica intra-registros

As condições de imputação dos aplicativos do sistema DIA, para as funções de crítica intra-registros dos dados do Conjunto Universo, são apresentadas a seguir.

a) Aplicativo Características do Domicílio

Neste aplicativo poderiam ser imputadas todas as variáveis do bloco 2 – Características do Domicílio - com exceção da V0201 "espécie do domicílio", já consistente, conforme explicado anteriormente e, portanto, mantida fixa durante a execução do aplicativo.

As variáveis foram imputadas de acordo com as respectivas distribuições marginais formadas pelas freqüências dos registros não suspeitos, através do método proporcional.

Para as variáveis V0203 "condição de ocupação do domicílio" e V0207 "número de banheiros" foram atribuídos pesos 2 e 1, respectivamente, diferentemente das demais, cujos pesos foram mantidos em 5, peso médio da escala de confiança na variável. O motivo para essa alteração decorreu dos resultados das análises efetuadas, onde se constatou leve mudança em algumas distribuições dessas variáveis antes e depois da imputação.

b) Aplicativo Características da Pessoa Responsável pelo Domicílio ou Individual em Domicílio Coletivo

Neste aplicativo podiam ser imputadas as seguintes variáveis: idade, sabe ler e escrever, curso mais elevado que freqüentou, no qual concluiu, pelo menos uma série e a última série concluída com aprovação.

Considerando-se que a "idade" pode ser obtida através do mês e ano de nascimento, ou da idade em 31 de julho de 2000, ou ainda, da idade presumida, foi necessário criar um algoritmo que, levando em consideração critérios para esses três quesitos, chegava a informação da idade a ser tratada nesse aplicativo. Embora o algoritmo fizesse a escolha, era possível que a informação da idade passasse por correção, durante a execução do DIA, visto que o sistema poderia identificar inconsistências de acordo com as regras de crítica em que esta variável estivesse envolvida.

Historicamente, admite-se que a informação sobre a idade constitua-se num dado com elevado grau de confiança. Por esse motivo, como critério de imputação, atribuiu-se-lhe peso 1, enquanto que as demais variáveis tiveram o peso médio 5, garantindo-se, desse modo, que a idade fosse, em relação às demais, proporcionalmente, bem menos imputada.

Para a imputação da "idade", utilizou-se a distribuição conjunta, a partir da criação da variável auxiliar V4040 "faixa de idade do cônjuge" e da variável V0402.

A variável "sabe ler e escrever" foi imputada de acordo com a distribuição conjunta com a variável "idade", utilizando-se a distribuição dos registros não-suspeitos e o método proporcional.

As variáveis "curso mais elevado que freqüentou, no qual concluiu, pelo menos uma série" e "última série concluída com aprovação", foram imputadas pelo DIA, de acordo com a situação, através de método determinístico ou probabilístico. A imputação determinística passou a ser uma estratégia em virtude das eventuais inconsistências oriundas de informações errôneas entre a série e o grau, envolvendo mudanças no sistema de ensino brasileiro, ao longo do tempo. Nos casos em que a estratégia foi a imputação probabilística, utilizou-se a distribuição marginal dos registros não-suspeitos e o método proporcional.

c) Aplicativo Características das Demais Pessoas

Os procedimentos utilizados para a execução deste aplicativo foram os mesmos já descritos para o aplicativo anterior. Houve, apenas, a necessidade de se criar a variável auxiliar V4041 "faixa de idade da pessoa responsável ou individual em domicílio coletivo" para imputação da idade das demais pessoas.

12.3.2.6 Análise do processo de crítica e imputação

No Censo Demográfico 2000, os procedimentos de crítica e imputação dos dados foram constantemente monitorados a fim de evitar a alteração na estrutura da informação. Vários foram os instrumentos utilizados com esse objetivo, como as tabelas (conjunto de tabelas que envolvem o aplicativo), a análise demográfica, estudos de população e o controle das alterações nas respostas originais constantes do questionário.

a) Análise dos relatórios do Sistema DIA

O trabalho de análise do processo de crítica e imputação dos Resultados do Universo foi desenvolvido, para a crítica entre registros e para cada um dos aplicativos da crítica intra registros, em duas partes: a que permitia avaliar a correção automática dos erros detectados em nível de cada lote de trabalho e a outra, com o mesmo objetivo, abrangendo os municípios e alguns subdistritos selecionados. Os elementos para a realização dessa tarefa constam do Plano de Análise da Correção Automática e Elementos de Apoio para a Análise da Composição do Lote – CD 1.01 – Questionário Básico.

Obedecendo às determinações do plano, para a investigação em nível de lote eram emitidos relatórios que apresentavam dados gerais sobre o resultado da imputação, informando, por exemplo:

- os totais de registros, de registros bons e de registros com erros, em valores absolutos e relativos;

- a participação de cada tipo de erro em relação ao total de registros;

- os registros segundo o tipo de imputação, apresentando variáveis com valores inválidos, com inconsistências entre variáveis; e

- número de variáveis imputadas por número de registros.

Além dessas informações, toda vez que um lote era considerado suspeito em algum aplicativo, era também emitido o relatório Tablas, parte integrante do sistema DIA. Através desse relatório, era possível analisar, para cada variável, as distribuições de freqüências de entrada e de saída dos dados, assim como, as distribuições dos registros bons e os não-suspeitos, procurando identificar distorções significativas resultantes do processo de imputação.

Um lote era considerado suspeito quando apresentasse alguma variável fixa com valor inválido, ou atingisse o limite de tolerância estabelecido em, pelo menos, um dos seguintes indicadores;

- E - percentual de registros com erro em relação ao total de registros; e

- F - percentual de registros que falharam em cada regra de crítica em relação ao total de registros com erro.

Esses indicadores, calculados após a imputação para cada um dos aplicativos da crítica intra-registros, tinham como limites máximos 10% e 50%, respectivamente, para E e F.

No entanto, era necessário estabelecer um outro nível de investigação que permitisse uma análise mais desagregada dos registros, de modo a possibilitar a identificação de eventuais distorções proporcionadas pela imputação, não-sensíveis no nível agregado de lote.

Desenvolveu-se, então, no plano de análise, os critérios para a emissão de relatórios que permitissem efetuar a investigação para municípios ou subdistritos suspeitos. Isso permitiu realizar o trabalho em dimensão bem próxima ao das análises estruturais desenvolvidas pelos especialistas das diversas áreas temáticas da DPE.

Um município ou subdistrito foi considerado suspeito, caso alcançasse os limites de tolerância para, pelo menos, um dos indicadores seguintes:

E - já definido anteriormente;

Ij - percentual de registros em que a variável j apresentou valor inválido em relação ao total de registros;

$\max_i D_j(i)$ - maior distância em termos relativos, entre as freqüências marginais dos dados bons (FB) e dos dados depurados (FD), para o código i (valores possíveis) da variável j, onde:

$$D_j(i) = \left| \frac{FD_j(i)}{FD_j} - \frac{FB_j(i)}{FB_j} \right| x100$$

FD = nº total de registros depurados da variável j

FB = nº total de registros bons da variável j

$\max_i A_j(i)$ - maior distância em termos relativos, entre as freqüências marginais dos dados de entrada (FE) e dos dados depurados (FD), para o código i da variável j, onde:

$$A_j(i) = \left| \frac{FD_j(i) - FE_j(i)}{FD_j} \right| x100$$

FD = nº total de registros depurados da variável j

FE = nº total de registros de entrada da variável j

Tj - distância entre as freqüências dos dados bons (FB) e os dados depurados (FD), para a variável j, em termos relativos, sendo n o número de códigos possíveis para a variável j, onde:

$$T_j = \sum_{i=1}^{n} \frac{FD_i - FB_i}{\frac{FD}{2}}$$

Bj - distância entre as freqüências dos dados de entrada (FE) e dos dados depurados (FD) para a variável j, em termos relativos, sendo n o número de códigos possíveis para a variável j, onde:

$$B_j = \sum_{i=1}^{n} \frac{FD_i - FE_i}{\frac{FD}{2}}$$

O indicador Ij foi somente calculado para os aplicativos "Características da Pessoa Responsável pelo Domicílio ou Individual em Domicílio Coletivo" e "Características das Demais Pessoas". Quanto aos limites de tolerância, foram estabelecidos 10% e 5% para, respectivamente, E e Bj e 3% para os demais indicadores.

Tanto para o lote quanto para o município ou subdistrito, fundamentalmente o processo de avaliação da imputação através das tablas, concentrava-se na análise dos registros de entrada e depurados, procurando-se identificar alterações nas distribuições das variáveis. Como suporte a esse trabalho, caso necessário, eram consultadas também as informações sobre a composição do lote da crítica.

Qualquer problema – nos dados gerais ou nas tablas – encontrado no resultado da imputação, era encaminhado aos analistas temáticos para o exame da consistência das variáveis correspondentes ao aplicativo implementado.

b) Análise da consistência da imputação e validação dos resultados

É importante fazer uma primeira observação com respeito à Análise do processo de Crítica e Imputação e à dificuldade em separá-la do item validação dos resultados.

Procederam-se análises para algumas unidades da federação e, em alguns casos, alguns municípios onde era estudado o comportamento de cada quesito do questionário antes e depois da imputação automática. Para isso, foram utilizados a visualização de imagens, a listagem do registro completo das pessoas contendo as variáveis a serem analisadas e matrizes de contingência conforme modelo a seguir.

Figura 12.1 - Matriz de contingência com vetores antes e após o processo de imputação

VPOSTERIOR / VANTERIOR	Total 1	%	1	2	3	...	j	...	n
Total2									
%									
1									
2									
3									
...									
j									
...									
m									

Onde:

VANTERIOR = código do quesito antes da correção automática

VPOSTERIOR = código do quesito após a correção automática

A acumulação de valores na diagonal indica ausência de modificações no processo, portanto constituíram o alvo do estudo os casos em que houve significativo aumento, ou redução, de frequência de casos observados em alguma categoria fora da diagonal. As distorções, em geral, desapareceram com a revisão e alteração de algumas regras contidas no plano de crítica. Em alguns casos elas estavam justificadas por se tratar de correção de erros sistemáticos.

As listagens dos registros e as matrizes de contingência foram obtidas através da utilização do REDATAM+G4 (REcuperação de DAdos para Áreas pequenas por Microcomputador, 4ª Geração), um programa computacional desenvolvido pelo Centro Latino-americano e Caribenho de Demografia – CELADE. Com este objetivo foi feita a junção dos arquivos antes e depois da imputação.

Variáveis de domicílio

Na fase de crítica e imputação dos dados de domicílios do Conjunto Universo do Censo Demográfico 2000 foram realizadas análises de consistência dos resultados de cada característica investigada em relação aos obtidos no Censo Demográfico 1991, utilizando indicadores da Pesquisa Nacional por Amostra de Domicílios na década de 1990 como balizamento de tendência. Buscou-se verificar, também, a ocorrência de efeitos de registros inadequados que pudessem ter acontecido na etapa de coleta ou de crítica e imputação dos dados, considerando,

separadamente, os resultados obtidos por meio dos Questionários Básico e da Amostra para as parcelas urbana e rural e, ainda, o cruzamento de determinadas características investigadas. Com base nessas análises constatou-se que, a maior incidência de falhas oriundas da fase de coleta foram decorrentes da desobediência da seqüência dos quesitos nos casos em que a rota era definida em função das respostas registradas. Constatou-se, também, que apenas nos quesitos que geravam seqüências distintas em função das respostas registradas e para aqueles que deveriam ser seguidos somente em função de determinada respostas registradas no anterior, o procedimento geral de crítica e imputação aplicado para correção dos erros de seqüência e dos registros omitidos apresentava efeito perceptível na distribuição dos resultados dos itens. Para evitar esta ocorrência, foram adotados procedimentos especiais para correção e imputação, existentes no próprio sistema DIA, conforme descritos no item 12.3.2.5 a) Aplicativo Características do Domicílio.

Sexo e idade

Para validar o grau de precisão dos resultados das declarações de sexo e de idade dos entrevistados, calcularam-se alguns indicadores demográficos, objetivando verificar as imperfeições nas declarações e se as informações eram coerentes com a tendência observada ao longo dos censos. Os métodos de Myers, Bachi e Whipple (SHRYOCK et al., 1971) foram utilizados para avaliar o grau de atração e repulsão exercido pelos dígitos terminais.

Os índices foram calculados com as informações originais do campo, sem aplicação da correção da crítica qualitativa e com as informações corrigidas, quer dizer, após a passagem da correção automática para o ano 2000 e comparada com os censos a partir de 1940.

Índices preferenciais calculados

Uma forma de comprovar a coerência interna das respostas é verificar a tendência dos informantes em declarar determinados dígitos terminais para a idade. Geralmente é comum o homem declarar-se com 21 anos, porque corresponde à maioridade, enquanto que as mulheres procuram reduzir a idade. As informações errôneas também podem ser causadas por razões econômicas, sociais, políticas ou puramente individuais. Normalmente, existe uma tendência a arredondar a idade, acumulando-se, portanto declarações em idades terminadas em 0 e 5 anos.

O Índice de Myers pode assumir valores entre zero e cento e oitenta correspondendo, respectivamente, a informações de idade prestadas com exatidão e a todas as declarações de idades terminadas pelo mesmo dígito. Os índices calculados com os resultados do Censo Demográfico 2000 não revelaram qualquer variação entre as informações originais, sem tratamento de crítica e as corrigidas, após a passagem da correção automática. E, quando cotejados com o Censo Demográfico 1991, revelaram que passou a existir um ligeiro crescimento na atração por determinados dígitos terminais. Isto é natural, considerando que o Censo Demográfico 1991 foi o único que não foi realizado em ano finalizado em zero. Na análise por sexo, contrariando uma tendência dos censos, o menor grau de precisão da declaração de idade foi proveniente das informações dos homens.

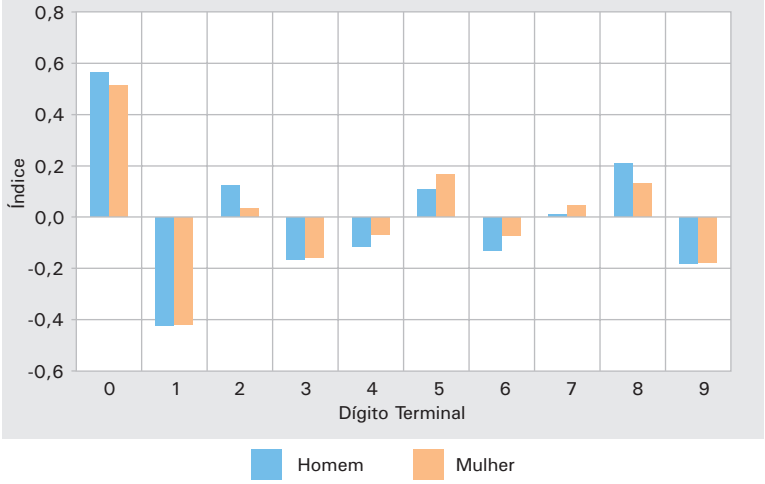
Tabela 12.2 - Índices preferenciais calculados para a população residente
Brasil

Ano	Índices preferenciais calculados		
	Myers	Bachi	Whipple
1940	17,9	12	148
1950 (1)	14,4	9,6	140
1960 (1)	17,5	11,1	143
1970	10,2	6,6	128
1980	4,1	2,6	111
1991	1,3	0,8	103
2000			
Informações originais	1,9	1	104
Informações corrigidas	1,9	1	104

Fonte: IBGE, Censo Demográfico 1940/2000.
(1) População presente.

A preferência por alguns dígitos terminais, no Censo Demográfico 2000 foi especialmente pelo zero, seguido pelos dígitos oito e cinco, e a repulsão foi pelo um, seguido pelo nove. O ano do levantamento apresenta uma certa influência na idade, porque nos censos terminados em zero a tendência é de dígito terminal atrativo zero e repulsivo um (1). No Censo Demográfico 1991, o dígito terminal preferido foi especialmente o cinco e o repulsivo foi o nove.

Gráfico 12.1 - Índice de Myers com as informações corrigidas



O Índice de Whipple tem como objetivo medir a concentração das declarações nas idades terminadas pelos dígitos zero e cinco. O índice de Whipple é o quociente entre duas distribuições de população. O numerador corresponde ao somatório do número de pessoas nas idades pontuais terminadas em zero e cinco a partir de 25 anos até 60 anos de idade multiplicada por cinco, assumindo a hipótese de linearidade no intervalo, e o denominador corresponde ao somatório das pessoas no intervalo de idade entre 23 e 62 anos.

$$IW = [\, 5\, [P(25)+P(30)+P(35)+.....+P(55)+P(60)]\, x\, 100] \,/\, {}^{62}\Sigma_{x=23}\, P(x)$$

onde,

P(x) = população na idade x.

A classificação utilizada pelo *Demographic yearbook* é a seguinte: IW < 105 correspondem a dados muitos precisos e IW > 175 correspondem a dados pouco refinados. Como o resultado foi IW = 104, considerou-se que os dados são precisos e não apresentaram diferença entre os sexos. Em geral existe a tendência à declaração da idade em certos números, especialmente nos que terminam em zero ou cinco, seja porque os entrevistados não conhecem exatamente sua idade ou porque não compreendem a importância de declarar a idade exata.

O método de Bachi aplica o método de Whipple repetidamente para determinar a extensão de preferência para cada dígito final e a partir daí baseia-se, tal como o Índice de Myers, na soma dos desvios, tomados positivamente, entre a freqüência relativa com que cada dígito de zero a nove ocorreu em um determinado levantamento e a freqüência relativa esperada, caso não houvesse preferência por nenhum dígito (distribuição uniforme). Os resultados obtidos pelo método de Bachi se assemelham aos obtidos pelo método de Myers.

As distorções e as falhas nas declarações de sexo e idade são menores de um censo para outro, em função da diminuição da dificuldade da população em informar sua idade com precisão, o que reflete positivamente na qualidade dos diversos indicadores. Os índices atingiram magnitude tão baixa que as oscilações podem ser consideradas desprezíveis.

Outros indicadores também foram calculados para avaliar as possíveis distorções da estrutura por sexo e idade da população, tais como as pirâmides etárias, razões de masculinidade, as razões de idade dentre outros.

Sexo ignorado

Em um primeiro momento, o critério para imputação do sexo, nos casos em que esta informação não foi coletada, levou em consideração somente a observação da variável "relação com o responsável pelo domicílio". Este procedimento não apresentou resultados satisfatórios, uma vez que levou à imputação, em maior quantidade, de sexo masculino nas idades mais avançadas, pois as pessoas envolvidas eram, na sua maioria, responsáveis por domicílio. O resultado não condizia com o esperado, que seria uma proporção maior de mulheres nessas referidas idades.

Foi observado que a ausência de declarações de sexo se concentrava nas primeiras idades e nas mais avançadas. Então, foi solicitada a inclusão da variável "grupos de idade qüinqüenais" como condicionante do sexo a ser imputado – ver item 12.3.2.3. Dessa forma, se o sexo ignorado fosse observado numa pessoa com idade mais avançada, haveria uma maior probabilidade de ser imputado sexo feminino, caso contrário, masculino. Por outro lado, nas primeiras idades a probabilidade de ser imputado o sexo masculino seria maior que a de ser imputado feminino.

Relação com o responsável pelo domicílio

Em alguns questionários de domicílios particulares, o quesito "Qual é a relação com o responsável pelo domicílio?" havia sido preenchido incorretamente. Ao invés de ter sido assinalada, para a primeira pessoa, a quadrícula "pessoa responsável", havia sido assinalada a quadrícula "individual em domicílio coletivo".

De acordo com a ordem lógica estabelecida no manual de crítica, esta pessoa passaria à posição de outro membro do domicílio, enquanto a pessoa que ocupasse a segunda posição no questionário passaria à primeira pessoa e, consequentemente, a ser a responsável pelo domicílio.

Fazendo-se um estudo das características das pessoa que ocupavam a primeira posição do questionário, comparando-a com as demais pessoas e considerando a estrutura domiciliar, observou-se que tratava-se de erro de preenchimento da informação referente à responsabilidade pelo domicílio.

Com o objetivo de que não fosse alterada a estrutura domiciliar, foi acrescido aos critérios de ordem lógica, a manutenção dessa pessoa como primeira pessoa moradora no domicílio. Assim os responsáveis com a quadrícula "individual em domicílio coletivo" indevidamente assinalada, sofreriam acerto nesta informação quando da consistência feita pelo sistema DIA.

Freqüências de imputação

As tabelas 12.3 a 12.6 mostram a freqüência de imputações das variáveis do Conjunto Universo, relativas a domicílios e pessoas. Apresentam também o número de domicílios e pessoas, que foram objeto de imputação.

Tabela 12.3 - Registros imputados, segundo as variáveis de domicílio - Brasil

Variável	Total	Sem imputação		Com imputação	
		Absoluto	Relativo (%)	Absoluto	Relativo (%)
V0201	45 507 516	45 337 228	99,63	170 288	0,37
V0202	45 507 516	44 920 635	98,71	586 881	1,29
V0203	45 507 516	44 906 225	98,68	601 291	1,32
V0204	45 507 516	44 651 605	98,12	855 911	1,88
V0205	45 507 516	45 121 014	99,15	386 502	0,85
V0206	45 507 516	45 044 603	98,98	462 913	1,02
V0207	45 507 516	45 204 565	99,33	302 951	0,67
V0208	45 507 516	44 942 066	98,76	565 450	1,24
V0209	45 507 516	45 260 342	99,46	247 174	0,54
V0210	45 507 516	45 263 848	99,46	243 668	0,54

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.4 - Imputação nos registros de domicílio, segundo o aplicativo - Brasil

Aplicativo	Absoluto	Relativo (%)
Total	**45 507 516**	**100,00%**
Sem imputação	42 348 158	93,06%
Imputação somente na V0201	52 568	0,12%
Imputação pelo Aplicativo "Características do Domicílio" (V0202 a V0210)	2 989 070	6,57%
Imputações V0201 e pelo Aplicativo "Características do Domicílio" (V0202 a V0210)	117 720	0,26%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.5 - Registros imputados, segundo as variáveis de pessoa - Brasil

Variável	Total	Sem imputação		Com imputação	
		Absoluto	Relativo (%)	Absoluto	Relativo (%)
V0401	169 799 170	167 980 331	98,93%	1 818 839	1,07%
V0402	169 799 170	168 434 764	99,20%	1 364 406	0,80%
V4322	169 799 170	169 620 079	99,89%	179 091	0,11%
V4344	169 799 170	169 794 559	100,00%	4 611	0,00%
V0406	169 799 170	167 643 042	98,73%	2 156 128	1,27%
V0407	45 507 516	45 170 336	99,26%	337 180	0,74%
V0408	45 507 516	45 328 064	99,61%	179 452	0,39%
V4093	45 507 516	44 701 199	98,23%	806 317	1,77%

Fonte: IBGE, Censo Demográfico 2000.

Nota: As variáveis V0407, V0408 e V4093 só foram investigadas para a pessoa responsável pelo domicílio ou individual em domicílio coletivo.

Tabela 12.6 - Imputação nos registros de pessoa, segundo o aplicativo - Brasil

Aplicativo	Absoluto	Relativo (%)
Total	**169 799 170**	**100,00**
Sem imputação	159 649 188	94,02
Imputação pelo Aplicativo "Características da Pessoa"	1 405 900	0,83
Imputação pelo Aplicativo "Características da Pessoa Responsável pelo Domicílio ou Individual em Domicílio Coletivo"	6 842 666	4,03
Imputação pelos Aplicativos "Características da Pessoa" e "Características da Pessoa Responsável pelo Domicílio ou Individual em Domicílio Coletivo"	122 251	0,07
Imputação pelo Aplicativo "Características das Demais Pes	1 709 415	1,01
Imputação pelos Aplicativos "Características da Pessoa" e "Características das Demais Pessoas"	69 750	0,04

Fonte: IBGE, Censo Demográfico 2000.

12.3.2.7 Imputação da Variável de Rendimento

Este item descreve o processo de imputação da variável de rendimento dos responsáveis por domicílios do conjunto universo do Censo Demográfico 2000. São apresentadas as motivações que levaram ao desenvolvimento de tal processo, bem como a metodologia desenvolvida para sua aplicação e os resultados obtidos.

Vale lembrar que estamos tratando do Conjunto Universo, cuja formação está descrita no início deste capítulo.

Dentre as perguntas aplicadas, havia a que indagava o valor do "rendimento bruto do mês de julho de 2000" (em R$), proveniente de trabalho e de outras fontes, obtido pela pessoa responsável pelo domicílio ou pelo morador individual em domicílio coletivo (por muitas vezes no texto nos referiremos a essas duas categorias apenas como responsável por domicílio). Da não-resposta a essa questão, podem surgir diversos efeitos sobre análises a serem feitas. Daí, surge a necessidade de que seja feita a imputação de valores de rendimento dos não-respondentes.

Dada a magnitude de uma pesquisa como o Censo Demográfico 2000, fez-se necessária a adoção de uma metodologia de imputação que permitisse o processamento rápido e automatizado da grande massa de dados existente, além de atingir o objetivo principal de corrigir os possíveis efeitos causados pela não-resposta. Com essa finalidade, foi desenvolvida uma metodologia baseada na técnica de Árvores de Regressão (BREIMAN et al., c1984).

Aspectos gerais

A não-resposta é um dos mais comuns erros entre os não-amostrais de uma pesquisa, sendo bem freqüente em países mais desenvolvidos, embora ultimamente venha crescendo em países como o Brasil. Em geral, perguntas sobre rendimentos são mais sujeitas à não-resposta do que as demais perguntas existentes em uma pesquisa como o censo demográfico.

A não-resposta pode ser de dois tipos: completa, quando o total das informações a serem obtidas de uma unidade de pesquisa não é coletada; parcial, quando apenas uma parte das informações não é coletada. No caso do Conjunto Universo do Censo Demográfico 2000, a não-resposta do rendimento do responsável pelo domicílio pode ser vista como parcial, pois as não-respostas que tenham ocorrido nas demais perguntas do questionário foram imputadas com o uso do sistema DIA antes da etapa de imputação de rendimento. A não-utilização do DIA para a imputação de rendimento, justifica-se pelo fato deste ser um sistema adequado à imputação de variáveis categóricas ou numéricas discretas.

Uma pergunta fundamental, cuja resposta justifica em muito a execução de um procedimento de imputação de rendimento, é: quais os efeitos de ignorar-se a não-resposta ao se fazer inferências ou análises sobre o rendimento dos responsáveis por domicílios? No caso da não-resposta não diferencial, isto é, ao acaso, o seu efeito é o de aumento de variância das estimativas obtidas para parâmetros populacionais. No caso da não-resposta diferencial, o seu impacto se dá sob a forma de vício nas estimativas obtidas, com esse vício crescendo com a taxa de não resposta e com a diferença entre respondentes e não-respondentes.

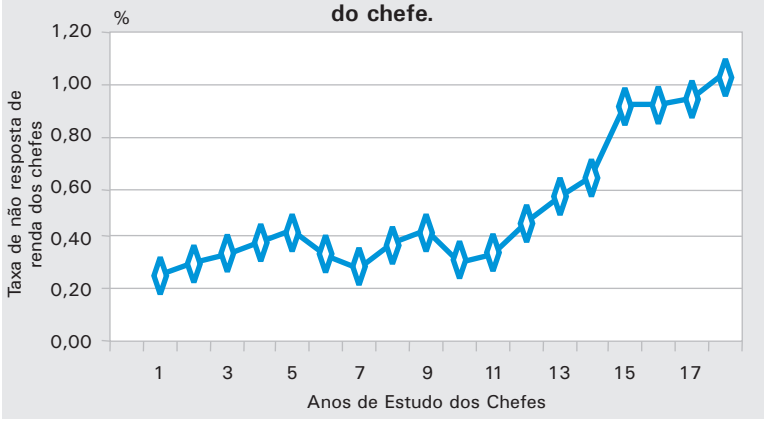
Análises realizadas com dados do Censo Demográfico 1991, mostraram que a não-resposta nos rendimentos dos chefes de domicílio se dava de forma diferencial em relação a algumas das variáveis presentes no questionário. O Gráfico 12.2 mostra que a taxa de não-resposta de rendimento dos chefes de domicílio no Censo de

1991, cresce conforme aumentam os anos de estudo dos chefes. Em geral, foi possível verificar com os dados do Censo de 1991 que as taxas de não-resposta mais elevadas estavam associadas a valores de variáveis que caracterizavam níveis mais altos de rendimento. Donde concluiu-se que quanto maior o rendimento dos chefes de domicílio maior era a probabilidade de não-resposta do rendimento.

São duas as alternativas para lidar com o problema da não-resposta diferencial: uso de estimadores adequados para dados faltantes (LITTLE; RUBIN, 1987) e uso de métodos de imputação (substituição de valores estimados em cada caso individual). No caso de não-resposta parcial de um item/variável, a preferência das agências de estatísticas oficiais é geralmente por métodos baseados em imputação. Tal preferência se deve à maior simplicidade dessa alternativa no processamento posterior dos dados, particularmente quando estes precisam ser publicados na forma de arquivos de microdados com as informações de cada pessoa, individualmente. Albieri (1992) investigou a aplicação de vários métodos para imputação do rendimento na Pesquisa Mensal de Emprego do IBGE.

O método de imputação adotado trabalha com a idéia de estabelecer uma relação entre rendimentos declarados dos responsáveis por domicílios e um grupo de variáveis do conjunto universo cujos valores são conhecidos para todos os domicílios pesquisados, e a partir dessa relação imputar valores de rendimento para os não-respondentes.

Gráfico 12.2 - Taxa de não-resposta de rendimento dos chefes de domicílio no Censo Demográfico 1991, segundo os anos de estudo do chefe.



Dentre as variáveis existentes para o Conjunto Universo algumas, foram selecionadas como possíveis variáveis explicativas do rendimento dos responsáveis por domicílios, sendo este conjunto diverso o suficiente para descrever de forma satisfatória as diferentes relações com o rendimento existentes ao longo do país. A seguir, é apresentado o conjunto dessas variáveis selecionadas para utilização na processo de imputação, com a descrição de cada variável precedida da respectiva nomenclatura adotada para ela durante :

1. IDADANO – Idade em anos do responsável pelo domicílio;

2. ANOSEST – Anos de estudo do responsável pelo domicílio;

3. SEXO – Sexo do responsável pelo domicílio;

4. ESPECIE – Espécie de domicílio: particular permanente; particular improvisado; coletivo;

5. TIPODOM – Tipo de domicílio: casa; apartamento; cômodo;

6. TOTPESDO – Total de moradores no domicílio;

7. EMPREDOM – Total de empregados domésticos residentes no domicílio;

8. CONDDOM – Condição do domicílio: próprio - já pago; próprio – ainda pagando; alugado; cedido por empregador; cedido de outra forma; outra condição;

9. QTDBANH – Quantidade de banheiros existentes no domicílio;

10. SANITAR – Indicadora de existência de sanitário no domicílio com zero banheiro;

11. ABASTEC – Tipo de abastecimento de água: rede geral; poço ou nascente (na propriedade); outra;

12. TIPOCAN – Tipo de canalização de água: canalizada em pelo menos um cômodo; canalizada só na propriedade ou terreno; não canalizada;

13. TIPOESC – Tipo de ligação do escoadouro do banheiro ou sanitário do domicílio: rede geral de esgoto ou pluvial; fossa séptica; fossa rudimentar; vala; rio, lago ou mar; outro escoadouro;

14. LIXO – Tipo de coleta de lixo do domicílio: coletado por serviço de limpeza; colocado em caçamba de serviço de limpeza; queimado (na propriedade); enterrado (na propriedade); jogado em terreno baldio ou logradouro; jogado em rio, lago ou mar; tem outro destino;

15. TIPOSET – Tipo de setor censitário em que se situa o domicílio.

Pode-se notar que parte das variáveis utilizadas refere-se diretamente ao responsável pelo domicílio (de 1 a 3); outra parte é referente a características do domicílio (de 4 a 10); enquanto as demais referem-se a local onde se situa o domicílio (de 11 a 15).

Metodologia

Como já foi dito anteriormente, a metodologia empregada baseia-se na técnica de Árvores de Regressão. A seguir, é dada uma breve idéia a respeito do funcionamento da técnica.

Considere-se a seguinte situação: é preciso predizer o rendimento de uma pessoa. Se for levado em conta apenas o fato de que essa pessoa viva e trabalhe no município do Rio de Janeiro, um preditor bastante "grosseiro" seria o rendimento médio da população desse município. O problema com esse preditor é que ele teria uma precisão muito pequena, ou seja, uma dispersão muito grande, isso se comparado a outros possíveis preditores que levassem em conta variáveis explicativas do rendimento das pessoas residentes no município do Rio de Janeiro.

Portanto, para melhorar a predição pode-se usar outras informações sobre a pessoa. Por exemplo, se for considerado não só o lugar onde ela reside, mas também informações como: idade; nível de instrução; sexo; etc., pode-se assim melhorar a qualidade do preditor. Novamente seria calculada uma média de rendimentos, mas agora sobre uma população bem mais restrita e homogênea.

Esta é a idéia básica das técnicas de regressão: calcular médias em subgrupos (estratos) definidos por variáveis explicativas (covariáveis), obtendo um preditor mais preciso da variável resposta do que o obtido caso não fossem usadas informações sobre essas covariáveis.

Continuando com o exemplo, que perguntas deveríam ser feitas a fim de melhor predizer o rendimento da pessoa? Ou seja, que variáveis explicativas escolher e como fazer a pergunta? Deve ser lembrado que não é permitido perguntar diretamente sobre o rendimento.

Suponha-se que só seja possível perguntar sobre um conjunto dado de variáveis explicativas. Mais ainda, as perguntas são específicas e só podem ser do seguinte tipo:

- no caso de variável numérica: se está abaixo de um valor escolhido (por exemplo: idade ≤ 27 anos; anos de estudo ≤ 7 anos, etc); e
- no caso de variável categórica: se pertence a um subconjunto de categorias.

A resposta a cada pergunta formulada será sim ou não. Mas como escolher as perguntas a fazer? Escolher uma pergunta implica duas escolhas: a da variável explicativa e a de como formular a pergunta a respeito da variável selecionada. Note que no caso de predições de rendimento para indivíduos, as respostas às perguntas definem, passo a passo, estratos cada vez menores de indivíduos. Portanto, para o objetivo é importante escolher as perguntas de modo que esses estratos sejam cada vez mais homogêneos em relação ao rendimento.

Na técnica de Árvores de Regressão, para estabelecer a melhor seqüência de perguntas definidoras dos estratos, parte-se de uma amostra onde sejam conhecidas para cada indivíduo o seu rendimento e os valores das variáveis explicativas. Essa amostra recebe o nome de amostra de treinamento, pois a partir dela é que "se entende" a relação entre o rendimento e as covariáveis adotadas.

Suponha-se, ainda, que seja conhecido um critério numérico D para comparar partições de grupos em dois subgrupos e que o valor de D só dependa dos valores dos rendimentos nos subgrupos definidos. Então, usando a amostra de treinamento, pode-se usar o seguinte procedimento:

- na amostra de treinamento, faz-se todas as perguntas possíveis sobre cada uma das variáveis explicativas, obedecendo as especificações acima definidas para as perguntas. Para cada partição definida por cada pergunta calcula-se o valor do critério. Escolhe-se a pergunta que minimize D. Observe que só é preciso um número finito de perguntas, pois os subgrupos definidos só seriam modificados quando um indivíduo mudasse de grupo, o que ocorreria quando o "ponto de corte" coincidisse com um valor da variável na amostra de treinamento;
- o mesmo procedimento acima seria aplicado em cada um dos dois subgrupos obtidos, sendo sucessivamente geradas partições binárias no grupo de indivíduos da amostra de treinamento. Note que nesse segundo passo, a cada definição de partição é necessário escolher em qual subgrupo particionar. Para isso, bastaria calcular os valores de D referentes às possíveis partições e selecionar aquela para a qual fosse minimizado o valor do critério; e

- por último, é preciso definir um critério de parada para o processo de partições. Possibilidades: limite inferior para o contingente nos subgrupos; ou o fato de que uma nova partição traga "pouca melhora" em termos do critério adotado.

O procedimento acima descrito é uma síntese do funcionamento da técnica de Árvores de Regressão, podendo ser representado por uma árvore binária. Na figura 12.2 é exemplificada uma árvore de regressão, onde para uma amostra de treinamento fictícia (ver tabela 12.7, a seguir) os seus componentes têm o rendimento explicado pelas seguintes covariáveis: sexo; idade; anos de estudo.

Tabela 12.7 - Informações individuais de uma amostra de treinamento fictícia

Sexo	Renda	Idade	Anos de estudo
Masculino	100,00	18	4
Masculino	200,00	20	8
Masculino	200,00	24	6
Masculino	150,00	25	4
Masculino	450,00	30	11
Masculino	300,00	32	8
Masculino	200,00	35	1
Masculino	200,00	46	4
Masculino	1 200,00	63	11
Feminino	200,00	17	8
Feminino	50,00	22	1
Feminino	80,00	25	-
Feminino	150,00	32	4
Feminino	200,00	33	8
Feminino	400,00	35	11
Feminino	300,00	39	8
Feminino	280,00	44	4
Feminino	280,00	49	8
Feminino	120,00	52	5
Feminino	100,00	71	4

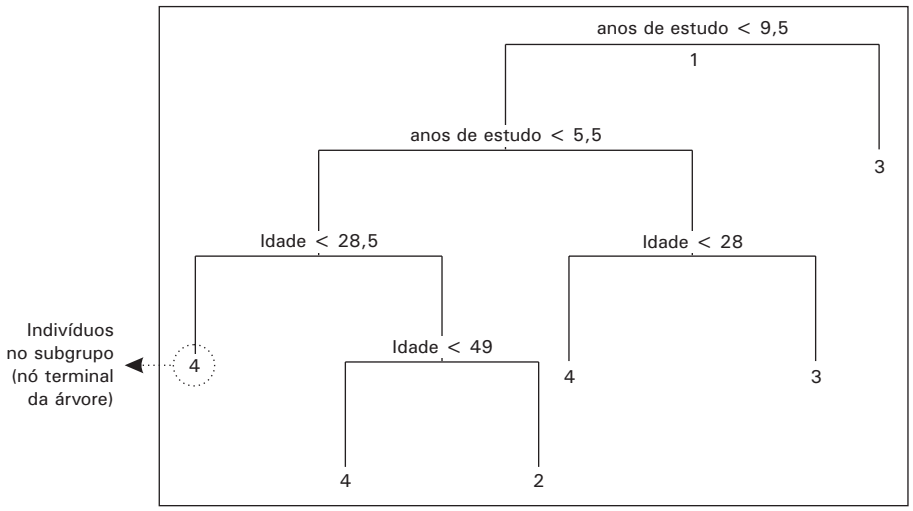
Fonte: IBGE, Diretoria de Pesquisas, Departamento de Metodologia.

Os principais aspectos a comentar sobre a figura 12.2 são:

- a primeira partição dá-se pela separação dos indivíduos com menos de 9,5 anos de estudo dos com mais de 9,5 anos de estudo;
- a segunda partição é feita dentro do grupo de indivíduos com menos de 9,5 anos de estudo, separando-se os que têm menos de 5,5 anos de estudo dos demais indivíduos do estrato;

- as partições são feitas sucessivamente até a condição de parada ser atingida, o que para este exemplo foi estabelecida como sendo a existência de um mínimo de dois indivíduos por nó terminal da árvore; e
- nota-se que a variável "sexo" não é utilizada para a construção da árvore. Isso ilustra o fato de que, na técnica de Árvores de Regressão, não necessariamente todas as covariáveis presentes na amostra de treinamento devam ser utilizadas. A técnica tem, por si só, a capacidade de selecionar as variáveis explicativas mais "poderosas" no sentido de explicar a variável resposta.

Figura 12.2 - Exemplo de Árvore de Regressão



O critério D adotado foi a deviance (soma de desvios quadráticos), que pode ser assim definida:

$$D = \sum_{j} \sum_{i} (y_{ij} - \bar{y}_{j})^2$$

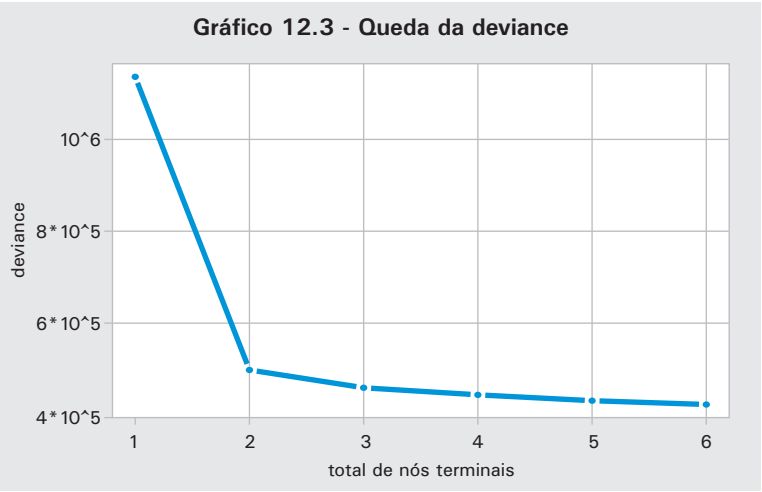
onde:

- y_{ij} é o valor da variável resposta observado para o indivíduo i pertencente ao estrato (nó terminal) j;
- $\bar{y}_{j}$ é a média da variável resposta no estrato j.

No Gráfico 12.3 é apresentado o comportamento da *deviance* com o aumento do número de partições na árvore de regressão construída para a amostra de treinamento adotada como exemplo. Conforme pode-se notar, há uma queda acentuada no valor de D ao particionar-se o grupo inicial em dois estratos,

com as partições seguintes trazendo "melhoras" cada vez menores na obtenção de estratos mais homogêneos em relação ao rendimento. Esses gráficos podem ser utilizados na escolha do número de nós terminais a ser adotado.

Gráfico 12.3 - Queda da deviance



Conhecidas as perguntas a fazer, como então predizer um rendimento desconhecido, isto é, como imputar os rendimentos dos não respondentes?

Dado que são conhecidos os valores das variáveis explicativas para os que não declararam seus valores de rendimentos, uma solução seria localizar essas pessoas nos nós terminais e, em seguida, imputar o rendimento de cada uma pelo rendimento médio em seu respectivo "nó". Porém, tal solução possui o inconveniente de não manter a distribuição original da variável resposta em cada estrato, visto que seria sempre imputado o rendimento médio no estrato. Por esse motivo adota-se o procedimento denominado hot-deck aleatório, onde para cada indivíduo não-respondente seleciona-se aleatoriamente um "doador de rendimento" dentro de seu nó terminal, e imputa-se seu rendimento pelo rendimento do doador.

Aplicação e conclusões

Para o processo de imputação de rendimento dos responsáveis, utilizaram-se os mesmos lotes de registros definidos para a crítica e imputação de dados do Censo Demográfico 2000, realizadas com o uso do sistema DIA. Esses lotes correspondem a uma partição do conjunto universo de respondentes, obedecendo os domínios das Unidades da Federação (UF), isto é, um mesmo lote não contém registros de mais de uma UF. Para o processo de imputação de rendimento foram utilizados somente os registros de responsáveis por domicílio ou morador individual em domicílio coletivo, além de terem sido excluídos de cada lote os registros cujos rendimentos estavam fora das cercas construídas para detectar outliers (valores atípicos). Nos 526 lotes utilizados na imputação de rendimento havia o total de 45.280.240 registros, com o menor lote possuindo 10.094 registros e o maior 103.248 registros. A distribuição da quantidade de lotes por Unidades da Federação pode ser vista na tabela 12.8.

Tabela 12.8 - Número de lotes para imputação, segundo as Unidades da Federação

Unidades da Federação	Lotes	Unidades da Federação	Lotes
Brasil	**526**	Alagoas	9
Rondônia	5	Sergipe	6
Acre	3	Bahia	36
Amazonas	8	Minas Gerais	54
Roraima	2	Espírito Santo	10
Pará	16	Rio de Janeiro	48
Amapá	2	São Paulo	117
Tocantins	4	Paraná	31
Maranhão	15	Santa Catarina	18
Piauí	8	Rio Grande do Sul	34
Ceará	21	Mato Grosso do Sul	7
Rio Grande do Norte	8	Mato Grosso	8
Paraíba	10	Goiás	16
Pernambuco	23	Distrito Federal	7

Fonte: IBGE, Censo Demográfico 2000.

Para cada um dos 526 lotes foi aplicado o procedimento de imputação baseado em árvores de regressão, descrito acima. Esse procedimento foi implementado com o software *S-Plus* e executado em ambiente operacional Windows 98. Como os lotes de registros residiam em arquivos do ambiente operacional OS/390 (mainframe IBM), foi desenvolvida, utilizando o software SAS e seus recursos para a conexão desses dois ambientes operacionais, uma rotina computacional para automatizar todo o processo de produção dessa imputação, constituído das seguintes etapas: a) preparação do arquivo de entrada para o *S-Plus*; b) ativação do *S-Plus* para a imputação propriamente dita; c) transferência dos resultados para o ambiente OS/390 e d) atualização dos registros nos lotes originais com os valores imputados.

A regra de parada na construção das árvores de regressão baseou-se no número máximo de nós terminais permitido nas árvores e no contingente populacional mínimo exigido em cada nó terminal. Visto que seria impraticável a análise dos gráficos de queda da *deviance* para cada uma das 526 árvores, uma das regras de parada adotada foi a da partição de cada lote em no máximo 25 estratos. Análises preliminares com dados do Censo de 1991 indicaram ser este um número de nós terminais para o qual, em geral, não haveria "ganhos consideráveis" com novas partições. Quanto aos contigentes populacionais de cada estrato, foi estipulado que estes deveriam ser de, no mínimo, 100 pessoas.

Tabela 12.9 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação de rendimento dos responsáveis por domicílios

Taxas de não-resposta nos lotes de imputação (%)					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
0,54	1,17	1,60	1,78	2,15	6,24

Fonte: IBGE, Censo Demográfico 2000.

Além das estatísticas referentes às taxas de não-resposta nos lotes, apresentadas na Tabela 12.9, há que se dizer ainda que a taxa geral de não-resposta de rendimento entre os responsáveis por domicílios foi de 1,75%.

O Gráfico 12.4 apresenta os resultados da participação de cada variável explicativa nas árvores de regressão construídas, com os dados utilizados na construção do gráfico podendo ser vistos na Tabela 12.10. Essa participação é vista sob o ponto de vista da ordem em que a variável gerou uma partição na árvore pela primeira vez. Por exemplo, as variáveis QTDBANH e ANOSEST geraram a 1ª ou a 2ª partição em 94,7% e 87,6% das árvores, respectivamente.

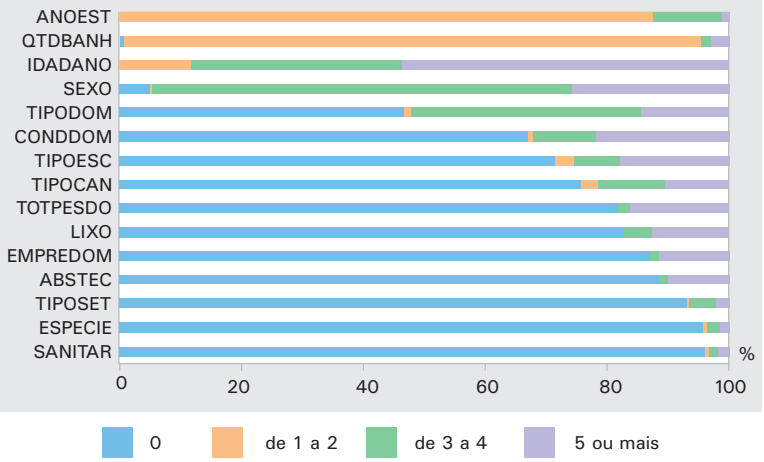
Tabela 12.10 - Resumo da participação das variáveis nas árvores de regressão

Variável	Ordem de entrada na variável árvore (%)				
	1	2	3 ou 4	5 ou mais	Não entrou
QTDBANH	53,80	40,87	1,71	2,85	0,76
ANOSEST	39,92	47,72	11,41	0,95	0,00
IDADANO	3,99	3,23	34,79	53,42	4,56
TIPOESC	1,33	2,09	7,41	17,87	71,29
TIPOCAN	0,57	2,28	11,03	10,46	75,67
TIPODOM	0,19	0,76	37,64	14,45	46,96
SANITAR	0,19	0,76	1,33	1,71	96,01
ESPECIE	0,00	0,76	2,09	1,52	95,63
SEXO	0,00	0,57	68,63	25,86	4,94
CONDDOM	0,00	0,38	10,46	21,86	67,30
TIPOSET	0,00	0,38	4,18	2,09	93,35
TOTPESDO	0,00	0,19	1,90	16,16	81,75
LIXO	0,00	0,00	4,75	12,36	82,89
ABASTEC	0,00	0,00	1,52	10,08	88,40
EMPREDOM	0,00	0,00	1,14	11,60	87,26

Fonte: IBGE, Diretoria de Pesquisas, Departamento de Metodologia.

Nota: As variáveis V0407, V0408 e V4093 só foram investigadas para a pessoa responsável pelo domicílio ou individual em domicílio coletivo.

Gráfico 12.4 - Ordem de entrada das variáveis



Ainda a respeito das variáveis ANOSEST e QTDBANH, nota-se que a primeira foi selecionada na construção das árvores de todos os lotes, enquanto a segunda não entrou em menos de 1% das árvores construídas. As variáveis IDADANO e SEXO entraram em mais de 95% das árvores, em geral foram a 3ª ou 4ª a ser selecionada, porém a variável SEXO foi com maior freqüência mais importante do que IDADANO. Por outro lado, as variáveis TOTPESDO; LIXO; EMPREDOM; ABASTEC; TIPOSET; ESPECIE e SANITAR, não foram incluídas nas árvores de mais de 80% dos lotes.

Conforme pôde-se notar pelos resultados apresentados, as variáveis diretamente relacionadas à pessoa do responsável pelo domicílio foram as que apresentaram maior poder de predição do rendimento desta pessoa, excetuando-se aí, é claro, a variável QTDBANH.

Como forma de avaliar a qualidade do resultado da imputação em cada lote foi aplicado o teste estatístico de Kolmogorov-Smirnov (LEHMANN, c1975). Este teste visa verificar se duas amostras de dados provêm de uma mesma população. No nosso caso, como se tinha o objetivo de não alterar a distribuição do rendimento em cada subgrupo formado, então o teste foi aplicado para se comparar os vetores de rendimentos em cada subgrupo antes e depois da execução do procedimento de imputação. Cada lote só teve seu respectivo processo de imputação aprovado se o teste de Kolmogorov-Smirnov indicasse que os rendimentos antes e depois da imputação, em cada estrato formado, apresentavam a "mesma distribuição".

Outro ponto a ser colocado diz respeito ao tratamento dado à imputação de rendimentos nulos. Visto que seria possível a categorização da variável de rendimentos em nulos e não-nulos, haveria a possibilidade do sistema DIA ser utilizado para a imputação de valores nulos de rendimentos. Porém, essa solução não foi adotada porque avaliou-se que, no caso do conjunto universo, o conjunto de variáveis explicativas disponíveis não possuía boa capacidade de predição da variável de rendimento dicotomizada (sem rendimento ou com rendimento positivo). Posto isso, optou-se por realizar a imputação de rendimentos nulos dentro do procedimento de imputação estabelecido.

Finalizando, são apresentadas a seguir algumas estatísticas referentes às distribuições nos lotes do percentual de responsáveis contido em cada estrato (tabela 12.11) e às distribuições das taxas de imputação nos estratos obtidos nas árvores de regressão construídas (tabela 12.12). Um aspecto a comentar no Tabela 12.12 diz respeito ao fato de ter sido observado extrato com 100% de não-resposta, o que acarretou na inexistência de "doadores de rendimento". Tal problema foi solucionado com a seleção de doador no grupo de responsáveis do qual foi gerado o estrato.

Tabela 12.11 - Estatísticas descritivas dos percentuais de responsáveis nos estratos em relação ao tamanho dos lotes de imputação de rendimento dos responsáveis por domicílios

Percentual de responsáveis no estrato em relação ao tamanho do lote (%)					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
0,001	0,580	1,530	3,990	4,220	61,120

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.12 - Estatísticas descritivas das taxas de não-resposta nos estratos das árvores de regressão dos lotes de imputação de rendimento dos responsáveis por domicílios

Percentual de responsáveis no estrato em relação ao tamanho do lote					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
0,00	0,67	1,15	1,54	1,89	100,00

Fonte: IBGE, Censo Demográfico 2000.

12.4 Resultados preliminares da amostra

Os resultados preliminares da amostra do Censo Demográfico 2000 foram divulgados em maio de 2002, correspondendo à Tabulação Avançada e à Fecundidade e Mortalidade Infantil.

Diferentemente da Fecundidade e Mortalidade Infantil, onde foram utilizados para a obtenção das estimativas os dados de toda a amostra do censo, na Tabulação Avançada recorreu-se à elaboração de uma subamostra.

Com exceção dos dados referentes à estrutura familiar e sexo, os demais, utilizados para a obtenção desses resultados, ainda não tinham sido submetidos ao processo de crítica e imputação.

12.4.1 Tabulação Avançada

O objetivo da Tabulação Avançada foi fornecer, antecipadamente à divulgação dos resultados da amostra do censo, estimativas para um conjunto de tabelas com variáveis do questionário da amostra para o total do País, Grandes Regiões e Unidades da Federação. Para isso, foi retirada uma subamostra da amostra do Censo Demográfico 2000, constituída por uma amostra de setores censitários, com os respectivos domicílios e pessoas neles pesquisados, que preencheram o questionário da amostra, cujas informações ainda não haviam sido submetidas a todos os processos de crítica eletrônica.

Portanto, a Tabulação Avançada contém resultados preliminares da amostra do Censo Demográfico 2000, que estavam sujeitos a alterações quando da versão definitiva.

Plano amostral da Tabulação Avançada

O plano amostral da Tabulação Avançada consistiu em uma amostragem estratificada simples de setores censitários em cada unidade da federação (cada um dos 26 estados e o Distrito Federal). Em cada Unidade da Federação foram definidos até três estratos de acordo com a situação e o tipo do setor, a saber: setores rurais, setores urbanos não-especiais, e setores urbanos de aglomerados subnormais, quando existentes. Dentro de cada estrato, os setores foram selecionados por amostragem aleatória simples. A seleção dos setores foi feita utilizando-se o algoritmo sugerido por Fan, Muller e Rezucha em 1962, que está descrito em Särndal, Swensson e Wretman (c1992, p. 66).

O tamanho da amostra de setores em cada Unidade da Federação (UF) foi definido com base em estudos descritos em Albieri, Martelotte e Duarte (2000) e em Albieri (1999b). Nestes estudos ficou definido que, para a estimação de características para Unidades da Federação (UF) com precisão razoável, o número de

setores a ser utilizado seria o equivalente ao da amostra da Pesquisa Nacional por Amostra de Domicílios – PNAD, e que o tamanho mínimo da amostra de uma UF seria de 50 setores. A alocação da amostra de setores nos três estratos foi proporcional ao número de setores existentes em cada estrato na população, sendo considerado um mínimo de dois setores, para permitir a estimação do erro amostral.

A Tabela 12.13 a seguir mostra o número de setores do Censo Demográfico e da Tabulação Avançada por Unidade da Federação.

Tabela 12.13 - Número de setores do Censo Demográfico 2000 e da Tabulação Avançada, segundo as Unidades da Federação

Unidades da Federação	Setores Do Censo 2000 (1)	Setores da Tabulação Avançada (1)				Fração amostral de setores (%)
		Total	Estratos			
			Rural	Urbano	Aglomerado subnormal	
Brasil	**214 319**	**4 359**	**1 312**	**2 897**	**150**	**2**
Rondônia	1 888	57	28	29	-	3
Acre (1)	552	51	21	30	-	9,2
Amazonas	3 236	104	33	64	7	3,2
Roraima	476	51	20	31	-	10,7
Pará	6 083	121	50	60	11	2
Amapá	452	51	13	35	3	11,3
Tocantins	1 364	58	26	32	-	4,3
Maranhão	6 398	114	65	46	3	1,8
Piauí	3 708	97	51	42	4	2,6
Ceará	7 947	193	65	117	11	2,4
Rio Grande do Norte	2 633	91	35	53	3	3,5
Paraíba	4 162	114	48	63	3	2,7
Pernambuco	8 541	233	85	142	6	2,7
Alagoas	2 600	94	38	53	3	3,6
Sergipe	2 220	87	33	51	3	3,9
Bahia	15 315	277	118	154	5	1,8
Minas Gerais	22 469	352	100	242	10	1,6
Espírito Santo	3 196	109	31	75	3	3,4
Rio de Janeiro	20 607	293	19	247	27	1,4
São Paulo	49 303	423	47	354	22	0,9
Paraná	13 005	212	68	139	5	1,6
Santa Catarina	6 794	173	52	118	3	2,5
Rio Grande do Sul	16 837	316	94	216	6	1,9
Mato Grosso do Sul	2 710	120	37	80	3	4,4
Mato Grosso	3 309	136	51	82	3	4,1
Goiás	5 960	263	75	185	3	4,4
Distrito Federal	2 554	169	9	157	3	6,6

Fonte: IBGE, Censo Demográfico 2000.

(1) Setores com pelo menos uma pessoa recenseada no Censo Demográfico 2000.

Em cada setor todos os domicílios e pessoas nele pesquisados, através do Questionário da Amostra, foram processados e incluídos na amostra da Tabulação Avançada.

Nas Tabelas 12.14 e 12.15 a seguir, apresenta-se o número de domicílios e de pessoas do Censo Demográfico e da Tabulação Avançada por Unidade da Federação.

Tabela 12.14 - Número de domicílios do Censo Demográfico 2000 e da Tabulação Avançada, segundo as Unidades da Federação

Unidades da Federação	Domicílios do Censo 2000 (1)	Domicílios da Tabulação Avançada (1)				Fração amostral de domicílios (%)
		Total	Estratos			
			Rural	Urbano	Aglomerado subnormal	
Brasil	**45 507 516**	**108 989**	**25 238**	**80 339**	**3 412**	**0,24**
Rondônia	354 391	1 149	448	701	-	0,32
Acre	131 580	1 539	575	964	-	1,17
Amazonas	580 900	2 092	556	1 383	153	0,36
Roraima	76 681	1 100	461	639	-	1,43
Pará	1 332 248	2 810	1 011	1 547	252	0,21
Amapá	100 765	1 203	137	916	150	1,19
Tocantins	285 701	1 710	377	1 333	-	0,60
Maranhão	1 246 715	2 463	960	1 430	73	0,20
Piauí	665 808	2 559	987	1 466	106	0,38
Ceará	1 773 393	4 866	1 510	3 127	229	0,27
Rio Grande do Norte	678 652	3 029	922	2 031	76	0,45
Paraíba	857 989	3 052	1 000	1 974	78	0,36
Pernambuco	1 994 041	5 946	1 496	4 268	182	0,30
Alagoas	658 873	2 805	870	1 902	33	0,43
Sergipe	442 256	2 138	675	1 394	69	0,48
Bahia	3 214 292	6 902	2 358	4 404	140	0,21
Minas Gerais	4 837 296	9 881	2 055	7 592	234	0,20
Espírito Santo	851 014	3 342	890	2 356	96	0,39
Rio de Janeiro	4 315 737	5 993	300	5 210	483	0,14
São Paulo	10 564 745	9 660	968	8 274	418	0,09
Paraná	2 709 523	5 516	1 218	4 221	77	0,20
Santa Catarina	1 518 651	5 028	1 270	3 714	44	0,33
Rio Grande do Sul	3 091 643	6 751	1 502	5 132	117	0,22
Mato Grosso do Sul	577 362	3 056	612	2 397	47	0,53
Mato Grosso	669 676	3 472	746	2 517	209	0,52
Goiás	1 420 822	7 286	1 186	6 010	90	0,51
Distrito Federal	556 762	3 641	148	3 437	56	0,65

Fonte: IBGE, Censo Demográfico 2000.

(1) Número de domicílios com pelo menos uma pessoa recenseada no Censo Demográfico 2000.

Tabela 12.15 - Número de pessoas do Censo Demográfico 2000 e da Tabulação Avançada, segundo as Unidades da Federação

(continua)

Unidades da Federação	Pessoas do Censo 2000	Pessoas da Tabulação Avançada				Fração amostral de pessoas (%)
		Total	Estratos			
			Rural	Urbano	Aglomerado subnormal	
Brasil	**169 799 170**	**423 049**	**108 938**	**300 268**	**13 843**	**0,25**
Rondônia	1 379 787	4 450	1 800	2 650	-	0,32
Acre	557 526	6 295	2 478	3 817	-	1,13
Amazonas	2 812 557	9 940	3 087	6 164	689	0,35
Roraima	324 397	4 534	1 908	2 626	-	1,40
Pará	6 192 307	13 395	5 141	7 185	1 069	0,22
Amapá	477 032	5 802	673	4 375	754	1,22
Tocantins	1 157 098	6 717	1 474	5 243	-	0,58
Maranhão	5 651 475	11 644	4 642	6 672	330	0,21
Piauí	2 843 278	10 939	4 339	6 141	459	0,38
Ceará	7 430 661	20 843	7 145	12 763	935	0,28
Rio Grande do Norte	2 776 782	12 711	4 252	8 127	332	0,46

Tabela 12.15 - Número de pessoas do Censo Demográfico 2000 e da Tabulação Avançada, segundo as Unidades da Federação

(conclusão)

Unidades da Federação	Pessoas do Censo 2000	Pessoas da Tabulação Avançada				Fração amostral de pessoas (%)
		Total	Estratos			
			Rural	Urbano	Aglomerado subnormal	
Paraíba	3 443 825	12 676	4 598	7 776	302	0,37
Pernambuco	7 918 344	24 575	7 283	16 508	784	0,31
Alagoas	2 822 621	12 746	4 528	8 071	147	0,45
Sergipe	1 784 475	9 231	3 227	5 697	307	0,52
Bahia	13 070 250	29 078	10 821	17 666	591	0,22
Minas Gerais	17 891 494	37 835	8 887	28 026	922	0,21
Espírito Santo	3 097 232	12 551	3 700	8 467	384	0,41
Rio de Janeiro	14 391 282	20 633	1 196	17 686	1 751	0,14
São Paulo	37 032 403	33 718	3 191	28 793	1 734	0,09
Paraná	9 563 458	19 439	4 451	14 701	287	0,20
Santa Catarina	5 356 360	18 151	5 169	12 763	219	0,34
Rio Grande do Sul	10 187 798	22 330	5 273	16 635	422	0,22
Mato Grosso do Sul	2 078 001	10 990	2 237	8 543	210	0,53
Mato Grosso	2 504 353	12 871	2 683	9 506	682	0,51
Goiás	5 003 228	25 749	4 243	21 203	303	0,51
Distrito Federal	2 051 146	13 206	512	12 464	230	0,64

Fonte: IBGE, Censo Demográfico 2000.

Expansão da amostra e cálculo dos pesos amostrais da Tabulação Avançada

Numa pesquisa por amostra o que se busca são estimativas dos valores de determinados parâmetros populacionais de interesse, por meio da investigação de apenas uma parte das unidades dessa população. As tabelas divulgadas na Tabulação Avançada são formadas pelo cruzamento de variáveis relativas a pessoas, domicílios e famílias, classificadas segundo categorias indicadoras de faixa etária, sexo, religião, rendimento, localização, etc. Cada uma das células das tabelas teve seu valor estimado, juntamente com uma medida de precisão de estimativa dada pelo coeficiente de variação.

O processo de cálculo das estimativas é também conhecido como expansão da amostra e depende da determinação dos pesos associados a cada unidade amostrada. Os pesos usuais dados pelo plano amostral, definidos como o inverso das probabilidades de inclusão de cada unidade da população na amostra, são os pesos mais simples, que podem ser utilizados para a expansão dos resultados de uma pesquisa por amostra. No caso do desenho amostral da Tabulação Avançada, estes pesos podem ser escritos como:

$$d_{hij} = \frac{1}{\pi_{hij}} = \frac{M_h}{m_h}\frac{N_{hi}}{n_{hi}} \quad ,$$

onde:

d_{hij} é o peso do domicílio j, do setor i, do estrato h na amostra;

π_{hij} é a probabilidade de inclusão do domicílio j, do setor i, do estrato h na amostra;

M_h e m_h são, respectivamente, o número de setores no estrato h no Censo Demográfico 2000 e na amostra da Tabulação Avançada;

N_{hi} e n_{hi} são, respectivamente, o número de domicílios no setor i do estrato h no Censo Demográfico 2000 e na amostra.

Em cada unidade da federação existem, no máximo, três estratos[3]. Esses pesos utilizam apenas informações do plano amostral.

Na Tabulação Avançada os pesos utilizados foram os pesos definidos como descrito, calibrados de modo a se ajustarem às informações auxiliares fornecidas pela investigação censitária, realizada pela aplicação das perguntas comuns aos dois tipos de questionário, que são feitas para todos os domicílios da população. Foram utilizados como variáveis de calibração os totais de homens e mulheres e o total de domicílios por estrato, já que tais informações estavam disponíveis para todos os domicílios da população e, conseqüentemente, da amostra.

Os pesos amostrais indicam quantas unidades da população cada unidade da amostra representa. A calibração dos pesos foi feita de maneira que os pesos calibrados ficassem o mais próximo possível dos pesos dados pelo desenho amostral, impondo-se como limite mínimo o valor 1, ou seja, nestes casos cada unidade representa apenas ela mesma.

A razão para se optar pelo uso de pesos calibrados vem do fato de estes produzirem estimativas mais precisas e mais consistentes com os valores conhecidos do Censo Demográfico 2000. Os totais, para os estratos, das variáveis usadas na calibração, quando estimados pelos pesos calibrados, coincidem com os valores conhecidos para toda a população.

O método de calibração usado é um processo com restrições não-lineares, nas variáveis de calibração com limites nos valores dos pesos calibrados. Detalhes podem ser vistos na publicação *Generalised estimation system* (1998), do Statistics Canada.

Os pesos foram calculados no nível do domicílio, sendo atribuídos também a cada um de seus moradores.

Cálculo das estimativas pontuais da Tabulação Avançada

As estimativas de totais para as células das tabelas da Tabulação Avançada foram calculadas por:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} d^{*}_{hij}\ y_{hij}$$

onde:

$d^{*}_{hij} = peso\ \ calibrado\ \ para\ a\ \ unidade\ \ j\ \ do\ setor\ \ i\ do\ estrato\ \ h$;

$H = número\ \ de\ \ estratos$; e

a) Para variáveis categóricas:

$$y_{hij} = \begin{cases} 1,\ se\ a\ unidade\ \ j\ do\ setor\ i\ do\ estrato\ h\ \ pertence\ à\ categoria\ em\ questão \\ 0,\ se\ a\ unidade\ \ j\ do\ setor\ i\ do\ estrato\ h\ não\ \ pertence\ à\ categoria\ em\ questão \end{cases}$$

[3] Em algumas Unidades da Federação, como Rondônia, Acre, Roraima e Tocantins, não havia setor, com pelo menos uma pessoa recenseada com o Questionário da Amostra, no estrato de setores urbanos de aglomerados subnormais..

b) Para variáveis contínuas:

$y_{hij} = valor\ da\ variável\ de\ estudo\ na\ unidade\ j\ do\ setor\ i\ no\ estrato\ h.$

Cálculo das estimativas de precisão das estimativas pontuais da Tabulação Avançada

Para avaliar a precisão das estimativas de totais, foram calculados os respectivos Coeficientes de Variação (CVs), definidos como:

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

onde $v(\hat{Y})$ é a estimativa da variância da estimativa de total $\hat{Y}$, que foi calculada conforme fórmulas definidas no capítulo 8 de Särndal, Swensson e Wretman (1992).

A partir do coeficiente de variação, pode-se construir um intervalo de confiança, para um total em questão, dado pela expressão:

$$\hat{Y} - z_\alpha \times \hat{Y} \times cv(\hat{Y}) \leq Y \leq \hat{Y} + z_\alpha \times \hat{Y} \times cv(\hat{Y});$$

onde:

Y é o valor verdadeiro do total em questão;

$\hat{Y}$ é a estimativa amostral do total;

$cv(\hat{Y})$ é o coeficiente de variação da estimativa;

z_α é o valor da ordenada da distribuição normal padrão para um nível á de significância.

Como nas tabelas divulgadas na Tabulação Avançada existe um número muito grande de estimativas de total, optou-se por não colocar os valores estimados dos CVs, e sim uma letra ao lado de cada estimativa pontual, correspondente a uma faixa de valores para o coeficiente de variação. As faixas utilizadas foram as sugeridas em Albieri (1999a), apresentadas no Quadro 12.1, a seguir.

Quadro 12.1 - Indicadores de faixas de coeficiente de variação utilizados nas tabelas da Tabulação Avançada do Censo Demográfico 2000

Indicador	Faixas de CV (%)
z	zero
a	de 0,0 até 0,5
b	mais de 0,5 até 1,0
c	mais de 1,0 até 2,5
d	mais de 2,5 até 5,0
e	mais de 5,0 até 7,5
f	mais de 7,5 até 10,0
g	mais de 10,0 até 15,0
h	mais de 15,0 até 25,0
i	mais de 25,0 até 35,0
j	mais de 35,0 até 50,0
k	mais de 50,0

A partir das letras indicativas dos valores dos coeficientes de variação, é possível calcular um intervalo de confiança aproximado para o total desejado, usando os limites das faixas de valores. Exemplificando, pode-se construir um intervalo de confiança de 95% para um dado total, cujo coeficiente de variação da estimativa esteja na faixa b usando o limite superior dessa faixa, ou seja:

$$\hat{Y} - 1{,}96 \times \hat{Y} \times 0{,}01 \leq Y \leq \hat{Y} + 1{,}96 \times \hat{Y} \times 0{,}01$$

Tratamento dos dados da Tabulação Avançada

Convém ressaltar que uma vez selecionados os setores que pertencerem à amostra da tabulação, foram definidos procedimentos para que a sua apuração fosse realizada de forma prioritária, ou seja, esses setores tiveram prioridade nas primeiras etapas de apuração dos questionários, desde a remessa para os centros de captura, passando por todos os procedimentos relacionados com a própria captura, a saber, leitura ótica, verificação e crítica de quantidades. Todos os demais procedimentos de apuração dos questionários pertencentes aos setores da amostra da tabulação avançada foram realizados de forma separada e independente da apuração para a obtenção dos resultados definitivos. Para tanto, a base de dados desses questionários foi duplicada e neles foram aplicados apenas os procedimentos de validação para a geração dos resultados preliminares divulgados. Todas as etapas de apuração, após a captura, não foram incorporadas ao processo definitivo.

Excetuando as informações referentes à estrutura familiar e gênero, os dados divulgados para os demais temas da publicação não passaram pelo processo de crítica eletrônica, adotado na divulgação de resultados definitivos, e que tem por finalidade eliminar eventuais inconsistências entre as informações dos diversos quesitos do questionário, que podem ter origem na coleta de dados ou na fase de reconhecimento de marcas e caracteres.

Portanto, as informações que apresentaram alguma inconsistência, do tipo quesito omitido quando deveria estar preenchido, erro de seqüência no preenchimento do questionário ou impossibilidade de alocação de um valor em alguma célula de tabela, foram incluídas apenas nas colunas ou linhas de total. Assim, os valores dos totais de linhas e colunas não necessariamente coincidem com as somas dos valores das parcelas correspondentes.

Conteúdo da publicação da Tabulação Avançada

A publicação *Tabulação avançada do censo demográfico 2000: resultados preliminares da amostra* (2002) contém, para o total do País e para as grandes regiões e unidades da federação, um total de 38 tabelas com características dos domicílios e das pessoas, captadas por meio do questionário da amostra sobre os seguintes temas: características gerais da população, educação, migração, nupcialidade, trabalho, famílias e domicílios.

A publicação incluiu, além das notas metodológicas, comentários dos resultados sobre as características gerais: cor ou raça, religião e deficiência; características da educação, migração, nupcialidade, famílias e domicílios. Incluiu também um CD-ROM com as 38 tabelas divulgadas.

12.4.2 Fecundidade e mortalidade Infantil

No tocante aos temas Fecundidade e Mortalidade Infantil, a divulgação dos resultados preliminares da amostra teve um tratamento diferenciado dos demais temas que compuseram a *Tabulação avançada do censo demográfico 2000: resultados preliminares da amostra*. Em vez de se utilizar uma subamostra de 0,24% dos domicílios, o IBGE optou por uma apresentação em separado da Tabulação Avançada utilizando-se, para fins de geração das correspondentes estimativas, toda a amostra do censo. Esta foi uma decisão baseada nas especificidades das variáveis envolvidas que guardam certo grau de complexidade na coleta das informações e, sobretudo, porque os indicadores derivados são, pela sua natureza, bastante sensíveis a flutuações amostrais.

Além disso, a elaboração das estimativas dos parâmetros da Fecundidade e da Mortalidade Infantil requer a aplicação de metodologias específicas. Para maiores esclarecimentos, podem ser consultados: Brass (1971), Brass et al. (1968, 1975), Camisa (1975), Coale e Trussell (1974), Oliveira (1991) e Trussell (1975). Cabe ressaltar, também, que nesta etapa de divulgação dos resultados preliminares da amostra do Censo Demográfico 2000 fez-se uso das informações obtidas após o processo de captura de dados, portanto, ainda não submetidas à crítica quanto à sua consistência.

12.4.2.1 Cálculo das estimativas

Para o cálculo das estimativas da Fecundidade e da Mortalidade Infantil foram introduzidos alguns filtros que tinham como objetivo fornecer uma visão, ainda que muito preliminar, dos dados com algum tipo de crítica. Isso foi levado a efeito por tema, considerando-se os registros com erro de preenchimento e de não resposta referentes às mulheres, para as seguintes variáveis:

- Fecundidade: filhos tidos nascidos vivos, filhos sobreviventes ou filhos tidos nos últimos doze meses anteriores ao censo demográfico 2000, e

- Mortalidade Infantil: filhos tidos nascidos vivos ou filhos sobreviventes.

Uma descrição detalhada da implementação desses filtros pode ser encontrada no anexo do CD-ROM que acompanha a metodologia.

São descritos a seguir os pressupostos relativos às técnicas utilizadas para o cálculo das taxas de fecundidade e de mortalidade infantil.

a) Fecundidade

No tocante à Fecundidade, a técnica empregada é a metodologia proposta por Brass (1971), que consiste em combinar três informações:

- mulheres em idade fértil (de 15 a 49 anos de idade), classificadas por grupos qüinqüenais de idade [M (i); i = 1,2,3,....,7],

onde $i = 1 \rightarrow$ 15 a 19 anos

$i = 2 \rightarrow$ 20 a 24 anos

$i = 3 \rightarrow$ 25 a 29 anos

$i = 4 \rightarrow$ 30 a 34 anos

$i = 5 \rightarrow$ 35 a 39 anos

$i = 6 \rightarrow$ 40 a 44 anos

$i = 7 \rightarrow$ 45 a 49 anos

- filhos tidos nascidos vivos, declarados pelas mulheres, classificados segundo os mesmos grupos qüinqüenais de idade das mulheres [FNV (i)], e

- filhos tidos nascidos vivos nos 12 meses anteriores à data de referência do censo, declarados pelas mulheres, tabulados segundo os grupos qüinqüenais de idade das mulheres de 15 a 49 anos [FUA (i)], sendo esta uma informação derivada do quesito "data de nascimento do último filho tido nascido vivo".

Tais informações foram obtidas das mulheres de 10 anos ou mais de idade, mas a técnica foi concebida para ser aplicada ao contingente feminino de 15 a 49 anos.

Inicialmente, são calculadas:

- as parturições médias, P(i), que: representam a fecundidade retrospectiva acumulada das mulheres, obtidas mediante a seguinte relação:

P (i) = FNV (i) / M (i), i = 1, 2, ... , 7, e

- as taxas específicas de fecundidade por grupos de idade das mulheres, que expressam a fecundidade atual das mulheres, através do seguinte cálculo:

f (i) = FUA (i) / M (i), i = 1, 2, ... ,7.

Em síntese, a aplicação da técnica consiste na comparação das parturições [P(i)] com a fecundidade atual acumulada [F(i)], obtida dos nascimentos dos últimos 12 meses. Esta comparação é feita mediante a análise da série P (i) / F (i), da qual será extraído o fator de correção das taxas de fecundidade atuais f (i).

Isso porque, para a aplicação da técnica, também conhecida como técnica da razão P/F, o requerimento básico é a aceitação de duas hipóteses relacionadas com a informação de referência. A primeira diz que a estrutura por grupos de idade das taxas de fecundidade atual é aceitável, ainda que não seja o nível estimado a partir dela, isto é, a Taxa de Fecundidade Total. A segunda hipótese faz referência à melhor qualidade da informação sobre a fecundidade retrospectiva, associada aos grupos 20 a 24 e 25 a 29 anos de idade. A depender do caso específico, esses grupos etários são empregados como bons indicadores do nível da fecundidade.

A técnica de Brass traz implícitos, porém, dois pressupostos metodológicos que relacionam f(i) e P (i):

- que as mulheres sobreviventes à data do censo são regidas pela mesma lei de fecundidade das mulheres que já faleceram, e

- que a fecundidade tenha permanecido constante ao longo do tempo. As mulheres chegam a uma parturição média correspondente ao grupo 45 a 49 anos de idade expostas aos mesmos riscos de fecundidade prevalecentes nos últimos 12 meses.

Se estes pressupostos se cumprem, concomitantemente com a inexistência de erros de declaração da idade e do número de filhos, a razão P (i) / F (i) = 1.

Mas, em populações reais, sobretudo quando a fecundidade experimenta declínios, tal relação costuma ter um comportamento crescente à medida que aumenta a idade das mulheres, e assume valores superiores à unidade. O fator

de correção das taxas de fecundidade atuais será selecionado a partir da experiência reprodutiva das mulheres mais jovens, das de 20 a 24 anos ou das de 25 a 29 anos de idade, respectivamente, P (2) / F (2) ou P (3) / F (3).

b) Mortalidade infantil

Já no que concerne à Mortalidade Infantil, a técnica empregada foi a variante de Trussell (1975), da técnica originalmente proposta por Brass (1971).

A informação básica necessária para estimar a mortalidade infantil é a seguinte:

- mulheres em idade fértil (de 15 a 49 anos de idade), classificadas por grupos qüinqüenais de idade [M (i); i = 1 = 15 a 19 anos, i = 2 = 20 a 24 anos,....., i = 7 = 45 a 49 anos],

- filhos tidos nascidos vivos, declarados pelas mulheres, classificados segundo os mesmos grupos qüinqüenais de idade das mulheres [FNV (i)], e

- filhos sobreviventes, declarados pelas mulheres, classificados segundo os mesmos grupos qüinqüenais de idade das mulheres [FV (i)]. Esta informação está referida ao momento do censo.

Com estes dados podem ser calculadas, inicialmente, as proporções de filhos falecidos com respeito ao total de filhos nascidos vivos, segundo a idade das mulheres:

$$D(i) = 1 - [\, FV(i) / FNV(i) \,]$$

onde D (i), por si só, constitui uma medida da mortalidade, mas tem a limitação de não ser um indicador convencional, por estar referido à idade das mulheres e não à idade dos filhos. Nesse sentido, Brass desenvolveu um procedimento que permite transformar as proporções D (i) em medidas convencionais de mortalidade nos primeiros anos de vida. O autor demonstrou haver uma relação empírica entre D (i) e a probabilidade de morte desde o nascimento até uma idade exata x, Q (x). A relação entre estas medidas se estabelece mediante as seguintes relações:

Quadro 12.2 – Correspondência entre D(i) e Q(x)

i	Q (x)	=	K (i) * D (i)
1	Q (1)	=	K (1) * D (1)
2	Q (2)	=	K (2) * D (2)
3	Q (3)	=	K (3) * D (3)
4	Q (5)	=	K (4) * D (4)
5	Q (10)	=	K (5) * D (5)
6	Q (15)	=	K (6) * D (6)
7	Q (20)	=	K (7) * D (7)

K (i) é um fator muito próximo a 1 (um), o que permite transformar as D (i) em Q (x). Brass calculou um conjunto de valores de K (i) com base em um modelo teórico no qual intervêm uma função de fecundidade e uma lei de mortalidade. Foi demonstrado que, neste modelo, os multiplicadores dependem

principalmente da estrutura por idade da fecundidade, no sentido de que, quanto mais cedo as mulheres tiverem seus filhos, maior será o tempo de exposição ao risco de morte de seus filhos. Por esse motivo, os parâmetros de entrada para a obtenção dos valores de K(i) são indicadores dessa estrutura, P (1) / P (2) e P (2) / P (3), sendo P (i) a parturição média das mulheres no grupo etário i (i = 1 para 15 a 19, i = 2 para 20 a 24, etc.).

Cada Q (x) estimada corresponde a momentos distintos anteriores à data de referência do censo. Na medida em que se avança na idade das mulheres, a estimativa corresponde a um passado mais distante. Feeney (1976, 1980) foi o primeiro a desenvolver idéias a respeito de como localizar as estimativas no tempo. Descobriu que, ao supor um declínio linear, qualquer que seja a intensidade do declínio, a mortalidade é a mesma num momento de tempo anterior ao censo. A partir dessa idéia, Coale e Trussell (1977) desenvolveram um procedimento para determinar os valores de t (x) (número de anos anteriores ao censo) para cada Q (x) estimada. Baseando-se nos modelos de fecundidade de Coale e Trussell e nas quatro famílias (Norte, Sul, Leste e Oeste) de tábuas-modelo de mortalidade de Coale e Demeny (1966), Trussell elaborou quatro conjuntos de regressões para o cálculo de K (i) e t (x).

Com o propósito de se obter uma medida comparável no tempo, usando-se as tábuas-modelo de Coale e Demeny (1966), as tábuas-modelo Brasil de Frias e Rodrigues (1981) ou uma transformação logital, todas as Q (x) obtidas, mediante o emprego da técnica de Trussell, foram transformadas em Q (1), ou seja, em probabilidades de um recém-nascido falecer antes de completar o primeiro ano de vida, devidamente localizadas no tempo.

A técnica de Trussell foi aplicada às informações provenientes dos Censos Demográficos 1970, 1980, 1991 e 2000, formando um conjunto de estimativas de Q (1), que compreendia o período de 1960 - 1998. Tais estimativas foram suavizadas, mediante médias móveis, de maneira a eliminar possíveis flutuações que normalmente existem, derivadas, principalmente, da má declaração, por parte das mulheres, quanto ao número de filhos nascidos vivos e sobreviventes. A esta série, já suavizada, ajustou-se uma função logística. Deve-se esclarecer que se teve o cuidado para que os ajustes realizados não implicassem diferenças significativas dos valores observados, especialmente nos anos próximos a 2000, o que foi possível, simulando-se valores para as assíntotas inferior e superior da função logística. Ao proceder assim, pouca variabilidade foi encontrada entre os valores observados e ajustados, ao longo do período considerado. Isto proporcionou um ajuste bastante satisfatório das probabilidades de morte no primeiro ano de vida, possibilitando realizar projeções das mesmas para o ano 2000.

A fim de realizar as análises com conhecimento do significado das estimativas, são os seguintes os pressupostos implícitos da técnica utilizada:

- que a fecundidade tenha permanecido constante num passado recente. Segundo Feeney (apud MANUAL X...,1983), as estimativas são suficientemente robustas de modo que os desvios não têm importância se não se cumpre esse suposto,

- que a mortalidade na infância tenha uma evolução linear através do tempo,

- que as leis de mortalidade e fecundidade usadas no modelo representem as mesmas condições da população em estudo,

- que não haja associação entre a mortalidade das mulheres (mães) e de seus filhos. Obviamente, não se tem informação sobre a mortalidade dos filhos cujas mães já faleceram, e, no caso em que sua mortalidade fosse maior que a dos filhos com mães vivas, as estimativas da mortalidade nos primeiros anos de vida estaria subestimada, e

- que não exista associação entre a mortalidade infanto-juvenil e a idade das mulheres (mães).

Além destes pressupostos, a informação básica deve cumprir certas condições:

- que não haja omissão diferencial na declaração do número de filhos nascidos vivos e sobreviventes,

- que não haja mortalidade diferencial entre os filhos das mulheres que declaram e as que não declaram a informação, e

- que a declaração da idade das mulheres seja correta.

As taxas estimadas segundo esses critérios são apresentadas a seguir.

Tabela 12.16 - Taxas de Fecundidade Total e Taxas de Mortalidade Infantil com base em dados censitários - Brasil - 1980-2000

Ano	Taxa de fecundidade total	Taxa de mortalidade infantil (por 1 000 nascidos vivos)
1980	4,4	82,6
1991	2,9	47,7
2000	2,3	28,3

Fonte: IBGE, Censo Demográfico 1980/2000.

c) Conceitos e definições

No Censo 2000, os quesitos do bloco de fecundidade foram indagados a todas as mulheres com 10 anos ou mais de idade, na data de referência do censo, ou seja, nascidas até 31 de julho de 1990.

- Filhos tidos nascidos vivos até 31 de julho de 2000 – Considerou-se como filho tido nascido vivo aquele que, após a expulsão ou extração completa do corpo da mãe, independentemente do tempo de duração da gravidez, manifestou qualquer sinal de vida (respiração, choro, movimentos de músculos de contração voluntária, batimento cardíaco, etc.), ainda que tenha falecido em seguida. O número de filhos tidos nascidos vivos foi registrado segundo o sexo.

- Filhos tidos que estavam vivos em 31 de julho de 2000 – O número de filhos tidos que estavam vivos em 31 de julho de 2000 foi registrado segundo o sexo.

- Sexo do último filho tido nascido vivo até 31 de julho de 2000. As opções de resposta foram: masculino ou feminino.

- Data de nascimento ou idade presumida do último filho tido nascido vivo até 31 de julho de 2000 – Registrou-se o mês e o ano de nascimento do último filho tido nascido vivo até 31 de julho de 2000. Se, esgotados todos os esforços, não fosse possível a obtenção do mês e ano de nascimento do último filho tido nascido vivo, registrou-se sua idade presumida, fornecida pela pessoa entrevistada.

- Sobrevivência do último filho tido nascido vivo até 31 de julho de 2000 – As opções de resposta foram: sim; não; não sabe.

- Filhos tidos nascidos mortos – São os óbitos ocorridos de todo o produto da concepção, a partir da 28a semana de gestação, antes de sua extração ou expulsão completa do corpo da mãe. A informação foi coletada segundo o sexo.

- Taxa de Fecundidade Total – Expressa o número de filhos que, em média, teria uma mulher, pertencente a uma coorte hipotética de mulheres, que durante sua vida fértil tiveram seus filhos de acordo com as Taxas de Fecundidade, por Idade do período em estudo e que não estiveram expostas a riscos de mortalidade desde o nascimento até o término do período fértil.

- Taxa Específica de Fecundidade por idade – É geralmente calculada por grupo qüinqüenal de idade, desde os 15 até os 49 anos. A taxa resulta da divisão do número de filhos nascidos vivos de mulheres do grupo de idade, em um período de tempo próximo à data do Censo Demográfico, usualmente os últimos 12 meses, pelo total de mulheres do mesmo grupo etário.

- Coorte – Conjunto de indivíduos que estão experimentando um acontecimento similar no transcurso de um mesmo período de tempo.

- Coorte hipotética de mulheres – Num censo demográfico, a classificação das mulheres por grupos qüinqüenais de idade, dentro do período fértil, está associada a uma análise de período. Uma análise de coorte considera, por exemplo, um grupo de mulheres que ingressa no período fértil e, ao longo do tempo, observa-se o comportamento do mesmo frente aos riscos de procriação. Entretanto, em um único censo demográfico, mesclam-se distintas gerações de mulheres e, de acordo com o conceito da Taxa de Fecundidade Total, supõe-se o acompanhamento de como essas mulheres vão tendo seus filhos ao longo do tempo. Por esse motivo, na definição conceitual da Taxa de Fecundidade Total, é necessário enfatizar que o grupo de mulheres em questão trata-se de uma coorte hipotética.

- Taxa de Mortalidade Infantil – É definida como o número de óbitos de menores de 1 ano de idade (por mil nascidos vivos), em determinada área geográfica e período, e interpreta-se como a estimativa do risco de um nascido vivo morrer durante o seu primeiro ano de vida.

d) Expansão da amostra e divulgação dos resultados

É importante assinalar que os pesos preliminares para a expansão da amostra, que viabilizaram os cálculos dos indicadores de Fecundidade e Mortalidade Infantil, foram obtidos através do inverso da fração de amostragem, observada no setor censitário, e calibrados de tal forma que as estimativas de total de pessoas por sexo se igualassem aos valores correspondentes do Conjunto Universo – que compreende o conjunto de características básicas investigadas para o total da população e dos domicílios, em cada Unidade da Federação. Os pesos, assim determinados, foram atribuídos a cada domicílio; todas as pessoas residentes em um mesmo domicílio receberam peso idêntico ao do domicílio.

Os níveis geográficos para a divulgação dos Resultados Preliminares dos temas Fecundidade e Mortalidade Infantil foram o Brasil como um todo e suas cinco grandes regiões. Para o tema Fecundidade foram divulgados indicadores representativos do nível – de 1940 até 2000 – e do padrão etário – de 1980 até 2000. O nível da Fecundidade está representado pela Taxa de Fecundidade Total e o padrão etário pelas Taxas Específicas de Fecundidade por Idade. No caso da Mortalidade Infantil, foram apresentadas as respectivas séries históricas das Taxas de Mortalidade Infantil, abrangendo o período de 1990 a 2000.

12.5 Resultados da amostra

Comparativamente ao processo de apuração dos Resultados do Universo, a tarefa referente aos dados da amostra, pelo maior volume do trabalho de crítica e imputação e por incorporar as tarefas da Codificação e Expansão, apresenta-se bem mais intensa. A seguir, são tratadas cada uma dessas atividades.

12.5.1 Codificação

Tendo como referência o Censo de 1991, a Codificação pode ser apontada como uma das partes da apuração dos dados que incorporou um grande número de modificações para o Censo. A maior novidade foi a implantação de uma rotina de aplicação de códigos – codificação propriamente dita – reformulada, que tinha como objetivo garantir a qualidade do trabalho, com um prazo de execução bastante reduzido.

Assim, nesse contexto de procura de maior eficiência, foi também estabelecida a etapa de Verificação, que compreendeu um conjunto de ações que buscavam aprimorar a atividade de aplicação de códigos, vindo a constituir-se numa inovação no processo de trabalho da apuração. Entenda-se, portanto, a fase de Codificação do Censo Demográfico 2000, compreendendo duas fases: Aplicação de Códigos e Verificação de Códigos.

12.5.1.1 Formação dos lotes

Para a formação dos lotes a serem trabalhados, na Codificação e nas atividades de Crítica Intra-Registros das informações do CD 1.02, foram considerados, em boa parte, os critérios utilizados para os dados do Conjunto Universo. Assim, primeiramente, os setores foram classificados/ordenados levando-se em conta a Unidade da Federação, a situação do domicílio, a mesorregião, a microrregião, o município, o distrito e o subdistrito.

A diferença em relação ao Conjunto Universo ficou por conta do processo de escolha dos lotes. Partindo do mesmo limite superior de 90000 domicílios particulares ocupados – DPO –, foram definidos os lotes, buscando-se respeitar os vários níveis da classificação.

Para as Unidades da Federação, onde o total de DPO não ultrapassou os 90 000 domicílios, para cada situação urbana e rural, os lotes foram formados com todos os setores dessas UFs; caso contrário, cada lote foi composto pela mesorregião. As áreas urbanas dos municípios de Salvador, Belo Horizonte e Porto Alegre formaram, cada uma, um lote exclusivo. Para a área urbana do município de São Paulo, foram formados lotes através de áreas pré-definidas, agrupando-se os seus distritos. No caso do município do Rio de Janeiro, foram formados lotes através de grupamentos de subdistritos.

A tabela seguinte apresenta algumas informações para os 215 lotes que foram formados:

Tabela 12.17 - Número de lotes e de domicílios particulares ocupados e de pessoas recenseadas, informados no SIGC, referentes aos Questionários da Amostra, segundo as Unidades da Federação

Unidades da Federação	Número de lotes	DPO	Pessoas
Brasil	**215**	**5 247 272**	**20 199 963**
Rondônia	2	42 966	171 504
Acre	2	16 711	70 878
Amazonas	2	63 694	314 039
Roraima	2	9 663	41 621
Pará	3	145 106	689 625
Amapá	2	11 653	55 192
Tocantins	2	42 648	175 455
Maranhão	2	149 698	701 509
Piauí	2	94 195	404 761
Ceará	14	200 174	864 021
Rio Grande do Norte	2	92 118	388 864
Paraíba	2	116 895	486 823
Pernambuco	10	223 992	932 492
Alagoas	2	77 516	347 126
Sergipe	2	54 825	230 323
Bahia	15	376 477	1 593 721
Minas Gerais	25	608 183	2 339 334
Espírito Santo	2	98 067	368 075
Rio de Janeiro	19	437 509	1 503 726
São Paulo	40	1 118 165	4 017 990
Paraná	20	332 772	1 215 472
Santa Catarina	12	191 741	69 139
Rio Grande do Sul	15	361 390	1 206 575
Mato Grosso do Sul	2	68 316	25 017
Mato Grosso	2	85 402	325 034
Goiás	10	173 456	614 801
Distrito Federal	2	53 940	199 442

Fonte: IBGE, Censo Demográfico 2000, Sistema de Indicadores Gerenciais da Coleta.

12.5.1.2 Modelo de codificação automática e assistida

O sistema de codificação automática/assistida, disponível no IBGE para ser utilizado por pesquisas e censos, adotava um modelo idealizado em 1988 por ocasião da realização do Censo Experimental daquele ano. Esse modelo serviu de base para a confecção do sistema utilizado na codificação do Censo Demográfico 1991 e pelas PNADs da última década. Ao longo desses anos, o sistema foi aprimorado em suas funções, porém, nenhuma revisão foi efetuada no seu modelo conceitual.

Modelo utilizado no Censo Demográfico 1991

O modelo utilizado no Censo Demográfico 1991, que tinha sido implementado num sistema de codificação automática/assistida, apresentava, resumidamente, as características descritas a seguir:

a) para cada quesito a codificar, era necessário montar um arquivo (arquivo de descritores) com as várias descrições possíveis para o quesito em questão, contendo o código correspondente. A partir desses arquivos, era criado um banco de códigos contendo, para cada quesito a ser codificado, as descrições correspondentes e as palavras em sua forma normal e fonética e os códigos associados. Esta etapa preparatória era realizada, de forma centralizada, somente uma vez antes de iniciar a codificação e poderia sofrer atualizações, de forma centralizada, dependendo das necessidades surgidas ao longo do processo de codificação;

b) o processo de codificação era realizado sobre um arquivo magnético com os dados dos questionários capturados (via digitação ou reconhecimento ótico de caracteres);

c) a comparação do texto oriundo do quesito com os textos das descrições era feita palavra a palavra. Ao se dividir os textos em palavras, eram eliminadas as preposições e os artigos, além das palavras de uma lista opcional a serem eliminadas, fornecida pelo responsável pela codificação do quesito. O objetivo desta lista era possibilitar a eliminação de palavras que não contribuiriam para a codificação do texto (ex.: a palavra igreja na codificação do quesito religião). Na divisão do texto em palavras, os sinônimos eram, também, considerados. Toda palavra, num conjunto de palavras consideradas sinônimos, era convertida para uma palavra padrão. As palavras também sofriam uma transformação fonética que consistia, em resumo, na eliminação do plural, do gênero (masculino e feminino) e na substituição de uma letra por outra do mesmo som (esta técnica auxilia no reconhecimento das palavras e resolve também alguns erros de grafia ou de digitação). O método utilizado para tratamento do texto a ser codificado foi o mesmo utilizado para geração do banco de códigos a partir dos arquivos de descritores.

d) a aplicação do código era feita a partir da comparação dos textos obtidos dos questionários com os textos armazenados no banco de códigos, com objetivo de atribuir ao texto um código numérico. Para esta comparação, o texto do questionário era dividido em palavras e cada palavra pesquisada em sua forma normal. Caso não fosse encontrada, era feita a pesquisa em sua forma fonética. No caso do texto conter mais de uma palavra, ou uma palavra não reconhecida, esta era pesquisada no contexto das descrições, onde apareciam as outras palavras, podendo ser feitas sugestões de palavras semelhantes. Reconhecidas as palavras, eram pesquisadas nas descrições onde apareciam, podendo surgir três situações:

d.1) uma única descrição era encontrada: codificação automática;

d.2) mais de uma descrição era encontrada: o codificador devia escolher uma das descrições apresentadas;

d.3) nenhuma descrição era encontrada: o codificador digitava um novo texto.

O modelo de codificação implementado apresentava claramente duas etapas de codificação: uma primeira chamada de codificação automática ou pré-codificação (*batch*) e uma segunda chamada de codificação assistida ou complementar (*on-line*). Na codificação automática eram codificados, questionário a questionário, aqueles textos para os quais era encontrado um único código no banco de códigos (d.1), o que determinava o fim do processo de codificação para o quesito do questionário. Na codificação assistida, um codificador fazia, *on-line*, a escolha do código, entre os apresentados com base na descrição (d.2) ou digitava um novo texto para o quesito em questão (d.3).

Para todos os quesitos a codificar de um questionário, o sistema realizava a codificação automática na seqüência em que os quesitos estavam no questionário. O codificador atuava somente após a execução da pré-codificação (processamento *batch*) para completar a tarefa nos questionários com codificação pendente, obedecendo a ordem dos quesitos no questionário. É importante registrar que um mesmo codificador era responsável pela codificação assistida dos diferentes quesitos (ocupação, atividade, religião etc.).

Avaliação do modelo

Uma das maiores atividades no processamento de um censo é a codificação, talvez só inferior à captura de dados.

O modelo até então disponível era, naturalmente, superior à codificação manual, processo lento que requer um grande contingente de pessoas com perfil adequado e consome muito tempo.

O esquema utilizado no modelo disponível (separar o texto em palavras, eliminar algumas palavras, utilizar palavra padrão – sinônimo, abreviatura – transformação fonética, e criar o banco de códigos, levando em conta a freqüência das palavras e a forma de busca do código a ser associado), com certeza, trouxe maior eficiência ao processo.

Se o conjunto de descrições encontradas no Censo Demográfico 1991 para os distintos quesitos a codificar fosse incorporado aos arquivos de descritores, com certeza, aumentaria bastante a eficiência e a qualidade da codificação do Censo Demográfico 2000, utilizando-se o modelo disponível.

No modelo descrito, pode-se verificar que muito investimento foi feito na "máquina para obter um código a partir de um texto". O modelo chegou a um ponto em que, apesar de ser possível melhorar o custo/benefício, as modificações seriam difíceis de implementar e os ganhos pouco perceptíveis.

O que chamou a atenção dos responsáveis pela codificação do Censo Demográfico 2000 não foi a eficiência dessa máquina, mas a forma como foi utilizada.

Nesse modelo, a codificação era feita questionário a questionário. Tanto na codificação automática, como na assistida, os quesitos eram trabalhados, um a um, na seqüência em que apareciam no questionário. O codificador tratava todos os quesitos do questionário, seqüencialmente. Isto fez pensar o seguinte: criou-se uma bela "máquina" de codificar, mas que estava sendo utiliza-

da, simulando um processo de codificação manual. Nesse processo, um codificador ia retirando questionários de uma pilha e codificando os quesitos, um a um, sem precisar recorrer ao velho e tradicional manual de códigos (transformado no banco de códigos). Automatizou-se um processo, porém, não se o informatizou. Isto é, as tarefas manuais foram transferidas para o computador, retirando-se uma parte do trabalho manual do codificador. A outra parte desse trabalho continuava sendo feita manualmente, porém, assistida pelo computador, sem que nenhuma informação gerada durante o processo fosse utilizada para aumentar a eficiência do mesmo. Desta forma, a tarefa mais difícil era resolvida por uma única pessoa, que devia codificar todos os quesitos do questionário, uma vez que a "inteligência" colocada na "máquina" não conseguia resolver a codificação de forma automática. Por outro lado, a partir da existência de um sistema de codificação automática/assistida, criou-se a idéia de que com um pequeno treinamento (basicamente no uso do sistema) uma pessoa estaria, rapidamente, habilitada a codificar qualquer quesito.

A codificação apresenta diferentes graus de dificuldade que variam de um quesito para outro. Os índices de codificação automática do Censo Demográfico 1991, para quesitos religião (da ordem de 80%), Unidade da Federação ou país estrangeiro de nascimento (80%), município (80%), curso concluído (79%) são totalmente diferentes ,se comparados com alcançados em ocupação (cerca de 20%) e atividade (inferior a 40%).

Religião, local de nascimento, município e curso concluído causam menos problemas, em virtude de serem mais facilmente entendidos e apresentarem um conjunto de possibilidades bem limitado. Os maiores problemas resumem-se a erros de grafia, uso de sinônimos, abreviaturas das mais diversas formas, etc.

Por outro lado, os quesitos ocupação e atividade apresentam uma grande variedade de descrições, às vezes feitas de uma forma muito geral para que o sistema as codifique. Estes quesitos podem depender de conceitos e entendimentos muitas vezes restritos à esfera de conhecimento de especialistas no assunto, o que não se pode esperar de codificadores que passam por um pequeno período de treinamento. Este é, com certeza, o ponto mais crítico desse modelo.

Modelo utilizado no Censo Demográfico 2000

O modelo utilizado para aplicação de códigos no Censo Demográfico 2000 foi uma combinação entre a aplicação manual e a automática. Representa algo como uma pós-codificação, em que os códigos para cada quesito só foram atribuídos após se ter conhecimento de todas as descrições encontradas nos questionários de um determinado conjunto (lote de codificação).

A aplicação de códigos foi feita em cada um dos lotes constituídos para as descrições registradas no Questionário da Amostra – CD 1.02 – correspondentes às informações coletadas em aberto, para os quesitos Religião, Migração, Curso, Ocupação e Atividade. O tema Migração era constituído por 5 quesitos de codificação, a saber: “Qual é a unidade da federação ou país estrangeiro de nascimento?” “Qual é a unidade da federação ou país estrangeiro de residência anterior?” “Em que município residia em 31 de julho de 1995?” “Em que unidade da federação ou país estrangeiro residia em 31 de julho de 1995?”, e “Em que município e unidade da federação, ou país estrangeiro, trabalha ou estuda?”. Compreende-se assim que, exceto em Migração, tema e quesito de codificação sejam a mesma coisa, e que foram codificados 9 quesitos em cada um dos 215

lotes trabalhados. O tamanho do lote de codificação foi determinado em função do lote preparado para as etapas de codificação e crítica.

O modelo apresentou, para cada quesito a ser codificado, as seguintes características:

- as descrições encontradas nos questionários foram armazenadas para cada pessoa e transferidas para uma base de dados em separado. Nesta base, as informações semelhantes, para cada tema, foram unificadas e grupadas, segundo os critérios automáticos de aplicação de códigos. As descrições foram grupadas por semelhança (através de palavras-chave que determinavam o código ou códigos associados usando a "máquina");

- a aplicação de códigos foi feita, nesta base de descrições, por lote. Todos os registros desta base, com as mesmas características, foram "codificados" de uma única vez, apresentando a freqüência de descrições que eles representavam. O sistema atuou sempre com controle de freqüência, de ocorrências das descrições, individuais e grupadas no lote;

- a aplicação de um código era, inicialmente, confirmada pelo codificador do tema, mediante a validação do código proposto pelo sistema. Depois de atingir uma determinada freqüência acumulada e após a validação do orientador, a aplicação do código passava a ser feita de forma automática, sem precisar da confirmação do codificador; e

- terminada a aplicação de códigos do arquivo de descrições básicas do lote, os arquivos originais (lote de codificação), contendo registros individualizados e o arquivo com as descrições já codificadas, passaram por um batimento para atribuição do código aos registros individuais no lote.

As vantagens do modelo utilizado em relação ao modelo anterior, entre outras, são: maior automação do processo de aplicação de códigos, uniformidade do processo e utilização de codificadores especializados em cada tema.

Etapas do Sistema

O sistema implementado era composto pelas seguintes etapas:

Extração de textos

O sistema lia o arquivo com o lote a codificar e extraía dos registros de pessoas, os textos referentes ao quesito (tema) selecionado, guardando-os com sintaxe única, em arquivos do banco com a freqüência com que ocorreiam. Em seguida, os textos eram verificados no banco de códigos, separando os textos corretos (aqueles que tem codificação única ou múltipla) e os não corretos (aquele para os quais não existe código) em relação ao descritor. Criavam-se,assim, três grupos de textos: os que codificavam, os que tinham codificação múltipla, e aqueles para os que não haviam códigos associados. Durante a extração, o sistema contabilizava totais de domicílios e pessoas lidas, quantidade de textos extraídos e quantos textos distintos existiam (textos diferentes). Ao final do processo, eram apresentadas as estatísticas da extração.

Correção de caracteres

O sistema lia o arquivo com os textos extraídos. Os textos que não tinham código associado (codificação não é única nem múltipla) eram exibidos em caixa de lista, permitindo sua correção; se um texto em trabalho já tivesse passado por correção anteriormente, o sistema sugeria a sua última correção feita. Se a freqüência de alguma correção atingisse um valor determinado como limite e houvesse autorização do orientdor, era feita automaticamente (aprendizado através do uso). A freqüência para correção automática era única para o quesito durante o processamento dos lotes.

Durante esta etapa, os textos, corrigidos ou não, eram mantidos em um arquivo de correções para serem pesquisados no arquivo descritor. Essa pesquisa era feita sempre ao iniciar a correção. Caso a correção de caracteres fosse re-executada, o sistema fazia a pesquisa no arquivo descritor, retornando como errados aqueles que não tinham sido encontrados. Caso a correção de caracteres fosse concluída sem re-execução, este batimento era feito na próxima etapa (Agrupamento por códigos) e os textos não encontrados eram agrupados como ainda não codificados.

Agrupamento por códigos

Os textos corrigidos eram codificados dentro dos grupos em cada tema em códigos únicos, múltiplos e não codificados, e disponibilizados para confirmação. Durante o agrupamento, era feita uma contagem para cada uma dessas situações, sendo as estatísticas apresentadas ao final do processo.

Confirmação de códigos

Os textos grupados eram exibidos em uma lista para confirmação. Se o código fosse único, eram exibidos o código e as descrições associadas. Se houvesse códigos múltiplos, estes eram exibidos, bem como as descrições associadas a cada um deles, permitindo a seleção de um código. Se não existisse código, o sistema permitia digitar um novo texto. Se os textos que codificavam (únicos ou múltiplos) já tivessem sido confirmados anteriormente e a freqüência de confirmação tivesse atingido um valor previamente determinado como limite, e houvesse autorização do operador, o código era confirmado automaticamente (aprendizado através do uso), sendo exibidos somente os não confirmados automaticamente. A freqüência para confirmação automática era única para o quesito durante o processamento dos lotes. Se, nesta etapa, algum texto ainda não tivesse sido codificado, era levado para a etapa de resolução de códigos pendentes.

Atribuição de código aos registros individuais

Caso houvesse textos pendentes para encerrar a aplicação de códigos do quesito em questão, o sistema abria uma janela, que permitia realizar este trabalho de forma assistida, apresentando o texto e algumas variáveis auxiliares predefinidas para auxílio da codificação, possibilitando a solução da pendência.

Uma vez codificado o conjunto de textos extraídos do lote, os arquivos originais passavam por um batimento com o arquivo das descrições grupadas já codificadas para atribuição do código aos registros individuais.

Acompanhamento da codificação

O acompanhamento foi realizado por lote de codificação e tema. O lote era uma unidade de trabalho para aplicação de código e, assim sendo, o sistema gerava, para cada um, o quadro de status que permitia saber as fases concluídas para cada quesito e as estatísticas de codificação por quesito (percentuais de codificação automática e assistida). Os codificadores só podiam iniciar o trabalho em um novo lote, quando todas as fases de trabalho no lote atual tivessem sido completadas. Este controle era feito de forma automática pelo sistema.

12.5.1.3 Aplicação de códigos

Como foi visto, a aplicação de um código podia acontecer automaticamente, ou decorrer do trabalho do operador, assistido pela rotina de aplicação de códigos. Esse trabalho normalmente passava pelas seguintes etapas de execução: extração de textos, correção de caracteres, agrupamento por códigos, confirmação de códigos, atribuição de códigos e aplicação assistida. Essas etapas serão tratadas agora, sob o ponto de vista da operação da rotina de aplicação de códigos.

a) Rotina de Aplicação de Códigos

A rotina trabalhava os textos resultantes da digitalização e do reconhecimento ótico – textos originais – de uma só vez e com critérios uniformes. Recebido um lote e escolhido um quesito de codificação, a etapa extração de textos separava, entre todas as pessoas investigadas, os textos originais distintos para, em seguida, submetê-los a um interpretador que, após a utilização de recursos de divisão em palavras, eliminação de preposições, artigos, plurais e gênero, substituição de letras e uso de sinônimos, fazia uma pesquisa, normal e fonética, em um banco descritor temático de textos.

Todos os textos originais que não encontravam correspondência no banco descritor, por insuficiência deste ou por erros de grafia e/ou reconhecimento ótico, eram objeto de trabalho do operador na etapa seguinte, correção de caracteres. Nessa etapa, o operador podia manter ou adequar os textos, para que, depois disso, fossem novamente pesquisados junto ao banco.

Na etapa seguinte, agrupamento por códigos, o sistema classificava, sem o auxílio do operador, cada texto modificado ou não no estágio anterior, em uma das seguintes situações: código único, código múltiplo e sem código. Na seqüência, o operador realizava a confirmação de códigos, onde as sugestões oferecidas pela rotina eram analisadas, podendo ser confirmadas ou levadas, juntamente com os textos sem código, à condição de pendência, para serem resolvidas mais tarde, quase sempre, com o auxílio de variáveis auxiliares.

Passando à etapa atribuição de códigos, a rotina atribuía os códigos aos registros das pessoas que constituíam o lote original, permanecendo ainda sem informação os correspondentes aos textos pendentes, cuja solução acontecia durante o estágio subseqüente, o de aplicação assistida.

Na aplicação assistida - última etapa da aplicação de códigos – a aplicação de código era realizada pessoa a pessoa, e o operador podia, novamente, fazer a adequação dos textos através da correção de caracteres. Dependendo do quesito a ser codificado, era possível também contar com o auxílio de variáveis auxiliares, ou seja, de outros dados originalmente informados no questionário, a saber:

- quesito ocupação – as variáveis auxiliares eram: atividade – quesito 4.46; espécie de curso mais elevado concluído – quesito 4.35; rendimento de trabalho – quesito 4.51 e 4.52; posição na ocupação - quesito 4.47 e número de empregados - quesito 4.49;

- quesito atividade – eram as seguintes as variáveis auxiliares: ocupação – quesito 4.45; rendimento de trabalho – quesito 4.51 e 4.52; posição na ocupação – quesito 4.47 e número de empregados – quesito 4.49;

- quesito curso – as variáveis eram: curso mais elevado que freqüentou – quesito 4.32; última série concluída com aprovação – quesito 4.33; se concluiu o curso no qual estudou – quesito 4.34;

- quesito "município de residência em 31/07/1995" – a variável auxiliar era o quesito 4.26 – UF ou país estrangeiro de residência em 31/07/1995; e

- quesito "UF ou país estrangeiro de residência em 31/07/1995" – a variável auxiliar era o quesito 4.25 – município de residência em 31/07/1995.

Ainda na aplicação assistida, a necessidade de esclarecimento das situações de dúvida de aplicação de códigos podia determinar a consulta aos técnicos das unidades regionais e/ou aos especialistas temáticos da Diretoria de Pesquisas.

A automatização podia ocorrer, após autorização do orientador, nas etapas correção de caracteres e/ou confirmação de códigos, quando um mesmo procedimento se repetia a partir de um número de vezes. Nesse caso, o limite exigido para o início do tratamento automático variou, na correção de caracteres, de 3 a 5 ocorrências, dependendo do quesito; já para confirmação de códigos, foi de 5 ocorrências em Religião, Migração e Curso, enquanto em Ocupação e Atividade estabeleceu-se, inicialmente, 10 ocorrências, baixadas para 5 com o andamento do trabalho.

Especificamente para algumas declarações dos quesitos Ocupação e Atividade, houve a determinação de que os operadores utilizassem o recurso de aplicação de códigos genéricos – códigos alfanuméricos. No quesito Ocupação, fez-se uso dessa alternativa desde o início do processo de aplicação de códigos, enquanto no quesito Atividade a implantação aconteceu com o trabalho já em andamento.

A utilização desse recurso teve como objetivo viabilizar o cumprimento dos prazos para conclusão da fase Aplicação de Códigos. Foi bastante útil nas situações de textos muito freqüentes, aos quais estavam associados códigos múltiplos, cujas atribuições exigiam sempre a passagem pela etapa aplicação assistida, o que aumentava, em muito, o tempo de trabalho.

Para essas situações, os respectivos códigos específicos finais foram atribuídos após terminada a fase Aplicação de Códigos, através de um programa especial, que implementava uma tabela de conversão ou de atribuição de código, estabelecida em conjunto com especialistas temáticos.

Para essa conversão, partia-se do código genérico atribuído e, através da consulta às variáveis auxiliares, fazia-se a decodificação:

- no caso do genérico do quesito ocupação, eram consultadas sempre a posição na ocupação e o código da atividade e, em alguns casos, a variável número de empregados; e

- para o genérico do quesito atividade, utilizava-se, necessariamente, o código da ocupação e, dependendo da situação, também a variável posição na ocupação.

A fase Aplicação de Códigos do lote trabalhado só estava encerrada quando o operador fazia o envio ao sistema do lote já codificado, ocasião em que todas as pendências deveriam estar resolvidas. Em tempo de produção, o sistema podia ser consultado sobre as seguintes estatísticas sobre o lote:

- datas de início e término do trabalho;

- número de registros lidos;

- números de textos extraídos, distintos e corrigidos;

- números de textos com códigos únicos, múltiplos e sem códigos; e

- números de textos distintos levados à aplicação assistida e de pessoas correspondentes

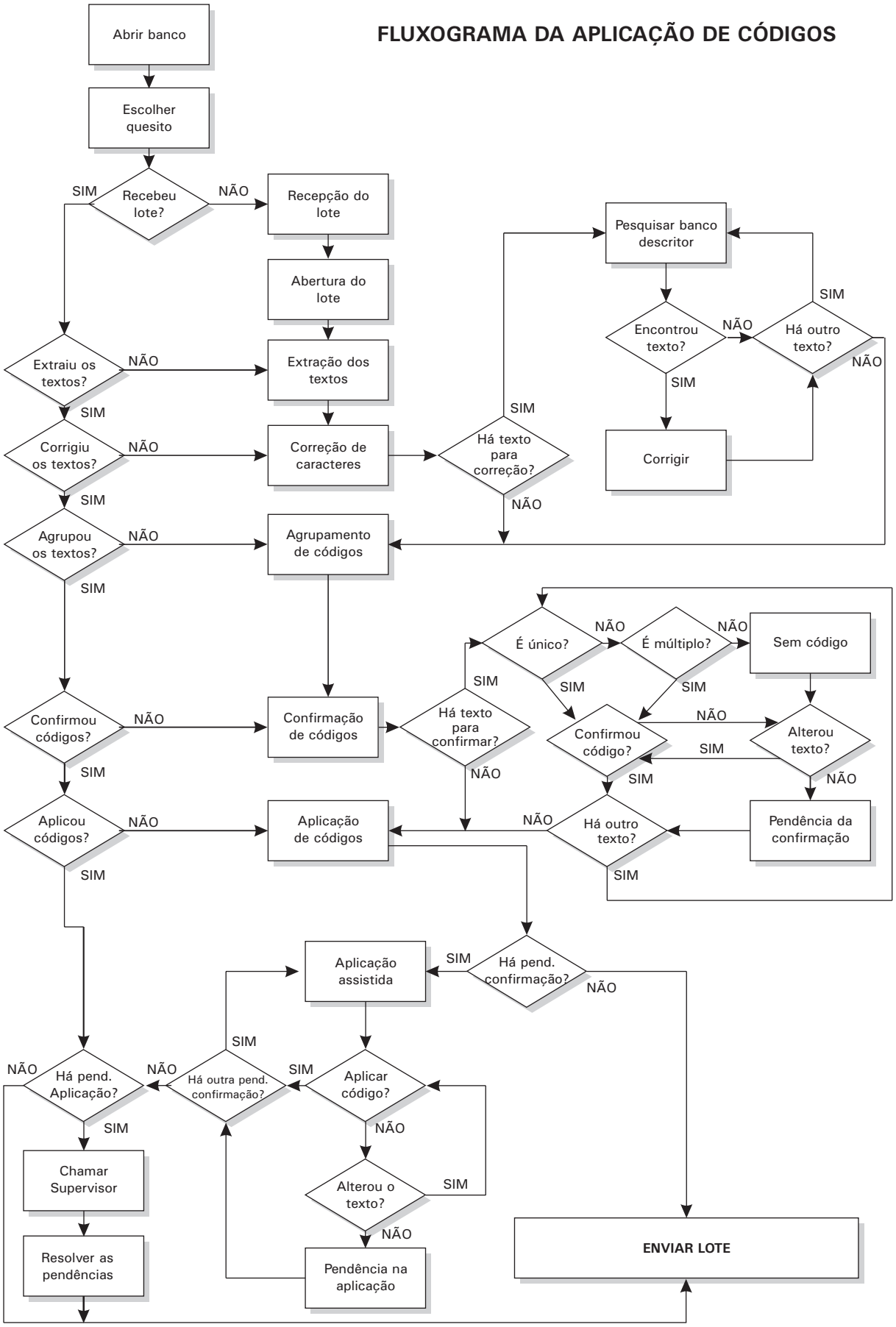
O esquema seguinte ilustra o fluxo de trabalho da fase Aplicação de Códigos, enquanto a Tabela 12.18 apresenta valores para alguns indicadores do processo.

Tabela 12.18 - Alguns indicadores relativos ao processo de aplicação de códigos

Indicadores	Quesito								
	Religião	Unidades da Federação/ País nasc.	Unidades da Federação/ País ant.	Município 5 anos	Unidades da Federação/ País 5 anos	Mun/Unidades da Federação/p. est. e trab.	Curso	Ocupação	Atividade
Tempo médio (em dias) de duração da aplicação de códigos por lote	3,2	1,2	1,3	2,4	1,2	1,2	1,4	33,7	32,8
Média de Registros trabalhados por dia	29 405	20 409	18 660	9 990	20 570	19 495	2 090	1 076	1 094
Média de textos extraídos trabalhados por dia	170	49	50	99	25	75	69	70	112
% textos corrigidos, em relação aos textos extraídos	45,22	55,29	61,23	24,94	65,08	29,14	20,72	24,59	38,43
% textos confirmados, em relação aos textos extraídos	7,39	6,57	4,68	24,00	9,80	50,15	15,05	7,21	4,40
% de textos distintos levados a aplicação assistida, em relação aos textos extraídos	55,32	95,11	96,78	75,01	93,88	83,35	91,74	85,38	59,42
% de registros levados a aplicação assistida, em relação ao total de registros	3,26	3,21	2,89	5,78	1,60	16,37	89,02	50,14	44,82

Fonte: IBGE, Censo Demográfico 2000.

Figura 12.3 – Fluxograma da aplicação de códigos



b) Treinamento e equipes de trabalho

As equipes para o trabalho de aplicação de códigos eram formadas por operadores/codificadores e por orientadores. Os operadores eram técnicos temporários, contratados para as atividades de apuração do censo, enquanto os orientadores faziam parte do quadro permanente do IBGE. Cada equipe desenvolvia a aplicação de códigos em um determinado tema. Além das tarefas de acompanhamento e supervisão dos trabalhos, cabia também aos orientadores avaliar as situações de automatização da correção de caracteres e da confirmação de códigos, cuja validação dependia de sua autorização, mediante o uso de senha específica.

A aplicação de códigos foi realizada no Rio de Janeiro, em duas instalações do IBGE e em dois turnos de trabalho, com exceção do tema Religião, cuja equipe atuou em regime de turno único. Especificamente no quesito Atividade, só houve um segundo turno durante os últimos três meses de trabalho. A tabela seguinte apresenta os quantitativos das equipes e os tempos gastos no trabalho, segundo cada quesito de codificação.

Tabela 12.19 - Tamanhos das equipes de aplicação de códigos e períodos de realização do trabalho, segundo os temas de codificação

Temas de codificação	Codificadores	Orientadores	Data	
			Inicial	Final
Total	**102**	**17**		
Religião	5	2	03/08/01	21/11/01
Curso	10	2	01/08/01	01/11/01
Migração	10	2	13/08/01	09/11/01
Ocupação	34	4	13/08/01	07/02/02
Atividade	43	7	18/07/01	11/03/02

Fonte: IBGE, Censo Demográfico 2000.

Os codificadores foram contratados em maio de 2001, com uma antecedência média de dois meses do início efetivo da aplicação de códigos Esse prazo atendeu às peculiaridades do treinamento das equipes, que foi realizado, inicialmente, para duas turmas, em duas etapas. A primeira delas, com a duração de 5 dias, abrangia alguns aspectos teóricos do censo e o conhecimento da rotina de aplicação de códigos, tendo como base o manual Procedimentos Operacionais para o Sistema de Codificação. A segunda etapa de capacitação, que durou em média 60 dias, estava totalmente voltada para a atividade prática do trabalho, tendo como ênfase qualificar os operadores no conhecimento dos bancos descritores e dar-lhes agilidade no uso da rotina; para tanto, durante esse período, foi realizada a aplicação de códigos, utilizando-se as declarações obtidas nos questionários do Censo Experimental.

Posteriormente, uma nova turma de operadores foi treinada, visando à criação do segundo turno de trabalho para aplicação de códigos no tema Atividade.

12.5.1.4 Verificação de códigos

A Verificação de Códigos teve como objetivo contribuir para a melhoria da qualidade do trabalho da fase Aplicação de Códigos. Operacionalmente, a tarefa consistiu em fazer, por amostragem, uma segunda atribuição de códigos

para todos os lotes de cada quesito de codificação, visando identificar as situações de divergência entre os códigos aplicados pelos dois operadores, codificador e verificador.

O trabalho de verificação foi realizado por 15 técnicos que, diferentemente da fase Aplicação de Códigos, atuavam em qualquer dos nove quesitos de codificação. A tarefa tinha início após a liberação do lote de codificação e da constituição dos correspondentes arquivos intermediários. Uma rotina especial fazia a gravação desses arquivos, um para cada lote, cujos registros continham a identificação da pessoa, o texto original, o texto após o estágio correção de caracteres, o código atribuído e, dependendo do quesito, as variáveis auxiliares. Constava, também, desse arquivo uma classificação que mostrava como a rotina e/ou o(s) operador(es) tratou(aram) - indicando se houve ou não procedimento automático – o texto de cada pessoa, em cada estágio do trabalho da Aplicação de Códigos.

a) Rotina de Verificação de códigos

Para a atividade de verificação de códigos, desenvolveu-se um aplicativo em cuja operação acontecia, em linhas gerais, o seguinte:

- liberado um lote de codificação, através do seu arquivo intermediário, era constituído o correspondente lote de verificação, que continha somente os registros das pessoas em que, pelo menos, numa das etapas, correção de caracteres e confirmação de códigos, houve a necessidade de intervenção do operador;

- no lote de verificação, era selecionada uma amostra de registros onde o verificador, utilizando a mesma rotina de trabalho do operador, com exceção dos mecanismos de correção e confirmação automáticos, fazia a nova aplicação de códigos;

- para cada registro da amostra, após o trabalho de verificação, eram acrescentadas as informações do arquivo intermediário, o texto após a correção de caracteres e o código aplicado pelo verificador;

- eram identificadas as situações de divergência entre os códigos aplicados pelo codificador e verificador, e calculada sua proporção na amostra;

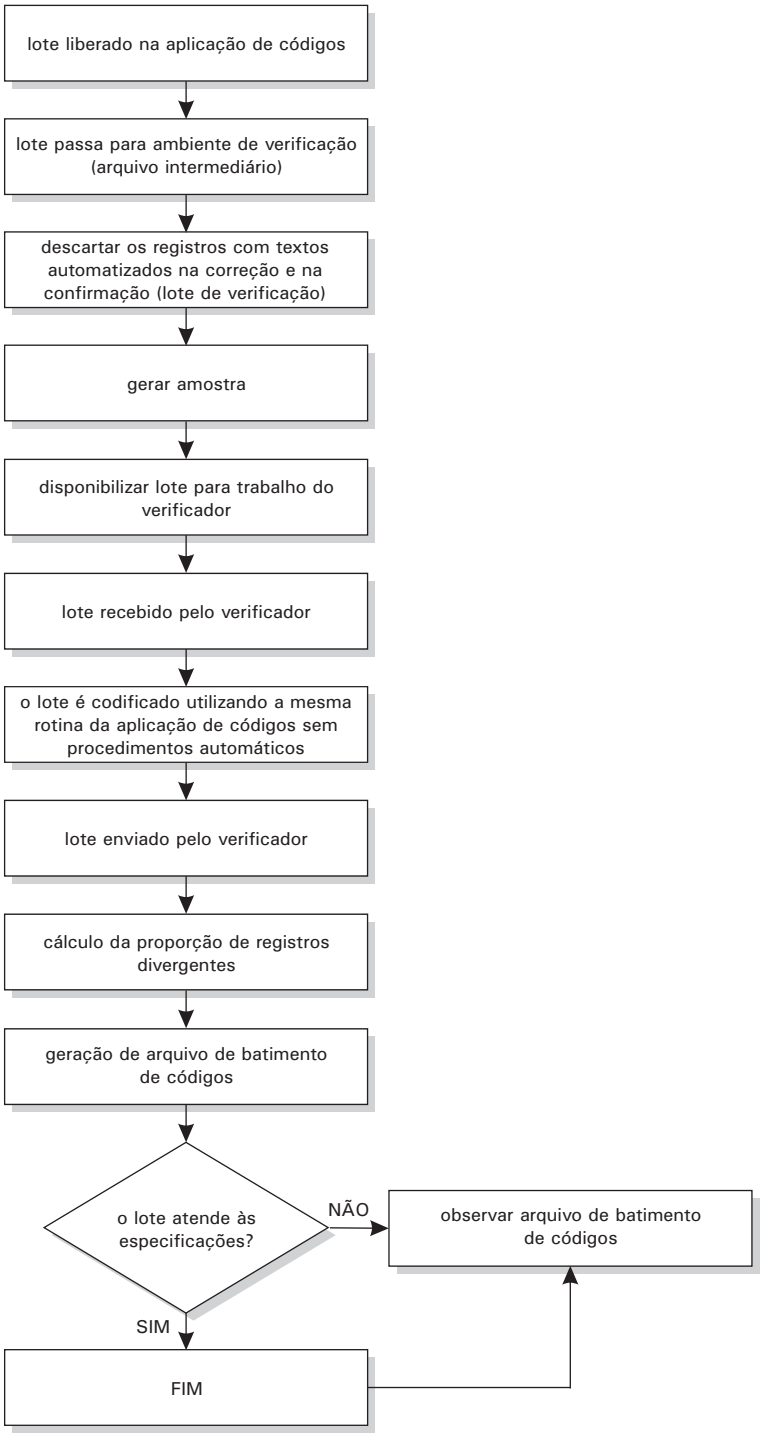
- constituía-se o arquivo de batimento, que continha, apenas, os registros onde ocorreram divergências de códigos, cujo conteúdo apresentava as mesmas informações do arquivo intermediário, pós trabalho de verificação; e

- dependendo do valor da proporção de códigos divergentes, era impresso, para investigação, o arquivo de batimento do lote de verificação.

O fluxo de trabalho, na etapa Verificação de Códigos, é apresentado a seguir.

Figura 12.4 – Fluxograma da verificação de códigos

Fluxograma da etapa de verificação



Por fim, o programa fazia a atualização dos seguintes relatórios de acompanhamento e avaliação da etapa Verificação de Códigos:

- gerenciamento de lotes 1 – apresenta o lote, segundo o quesito de verificação e o estágio de trabalho (codificado, disponível para verificação, amostrado, em trabalho, enviado e calculado);

- gerenciamento de lotes 2 – identifica o verificador, quesito de verificação e as datas de início e término do trabalho do lote;

- quantitativo de lotes – mostra a quantidade de lotes, segundo o estágio de trabalho e o quesito de verificação; e

- tamanhos dos lotes e amostras, segundo o quesito de verificação;

- proporção de códigos divergentes no lote; e

- relatório de lotes não classificados.

b) Critérios e Procedimentos

Levando-se em conta amostragem aleatória simples, a proporção Pvi de códigos divergentes em um determinado lote de verificação i foi estimada por

$$p_{vi} = \frac{d_{vi}}{n_i}$$

onde :

p_{vi} = proporção de códigos divergentes na amostra do lote i;

d_{vi} = número de códigos divergentes na amostra do lote i; e

n_i = número de códigos aplicados na amostra do lote i.

Definiu-se como "classificado", o lote de verificação cuja proporção máxima estimada de registros com códigos divergentes fosse da ordem de 5%. Ainda na amostragem aleatória simples sem reposição, para a estimação da proporção de 3% com um coeficiente de variação – CV – de 25% e o grau de confiança de 99%, tem-se um limite superior para o intervalo de aceitação de 5,25%, o que atendia, aproximadamente, ao limite estabelecido para julgamento do lote.

A especificação de um valor para a proporção de registros com códigos divergentes – ao redor de 5% – como parâmetro, que classificava um lote de verificação, foi feita arbitrariamente. Já a escolha dos valores de p (3%) e CV (25%), obedeceu à disponibilidade de pessoal e à relação entre estimativas dos tempos médios de trabalho dos operadores e verificadores. Mesmo nesse último caso, um certo grau de arbitrariedade aconteceu, pois seria possível outras combinações dos parâmetros p e CV que resultassem em tamanhos de amostra também adequados.

Para todos os lotes identificados como "não classificados", procedeu-se à investigação dos respectivos arquivos de batimento, objetivando esclarecer o que acarretou as divergências entre os códigos aplicados: se erro do operador e/ou do verificador. Assim, a(s) fonte(s) de erros identificada(s), associada(s) a outros critérios de acompanhamento do trabalho, determinavam quais as ações - alertas, retreinamento e implantação de novos procedimentos – seriam efetivadas junto aos operadores, orientadores e verificadores, tendo em vista melhorar a qualidade do trabalho.

Como foi visto, os pvi estimam, para todos os quesitos, as proporções de códigos divergentes nos lotes de verificação. No entanto, é necessário também fornecer uma indicação do resultado do trabalho de aplicação de códigos, tendo como referência os correspondentes lotes de codificação. Ou seja, devemos levar em conta para o cálculo da nova medida, a parcela dos códigos aplicados através de mecanismos automáticos de correção de caracteres e confirmação de códigos, simultaneamente.

Quando se toma somente os registros das pessoas em cujos códigos houve, necessariamente, a participação do operador, ou seja, sem o apoio de mecanismos automáticos – lotes de verificação – tem-se que a proporção de registros com códigos divergentes no respectivo lote de codificação será, no máximo, igual a estimada para o lote de verificação.

Assim, tendo em conta que na Aplicação de Códigos, os textos só são passíveis de tratamento automático após as situações superarem determinados limites de freqüência e serem avaliadas pelos orientadores, os códigos decorrentes dos procedimentos simultâneos de correção de textos e confirmação de códigos podem ser considerados corretos. Com base nessa hipótese, a proporção de códigos divergentes estimada para o lote de codificação i será

$$P_{ci} = p_{vi} \cdot N_i / M_i$$

onde N_i e M_i são, respectivamente, o tamanho do lote de verificação e de codificação, ou seja, o número de códigos obtidos sem mecanismos automáticos simultâneos e o número total de códigos atribuídos no lote.

A seguir, são apresentados gráficos para cada quesito de codificação que mostram as freqüências relativas – fr – dos lotes de codificação, segundo as proporções estimadas de códigos divergentes, relativas aos 215 lotes trabalhados. Deve-se registrar que essas proporções – eixo horizontal do gráfico – não representam equívocos de aplicação de códigos, mas tão somente as proporções estimadas de não coincidência entre códigos aplicados pelo operadores e verificadores.

Gráfico 12.5 - Quesito Religião
Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

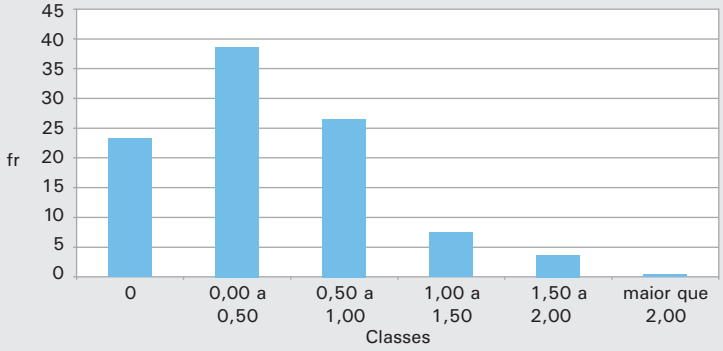


Gráfico 12.6 - Quesito Curso

Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

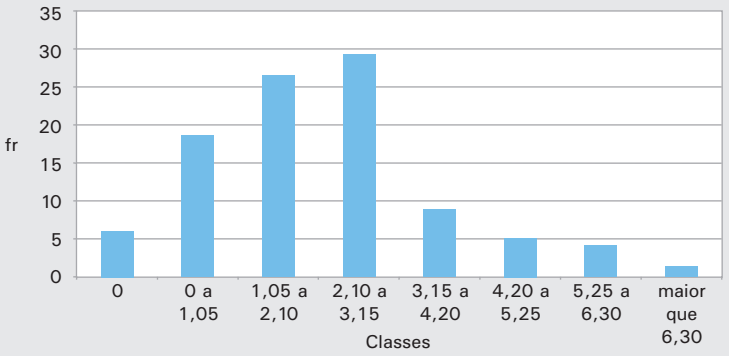


Gráfico 12.7 - Tema Migração - Quesito 4.21

Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

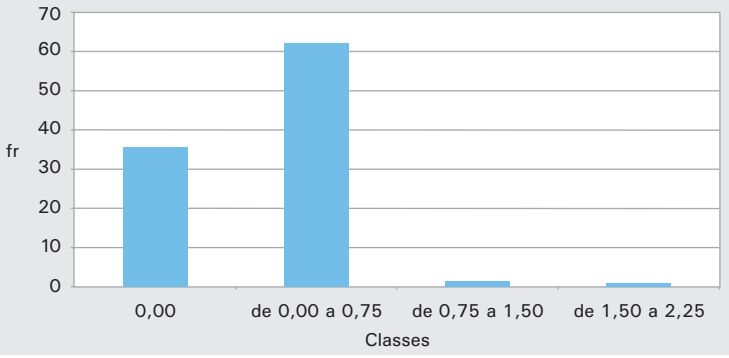


Gráfico 12.8 - Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

Tema Migração - Quesito 4.23

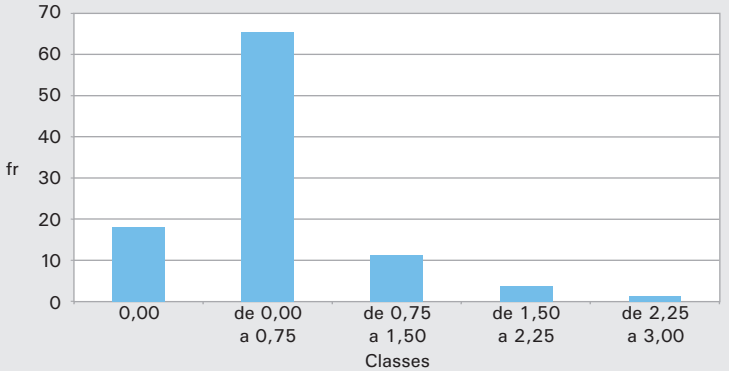


Gráfico 12.9 - Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

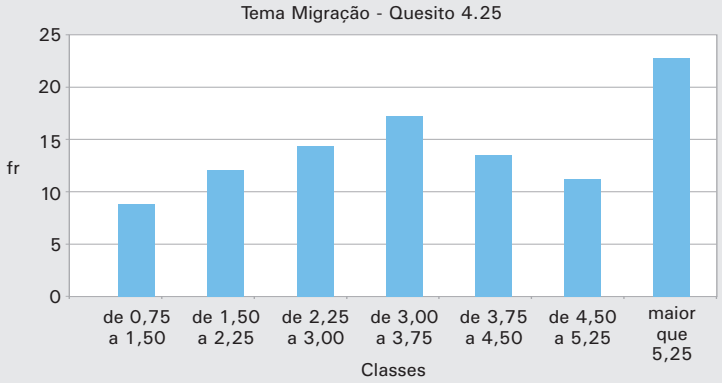


Gráfico 12.10 - Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

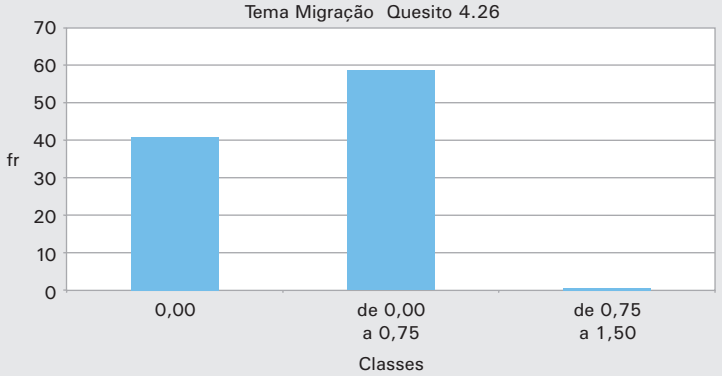


Gráfico 12.11 - Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

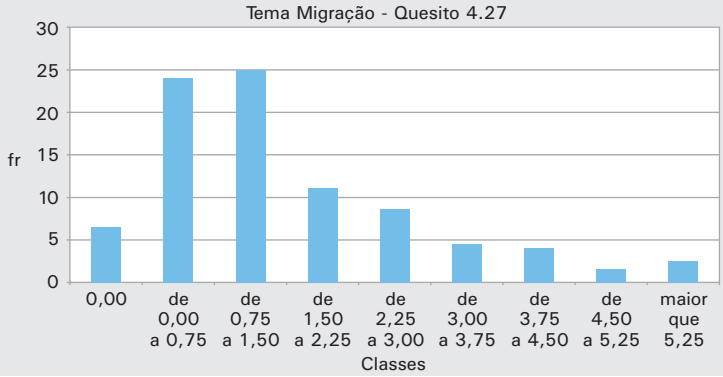


Gráfico 12.12 - Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes

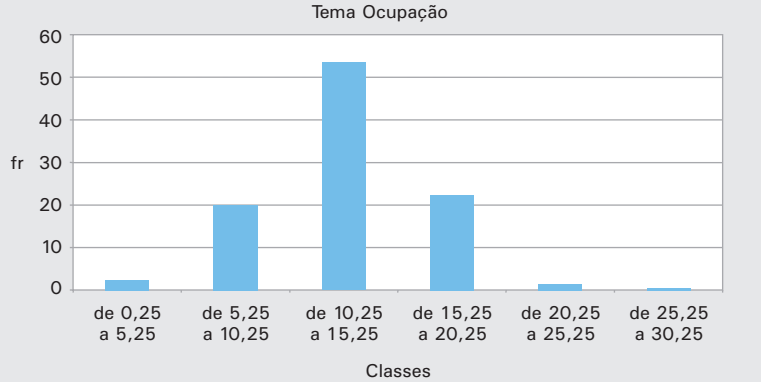
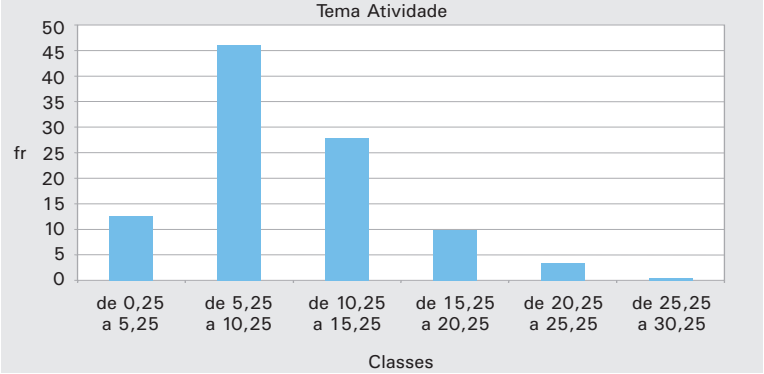


Gráfico 12.13 - Freqüências relativas - fr - dos lotes de codificação, segundo as proporções estimadas de códigos divergentes



12.5.2 Crítica e imputação dos dados

O planejamento dos trabalhos de apuração dos dados do Questionário da Amostra do Censo Demográfico 2000, com vistas à execução da crítica de incompatibilidades, partiu do princípio de que o sistema de detecção e imputação dos erros seria o DIA, ou seja, o mesmo utilizado no Censo Demográfico de 1991. Naquela oportunidade, as críticas de incompatibilidades foram, em quase sua totalidade, desenvolvidas de forma centralizada e apenas um conjunto mínimo, o correspondente às críticas entre registros, foi corrigido de forma manual, descentralizadamente, em pólos de apuração estruturados em vinte Unidades da Federação.

O objetivo traçado para o Censo 2000 era tratar toda a crítica dos dados do CD 1.02 de maneira centralizada, buscando a eliminação total do processo manual de correção, executado no Censo de 1991, mesmo sabendo das dificuldades do DIA em trabalhar com regras de crítica entre registros. O ponto de partida para consecução dessa meta foram os testes com os arquivos das informações da Segunda Prova Piloto.

Assim, inicialmente, foi escrito o conjunto de regras de crítica, descrevendo as inconsistências relativas às variáveis envolvidas nas críticas entre registros: sexo, relação com a pessoa responsável pelo domicílio, relação com a pessoa responsável pela família, número da família, idade do entrevistado, natureza da união e estado civil.

Durante esse processo de elaboração do conjunto de críticas, tendo em vista contornar a dificuldade do sistema DIA em trabalhar com regras de crítica entre registros, houve a necessidade de criação de um número razoável de variáveis auxiliares, a partir das variáveis originais do questionário.

Terminado o trabalho de elaboração das regras de crítica entre registros, verificou-se que o sistema DIA não obtinha êxito no procedimento de geração do conjunto completo de regras de crítica. Após várias tentativas, alternando-se as estratégias de criação das regras de incompatibilidades, permanecendo essa dificuldade, decidiu-se pelo abandono do sistema DIA para tratar as críticas entre registros.

Todo o desenvolvimento desse trabalho pode ser melhor conhecido, consultando-se o texto *Relatório sobre a experiência de tratamento automático de crítica entre registros, com vistas ao censo demográfico do ano 2000* (1999).

O abandono do sistema DIA determinou estudos para a busca de outra alternativa para a crítica entre registros dos dados do Questionário da Amostra. Os resultados desses estudos apontaram para duas vertentes metodológicas. A primeira, através de procedimentos automáticos de imputação, utilizando o sistema New Imputation Methodology – NIM para corrigir as inconsistências, considerando-se apenas os domicílios com até 8 moradores; a segunda, através do sistema Integrated Microcomputer Processing System – IMPS, atuando nos domicílios com 9 até 38 moradores, a partir de imputação manual.

O sistema NIM executa a correção dos dados, a partir da obtenção de domicílios doadores, selecionados dos estoques de domicílios sem nenhuma inconsistência nas informações, segundo o número de moradores. Nesse caso, foi necessária a formação de oito estratos de depuração.

Já o IMPS é um sistema de detecção de erros e, por essa razão, a correção deve ser realizada manualmente, em tela de microcomputador, através da análise das mensagens de erro. Técnicos treinados estiveram encarregados da execução desse trabalho, sendo os domicílios agrupados em um único estrato.

Ainda com relação a crítica entre registros, é necessário dizer que os cortes estabelecidos - 8 e 38 moradores – não se deram de forma arbitrária. Estabeleceu-se o critério, tendo em vista a preservação, a partir dos estoques de domicílios doadores , das características originais dos domicílios a serem corrigidos.

O critério de corte, domicílios com até oito moradores, levou em conta os resultados dos testes realizados com os arquivos da Segunda Prova Piloto. Em princípio, pretendia-se utilizar unicamente o NIM para tratar as críticas entre registros, envolvendo todos os domicílios. No entanto, durante esses testes, verificou-se que os resultados da imputação para domicílios com 12 e 16 moradores não foram satisfatórios. Aconteceu que, após a imputação, os domicílios viriam a ter suas características originais bastante alteradas. Houve, também, uma avaliação para os domicílios com 9 e 10 moradores, sendo considerados insuficientes os estoques de domicílios doadores para essas dimensões.

De acordo com o planejamento dos trabalhos para a crítica através do NIM, caso, em algum estrato, não se conseguisse um domicílio doador, o domicílio com erro era tratado pela alternativa IMPS.

12.5.2.1 Crítica entre registros

O objetivo deste item é descrever o procedimento utilizado para a crítica e imputação dos dados investigados, no Questionário da Amostra, sob a ótica da comparação entre registros de um mesmo domicílio. Além disso, são apresenta-

dos alguns resultados de uma exploração inicial feita nos arquivos de registro de execução dessa atividade, que indicam os tipos de análise que podem ser realizadas sobre o processo. Essa análise inicial oferece uma idéia da relevância das informações disponíveis para uma avaliação mais detalhada que tenha por objetivo entender o que aconteceu no Censo 2000 em termos de imputação, e ajudar no planejamento do próximo censo demográfico, ou de pesquisas correlatas.

Como já descrito em itens anteriores, na etapa de crítica de consistência do Censo Demográfico 2000, aproveitando a experiência adquirida no Censo de 1991, foi definida a utilização do programa DIA (Detección e Imputación Automática). O DIA trabalha muito bem no que se refere à consistência de variáveis dentro do mesmo registro (seja de domicílios ou pessoas), mas tem a limitação de não possuir ferramentas para verificar as regras entre registros (isto é, entre pessoas distintas, ou entre uma variável de domicílio e variáveis de pessoas).

A utilização do sistema DIA na crítica entre registros do Conjunto Universo foi possível em função do pequeno número de variáveis a serem criticadas e da estratégia adotada, que foi a de construir um único registro com as informações de todas as pessoas que seriam objeto dessa crítica. Esse procedimento está descrito no item 12.3.2.4, que trata da apuração do Conjunto Universo.

Para enfrentar este problema, buscou-se o NIM – Nearest-neighbour Imputation Methodology, do Statistics Canada, na época somente um protótipo, cujo nome era New Imputation Methodology, mas que apresentava uma série de vantagens técnicas e operacionais, a saber:

- a imputação é totalmente automática, bastando definir as regras de consistência (isto é, não é necessário definir também as regras de imputação);
- a imputação é baseada em um único doador por questionário falhado (domicílio);
- segue a filosofia de Feleggi e Holt, no sentido de alterar o menor número de variáveis do questionário;
- trabalha com arquivos comuns de tipo texto; e,
- o software foi cedido ao IBGE "em aberto", isto é, com os programas-fonte.

Por outro lado, devido à própria característica de imputação dos domicílios com erro a partir de doadores (domicílios "bons"), essa imputação baseia-se na busca de um doador que tenha a menor distância, no sentido de maior semelhança do conteúdo das variáveis. Por esse motivo, o NIM trabalha separando os domicílios por estratos de número de pessoas (domicílios com 1 pessoa, domicílios com 2 pessoas, domicílios com 3 pessoas, ..., domicílios com n pessoas), posto que este procedimento padroniza a utilização de doadores de mesmo tamanho (mesmo número de pessoas) para o domicílio com erro.

Ao processar um estrato qualquer, é de se esperar que este tenha um número mínimo de doadores tal que possa assegurar uma imputação de boa qualidade (distâncias pequenas, não utilização de mesmo doador, etc.). Isto não é possível de se garantir para os estratos com um maior número de pessoas (domicílios com 12 pessoas, por exemplo). Para estratos maiores, é possível não conseguir doador, impossibilitando o uso do método.

Nesse sentido, limitou-se a oito o número de estratos a serem trabalhados pelo NIM (desde domicílios com 1 pessoa, até domicílios com 8 pessoas). Os domicílios com mais de 8 pessoas foram tratados por um outro método, o IMPS

(Integrated Microcomputer Processing System), fornecido pelo Census Bureau, dos Estados Unidos. O IMPS necessita que sejam estipuladas tanto as regras de consistência como as de imputação, ou que a imputação seja operada manualmente por pessoal qualificado (os questionários com erro são apresentados na tela com seus respectivos erros, e os operadores definem ações de imputação para "limpá-los").

O ideal seria que todas as imputações pudessem ser definidas de maneira automática, sem intervenção humana, mas isto não foi de todo possível, dada a complexidade das relações de estrutura do questionário (relações entre pessoas). Isto obrigou que se dispusesse de um conjunto de operadores que processassem os questionários com mais de 8 pessoas.

Descrição dos procedimentos

Resumidamente e, em termos lógicos, o processo consistia na formatação do arquivo para extração das variáveis que seriam trabalhadas (somente aquelas envolvidas com a estrutura do domicílio). Em seguida, este arquivo era separado em 9 arquivos, um para cada estrato a ser tratado pelo NIM, e um único estrato com os domicílios com mais de 8 pessoas, a ser tratado pelo IMPS. Esses procedimentos foram realizados independentemente em cada um dos 67 lotes de questionários formados especificamente para essa etapa de crítica entre registros. Os lotes foram definidos, considerando os seguintes critérios:

- possuir, pelo menos, 1000 questionários em cada estrato, com o objetivo de formar massa crítica suficiente para a alocação de registros doadores similares aos que falhassem em alguma regra de crítica. esse valor foi definido com base nos testes de uso do NIM realizados pelo Statistics Canada. Por outro lado, aumentar a quantidade mínima de questionários implicaria a necessidade de agregação de Unidades da Federação menos populosas em um único lote. Também não seria possível, na maioria dos casos, obter pelo menos um lote de domicílios com situação "rural" por Unidade da Federação;

- considerar a separação por Unidade da Federação, desde que o primeiro critério continuasse válido;

- considerar a separação por situação do domicílio (urbano ou rural), desde que também o primeiro critério continuasse válido;

- ter um limite máximo de registros para um estrato (aproximadamente 20 000), de acordo com a capacidade e tempo de processamento do sistema, porém, não foi definido o tamanho máximo para um lote;

- obter o menor número possível de lotes, porém respeitando os critérios propostos.

A composição final dos lotes usados para a execução dessa etapa de apuração está apresentado no Quadro 12.3.

Quadro 12.3 - Composição dos lotes de questionários para a realização da etapa de crítica e imputação entre registros e em número de domicílios particulares ocupados na amostra

Unidades da Federação	Lote	Mesorregião	Número de domicílios particulares ocupados na amostra
11	1	lote único	42 966
12, 14 e 16	1	lote único	38 027
13	1	lote único	63 694
15	1	Urbana	96 317
	2	Rural	48 789
17	1	lote único	42 648
21	1	Urbana	85 656
	2	Rural	64 042
22	1	Urbana	54 851
	2	Rural	39 344
23	1	2 e 3 (urbana)	82 445
	2	1, 4 , 5, 6 , 7 (urbana)	56 852
	3	Rural	60 877
24	1	lote único	92 118
25	1	Urbana	77 603
	2	Rural	39 292
26	1	1, 2, 4 (urbana)	45 488
	2	3 (urbana)	36 537
	3	5 (urbana)	84 741
	4	Rural	57 226
27	1	Urbana	51 925
	2	Rural	25 591
28	1	lote único	54 825
29	1	1, 2, 3 (urbana)	55 675
	2	4, 5 (urbana)	111 854
	3	6, 7 (urbana)	76 689
	4	1 a 5 (rural)	75 205
	5	6 e 7 (rural)	57 054
31	1	1, 2, 3, 4 (urbana)	57 329
	2	5, 9 (urbana)	77 556
	3	6, 8, 10 (urbana)	34 533
	4	7 (exceto município 6200) (urbana)	86 121
	5	7 (município 6200 (BH)) (urbana)	63 308
	6	11, 12 (urbana)	71 208
	7	1, 2, 3, 4, 5, 6, 9, 10 (rural)	82 207
	8	7, 8, 11, 12 (rural)	40 561
32	1	lote único	98 067
33	1	1, 2, 3, 4, 5 (urbana e rural)	89 926
	2	6 (micros 14, 15, 16, 17,18 - município 456 a 3302) (urbana)	91 511
	3	6 (micro 18 – município 4557)	180 173
	4	6 (micro 18 – demais municípios – urbana) e rural meso 6	75 899
35	1	1, 8 (urbana)	76 351
	2	2, 9, 14 (urbana)	85 102
	3	3, 4, 5 (urbana)	81 757
	4	6, 13 (urbana)	86 205
	5	7 (urbana)	88 892
	6	10, 11, 12 (urbana)	127 101
	7	15 (micros 57, 58, 59 -urbana)	83 139
	8	15 (micros 60, 62, 63) (urbana)	87 806
	9	15 (micro 61 - exceto município 50308) (urbana)	65 925
	10	15 (micro 61 (município 50308 - áreas 1, 2, 3)	84 902
	11	15 (micro 61 - município 50308 - áreas 4, 5) (urbana)	93 459
	12	15 (micro 61 - município 50308 - áreas 6, 7, 8) (urbana)	104 579
	13	Rural	88 436
41	1	1, 2, 3, 4, 6 (urbana)	134 828
	2	5, 7, 8, 9, 10 (urbana)	122 985
	3	Rural	74 959
42	1	1, 2, 3 (urbana e rural)	92 233
	2	4, 5, 6 (urbana e rural)	99 508
43	1	1, 6 (urbana e rural)	107 017
	2	2, 3, 4, 7 (urbana e rural)	108 293
	3	5 (urbana e rural)	146 080
50	1	lote único	68 316
51	1	lote único	85 402
52	1	1, 2, 4, 5 (urbana e rural)	90 590
	2	3 (urbana e rural)	82 866
53	1	lote único	53 940

Fonte: IBGE, Censo Demográfico 2000, Sistema de Indicadores Gerenciais da Coleta.

Os estratos de 1 a 8 pessoas eram processados pelo NIM, que produzia como saída um arquivo imputado, de mesmo formato que o arquivo de entrada e um arquivo de controle ou ocorrências, aonde eram registradas, para cada domicílio, as seguintes informações (entre outras):

- regras falhadas, com a identificação da regra;
- variáveis imputadas, com a identificação da variável, o valor antigo e o valor novo.

O estrato de 9 e mais pessoas era tratado pelo IMPS, cujo processo também tinha como saída dois arquivos, o arquivo imputado e o arquivo de controle. Por possuir um componente manual (operado por pessoas), seu arquivo de controle não continha os registros de variáveis imputadas, somente o de regras falhadas.

Estes arquivos de controle são os arquivos usados para a exploração aqui descrita. É importante ressaltar que os arquivos de controle de imputação, recolhidos no processo do IMPS, não contêm informações sobre as variáveis imputadas, o que é uma limitação nas análises de imputação. Isso significa que, na maioria dos casos, o universo de análise é o dos domicílios com até 8 pessoas. Também não se encontram nestes arquivos as informações sobre a imputação dos domicílios em outros processos posteriores, como o DIA e a imputação dos rendimentos.

Tratamento das omissões da variável "espécie do domicílio"

Embora a variável "espécie do domicílio" estivesse relacionada, mais diretamente, ao Aplicativo Domicílio, fez parte também do conjunto de tabelas de decisão lógica do NIM. Em algumas dessas tabelas, a crítica relacionava a categoria da variável "espécie do domicílio" com as variáveis "relação com a pessoa responsável pelo domicílio" e "número da família. Por esse motivo, era preciso que houvesse informação para essa variável.

Para corrigir as omissões de informação da variável "espécie do domicílio", foi utilizada a mesma estratégia definida para os dados do Conjunto Universo, descrita no item 12.3.2.4. A omissão era detectada e corrigida durante o processo de formação dos estratos a serem submetidos ao NIM.

As variáveis envolvidas na estrutura do domicílio e que foram objeto dessa etapa de crítica entre registro são:

Quadro 12.4 - Variáveis envolvidas no tratamento das omissões

V0201	Espécie do domicílio (variável do domicílio, somente para consulta, nunca imputada)
V0401	Sexo da pessoa
V0402	Relação com o responsável pelo domicílio
V0403	Relação com o responsável pela família
V0404	Número da família
V0436	Vive em companhia de cônjuge
V0437	Natureza da última união
V0438	Estado civil
V4007	Faixa de idade
V4667	Indicador de fecundidade

Para dar uma idéia do tipo de crítica que foi realizada, envolvendo as variáveis acima, seguem exemplos descritivos das relações verificadas:

- duas pessoas classificadas como responsável pelo domicílio, V0402, (ou pela família, V0403) e cônjuge do responsável pelo domicílio (ou da família) devem ter informações iguais para a variável estado civil (V0438);
- duas pessoas classificadas como responsável pelo domicílio (ou pela família) e cônjuge do responsável pelo domicílio (ou da família) devem ter informações diferentes para a variável sexo (V0401);
- pessoa menor de 10 anos de idade não pode ter informações nas variáveis que investigam as características de nupcialidade (V0435, V0437 e V0438);
- o número de famílias em um domicílio deve ser igual ao número de pessoas classificadas como responsáveis pela famílias;
- todo domicílio tem que ter apenas um responsável pelo domicílio;
- nenhuma variável pode estar em branco.

Exploração de resultados

Pode-se dizer que, para o total Brasil, 28% dos domicílios tiveram, pelo menos, um erro na consistência de sua estrutura, seja este erro apontado pelo NIM ou IMPS[4]. A Tabela 12.20 traz estes percentuais de erro por UF, os quais também são mostrados no Gráfico 12.14 a seguir. Neste, pode-se ver que a UF com maior percentual de erros foi o Amazonas, com 40%, e a menor, o Rio Grande do Sul, com cerca de 19%.

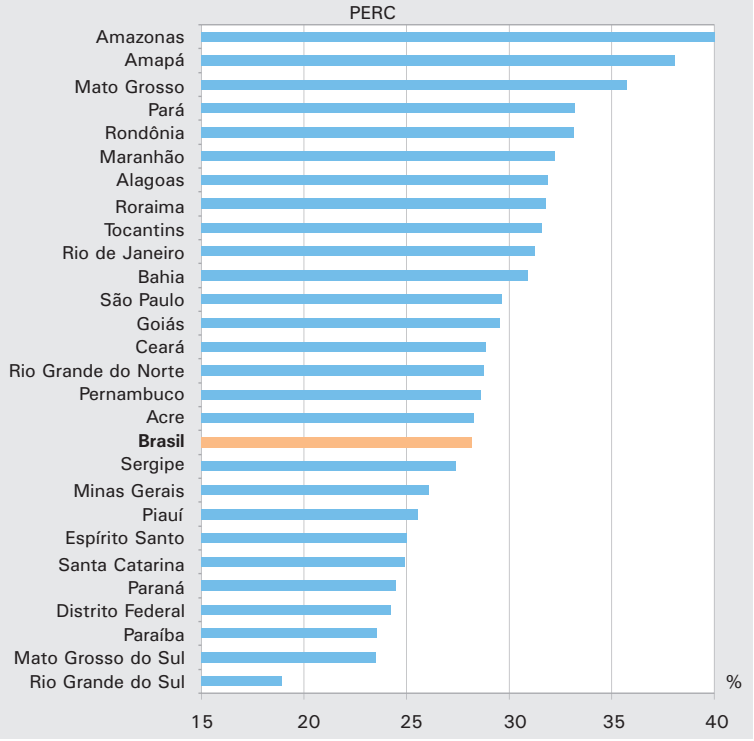
Tabela 12.20 - Número total de domicílios e de domicílios com erro, segundo as Unidades da Federação

Unidades da Federação	Número total de domicílios	Número de domicílios com erro	Percentual de domicílios com erro
Brasil	**5 304 711**	**1 494 397**	**28,17**
Rondônia	43 293	14 350	33,15
Acre	16 818	4 757	28,29
Amazonas	63 970	25 623	40,05
Roraima	9 857	3 133	31,78
Pará	145 992	48 514	33,23
Amapá	11 821	4 502	38,08
Tocantins	43 043	13 599	31,59
Maranhão	150 441	48 536	32,26
Piauí	94 534	24 156	25,55
Ceará	201 143	58 078	28,87
Rio Grande do Norte	92 673	26 669	28,78
Paraíba	117 577	27 698	23,56
Pernambuco	225 649	64 559	28,61
Alagoas	77 896	24 848	31,9
Sergipe	55 161	15 114	27,4
Bahia	378 907	117 216	30,94
Minas Gerais	615 101	160 622	26,11
Espírito Santo	98 820	24 716	25,01
Rio de Janeiro	442 976	138 408	31,25
São Paulo	1 137 154	337 148	29,65
Paraná	336 151	82 235	24,46
Santa Catarina	193 633	48 261	24,92
Rio Grande do Sul	365 827	69 217	18,92
Mato Grosso do Sul	69 401	16 332	23,53
Mato Grosso	86 946	31 053	35,72
Goiás	175 132	51 777	29,56
Distrito Federal	54 795	13 276	24,23

Fonte: IBGE, Censo Demográfico 2000.

[4] Aqui não foram computados os possíveis erros encontrados posteriormente pelos outros processos de consistência (DIA, imputação dos rendimentos, etc.).

Gráfico 12.14 - Percentual de domicílios com erro, por Unidade da Federação



A Tabela 12.21 mostra os totais de erros, segundo os estratos (número de pessoas por domicílio), também mostrados no Gráfico 12.15 abaixo. Os valores variam desde cerca de 13% para o estrato 1 até cerca de 54% para o estrato de 9 ou mais pessoas (processado pelo IMPS).

Tabela 12.21 - Número total de domicílios e de domicílios com erro, segundo os estratos

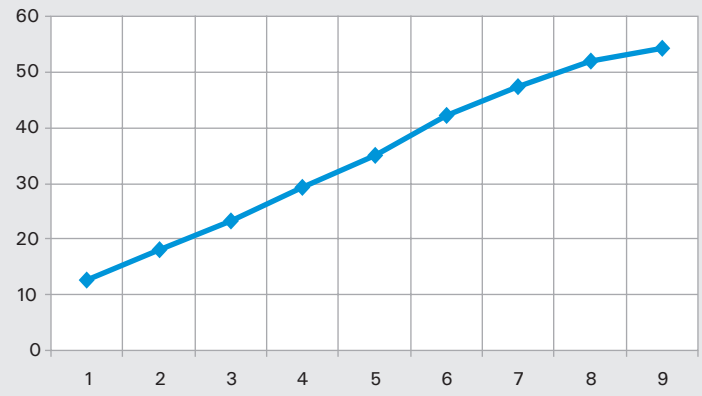
Estrato	Número total de domicílios	Número de domicílios com erro	Percentual de domicílios com erro
Total	**5 304 711**	**1 494 397**	**28,17**
1 pessoa	497 673	63 571	12,77
2 pessoas	873 061	157 554	18,05
3 pessoas	1 134 064	264 849	23,35
4 pessoas	1 210 216	353 623	29,22
5 pessoas	759 139	266 754	35,14
6 pessoas	388 930	164 640	42,33
7 pessoas	192 825	91 115	47,25
8 pessoas	108 764	56 367	51,83
9 ou mais pessoas	140 039	75 905	54,20

Fonte: IBGE, Censo Demográfico 2000.

Nota: Ocorreram 19 casos de registros pertencentes aos estratos de 1 até 8 pessoas que foram detectados como possuidores de algum erro ou inconsistência, mas para os quais não foi possível encontrar doador pelo sistema NIM. Esses 19 casos foram a valiados e imputados pelo sistema IMPS com os 75 905 casos do estrato de 9 ou mais pessoas

Neste caso, é razoável esperar que a percentagem de erros seja proporcional ao tamanho do estrato (quanto maior o estrato, maior será a percentagem de erros). A importância desta informação e, em especial, a relativa ao estrato de 9 ou mais pessoas, é que ela indica a quantidade de domicílios com erro que se pode esperar para serem corrigidos manualmente, caso esta estratégia seja, de novo, usada Ou seja, neste último estrato, um pouco mais da metade dos questionários tem algum erro de estrutura e, na hora de definir os lotes de produção, deve-se levar em conta que somente 50% dos questionários poderiam ser considerados como doadores em potencial.

Gráfico 12.15 - Totais de erros, segundo os estratos



A análise de erros (imputações) por domicílio é apresentada na Tabela 12.22, e mostra que 60% dos domicílios teve somente uma variável imputada (esta percentagem teria sido muito maior se houvesse uso de uma estratégia de imputação automática do número da família (ver mais adiante). Por curiosidade, pode-se ver que, ao final da lista, aparece um domicílio com 43 variáveis imputadas.

Tabela 12.22 - Número de domicílios, por total de erros por domicílio

(continua)

Número de erros por domicílio	Freqüência absoluta	Freqüência relativa (%)	Freqüência relativa acumulada (%)
Total	**1 418 473**	**100,00**	**100,00**
1	848 007	59,78	59,78
2	275 299	19,41	79,19
3	124 473	8,78	87,97
4	69 938	4,93	92,90
5	37 404	2,64	95,53
6	23 212	1,64	97,17
7	13 621	0,96	98,13
8	8 912	0,63	98,76
9	5 576	0,39	99,15
10	3 815	0,27	99,42
11	2 543	0,18	99,60
12	1 785	0,13	99,73
13	1 205	0,08	99,81
14	796	0,06	99,87
15	588	0,04	99,91

Tabela 12.22 - Número de domicílios, por total de erros por domicílio

(conclusão)

Número de erros por domicílio	Freqüência absoluta	Freqüência relativa (%)	Freqüência relativa acumulada (%)
16	432	0,03	99,94
17	251	0,02	99,96
18	154	0,01	99,97
19	117	0,01	99,98
20	94	0,01	99,98
21	63	0,00	99,99
22	49	0,00	99,99
23	32	0,00	99,99
24	25	0,00	99,99
25	26	0,00	100,00
26	9	0,00	100,00
27	13	0,00	100,00
28	9	0,00	100,00
29	4	0,00	100,00
30	4	0,00	100,00
31	5	0,00	100,00
32	3	0,00	100,00
33	1	0,00	100,00
34	1	0,00	100,00
35	5	0,00	100,00
37	1	0,00	100,00
43	1	0,00	100,00

Fonte: IBGE, Censo Demográfico 2000.

Evidentemente que o número de variáveis imputadas deve ser proporcional ao estrato, ou melhor, ao número de pessoas por domicílio. A Tabela 12.23 apresenta a razão de erros por pessoa, e pode-se ver que quase 58% dos domicílios têm menos de ½ erro por pessoa, e 73% têm até ½ erro por pessoa.

Tabela 12.23 - Número de domicílios, segundo a razão de erros por pessoa

Razão de erros por pessoa	Número de domicílios com erros	Percentual de domicílios com erros
Total	**1 418 473**	**100,00**
0. Menos de 0,5 erros por pessoa	819 726	57,79
1. 0,5 erros por pessoa	222 321	15,67
2. Mais de 0,5 até 1 erro por pessoa	285 141	20,10
3. Mais de 1 erro até 2 erros por pessoa	77 721	5,48
4. Mais de 2 erros por pessoa	13 564	0,96

Fonte: IBGE, Censo Demográfico 2000.

A Tabela 12.24 mostra a percentagem de imputação por variável. Vê-se que quase 1/3 (33%) do total de imputações corresponde à variável V0404 "número da família". A outra grande freqüência concentra-se na V0403 "relação com o responsável pela família", com cerca de 24%. Isto demonstra que não foi fácil responder a estas duas perguntas e que, provavelmente, seja necessário rever estes conceitos e/ou estas perguntas, nas próximas pesquisas.

Tabela 12.24 - Freqüência de imputação, segundo a variável

Variável	Número de imputações	Percentual de imputações
Total	**2 742 154**	**100,00**
V0401	247 174	9,01
V0402	227 268	8,29
V0403	647 312	23,61
V0404	907 712	33,10
V0436	248 873	9,08
V0437	218 168	7,96
V0438	123 565	4,51
V4007	5 367	0,20
V4667	116 715	4,26

Fonte: IBGE, Censo Demográfico 2000.

No caso da variável V0404, a Tabela 12.25 mostra a distribuição de freqüências dos valores errados (antes da imputação). A categoria 10 significa "valor em branco", e responde por quase 68% dos valores errados. Em outras palavras, 68% das pessoas que tiveram a variável V0404 imputada, o foram porque simplesmente estava em branco (não informada).

Tabela 12.25 - Freqüência dos valores errados da variável V0404, segundo seus valores antes da imputação

Valor da variável V0404 antes da imputação	Freqüência	Freqüência relativa (%)
Total	**907 712**	**100,00**
0	3 953	0,44
1	35 556	3,92
2	208 248	22,94
3	30 706	3,38
4	7 047	0,78
5	2 760	0,30
6	1 539	0,17
7	975	0,11
8	336	0,04
9	143	0,02
Branco	616 449	67,91

Fonte: IBGE, Censo Demográfico 2000.

Mais curioso, ainda, é que na Tabela 12.26 – matriz de imputação da variável V0404 – pode-se ver que a maioria dos valores em branco (606.128 de 616.449, ou 98%) foram imputados para o valor 1 (família 1). Isto era espera-

do, visto que 93% dos domicílios têm somente uma família. O importante é que esta variável poderia ser objeto de uma pré-imputação, transformando, por exemplo, todos os brancos no valor 1. Isto facilitaria os trabalhos de imputação, porque, possivelmente, estes domicílios seriam considerados corretos e aumentariam a massa de domicílios doadores em cada estrato, aumentando, assim, a chance de termos menores distâncias e uma qualidade maior na imputação.

Tabela 12.26 - Matriz de imputação da variável V0404: valores antes e depois da imputação

Valor da variável V0404 antes da imputação	Valor da variável V0404 depois da imputação					
	Total	0	1	2	3	4
Total	**3 953**	-	**3 918**	**35**	-	-
0	35 556	5 000	-	29 473	1 077	6
1	208 248	20	206 437	-	1 777	14
2	30 706	4	20 177	10 510	-	15
3	7 047	5	6 537	401	104	-
4	2 760	-	2 672	84	3	1
5	1 539	2	1 468	67	2	-
6	975	-	924	50	1	-
7	336	-	314	22	-	-
8	143	-	133	10	-	-
9	616 449	1 056	606 128	8 994	269	2
Branco	907 712	6 087	848 708	49 646	3 233	38

Fonte: IBGE, Censo Demográfico 2000.

Por outro lado, no caso da variável V0403 "relação com o responsável pela família", a Tabela 12.27 apresenta a distribuição de freqüências dos valores errados. A categoria 0 significa "valor em branco" e responde por 17% desses valores. O interessante aqui é a percentagem de casos imputados quando o entrevistado se declarou responsável pela família (categoria 1, com 48%). Isto significa que quase metade das imputações nesta variável foram causadas porque alguém respondeu que a pessoa era responsável pela família. Pelos conceitos do programa de imputação isto foi considerado errado.

Tabela 12.27 - Freqüência dos valores errados da variável V0403, segundo seus valores antes da imputação

Valor da variável V0403 antes da imputação	Freqüência	Freqüência relativa (%)
Total	**647 312**	**100,00**
1	314 325	48,56
2	81 259	12,55
3	78 823	12,18
4	9 615	1,49
5	4 916	0,76
6	6 785	1,05
7	29 688	4,59
8	5 257	0,81
9	870	0,13
10	1 698	0,26
11	1 432	0,22
12	486	0,08
Branco	112 158	17,33

Fonte: IBGE, Censo Demográfico 2000.

Ao verificar-se a Tabela 12.28, que mostra a matriz de imputação desta variável, pode-se ver que a grande maioria dos "supostos responsáveis" foi imputada para a categoria 2 (cônjuge), ou seja, 270.183 de 314.325. Possivelmente, isto pode significar que estas pessoas se considerariam dividindo a responsabilidade da família, o que seria uma razão a mais para uma revisão profunda destes conceitos.

Tabela 12.28 - Matriz de imputação da variável V0403: valores antes e depois da imputação

Valor da variável V0403 antes da imputação	Valor da variável V0403 depois da imputação												
	Total	1	2	3	4	5	6	7	8	9	10	11	12
Total	**647 312**	**107 146**	**334 709**	**146 069**	**7 360**	**21 359**	**7 091**	**14 650**	**1 861**	**538**	**825**	**21**	**5 683**
1	314 325	-	270 183	26 852	3 395	748	2 482	4 596	711	264	327	1	4 766
2	81 259	21 569	-	52 703	1 458	431	843	3 992	154	29	77	2	1
3	78 823	42 763	15 935	-	680	15 754	1 269	2 143	177	19	59	15	9
4	9 615	5 476	3 427	640	-	30	16	17	1	-	1	-	7
5	4 916	1 374	544	2 545	8	-	86	334	17	1	6	-	1
6	6 785	3 063	1 626	1 420	22	120	-	436	62	13	19	-	4
7	29 688	12 466	13 095	2 875	265	346	394	-	176	17	39	1	14
8	5 257	1 601	1 569	1 106	27	73	118	605	-	31	60	-	67
9	870	276	180	149	4	11	55	66	33	-	1	-	95
10	1 698	826	278	396	2	20	27	80	53	4	-	-	12
11	1 432	47	55	1 096	1	65	53	73	22	2	18	-	-
12	486	419	12	27	1	1	18	5	3	-	-	-	-
Branco	112 158	17 266	27 805	56 260	1 497	3 760	1 730	2 303	452	158	218	2	707

Fonte: IBGE, Censo Demográfico 2000.

Seguindo com a análise por variáveis, a variável V0402 “relação com o responsável pelo domicílio” não apresenta maiores peculiaridades, como mostra Tabela 12.29, a não ser a maior freqüência de imputados por falta de informação (representada pela categoria 0) em 25%.

Tabela 12.29 - Freqüência dos valores errados da variável V0402, segundo seus valores antes da imputação

Valor da variável V0402 antes da imputação	Freqüência	Freqüência relativa (%)
Total	**227 268**	**100,00**
1	15 526	6,83
2	32 172	14,16
3	33 702	14,83
4	2 569	1,13
5	34 248	15,07
6	8 708	3,83
7	23 149	10,19
8	11 976	5,27
9	2 244	0,99
10	2 308	1,02
11	2 249	0,99
12	527	0,23
Branco	57 890	25,47

Fonte: IBGE, Censo Demográfico 2000.

A análise da variável V0401 "sexo", cuja distribuição de freqüência está na Tabela 12.30, mostra que 68% das imputações são devidas à falta de informação (categoria 0): quase 170 mil respostas em branco na variável "sexo" que, em princípio, é de fácil preenchimento. Este número não representa muito no total de pessoas da amostra (um pouco mais de 20 milhões), mas é sintomático de que é preciso mais atenção para as próximas pesquisas.

Tabela 12.30 - Freqüência dos valores errados da variável V0401, segundo seus valores antes da imputação

Valor da variável V0401 antes da imputação	Freqüência	Freqüência relativa (%)
Total	**247 174**	**100,00**
1	35 151	14,22
2	42 627	17,25
Branco	169 396	68,53

Fonte: IBGE, Censo Demográfico 2000.

Pela Tabela 12.23, vê-se que a variável V0436 "vive em companhia de cônjuge ou companheiro" apresentou cerca de 9% das imputações realizadas. A Tabela 12.31 apresenta a freqüência absoluta e relativa dos valores possíveis da variável antes da imputação e a Tabela 12.32 apresenta, em valores percentuais, a matriz com a distribuição dos valores possíveis da variável antes e depois do processo de imputação.

Tabela 12.31 - Freqüência dos valores errados da variável V0436, segundo seus valores antes da imputação

Valor da variável V0436 antes da imputação	Freqüência	Freqüência relativa (%)
Total	**248 873**	**100,00**
Sim	151 391	60,83
Não, mas viveu	30 645	12,31
Nunca viveu	9 919	3,99
Branco	56 918	22,87

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.32 - Matriz de imputação da variável V0436: valores percentuais antes e depois da imputação

Valor da variável V0436 antes da imputação	Valor das variáveis V0436 depois da imputação				
	Total	Sim	Não, mas viveu	Nunca viveu	Branco
Total	**100,00**	**16,57**	**44,64**	**38,68**	**0,10**
Sim	60,83	-	41,00	19,75	0,08
Não, mas viveu	12,31	9,01	-	3,30	0,01
Nunca viveu	3,99	2,88	1,09	-	0,02
Branco	22,87	4,69	2,55	15,63	-

Fonte: IBGE, Censo Demográfico 2000.

Na Tabela 12.31, destaca-se o fato de 60% das imputações terem acontecido em registros de pessoas que responderam o código 1 – Sim, ou seja, informaram que viviam em companhia de cônjuge ou companheiro, mas que as informações das demais pessoas do domicílio não confirmavam essa informação. O mais importante ainda, apresentado na Tabela 12.31, é o fato de que do total das imputações realizadas nessa variável, 40% referem-se à troca do código 1 – Sim para o código 2 – Não, mas já viveu. Isso pode indicar algum viés na coleta da informação e carece de estudos para uma melhor compreensão da ocorrência.

O processamento através do sistema IMPS

No conjunto de críticas que seriam executadas por esse sistema, verificou-se que algumas eram determinísticas e que poderiam ser processadas através de programa. O documento interno *Procedimentos a serem implementados para a execução da crítica (NIM e IMPS), questionário da amostra/censo 2000* (2001), especifica as situações que deveriam ser corrigidas antes de se executar o IMPS.

Os procedimentos de correção determinística ocorreram em situações muito específicas, não esgotando, portanto, todos os acertos nessas variáveis, abrangendo as seguintes variáveis: sexo; relação com a pessoa responsável pelo domicílio e pela família; número da família; e a variável auxiliar faixa de idade, apenas na categoria inválido.

Para a correção determinística relativa à variável "sexo", nos casos de omissão do registro para as pessoas de dez anos e mais de idade, foram consideradas as especificações da variável auxiliar "marca da fecundidade". Já para as pessoas menores de dez anos de idade, foi desenvolvido um programa de imputação probabilística. Informações detalhadas sobre esse sistema podem ser conhecidas através do documento *Manual de crítica: sistema IMPS* (2002).

Ao submeter o lote de trabalho ao IMPS, o operador observava na tela do microcomputador as mensagens de erro, identificadas de acordo com o conjunto de críticas elaboradas para o tratamento das inconsistências entre registros. As correções só eram realizadas após a análise criteriosa de todas as variáveis envolvidas nas críticas entre registros, disponibilizadas para o operador. Alguns desses critérios, por razões de confiabilidade na variável, orientavam o operador a só alterar as variáveis faixa de idade e a marca da fecundidade em casos estritamente necessários. Por razões operacionais, na execução do sistema IMPS os domicílios com mais de 38 moradores não puderam ser criticados em tela do microcomputador. Esses domicílios, um de Pernambuco com 42 moradores e o outro de Sergipe com 43 moradores, foram criticados e corrigidos separadamente.

A equipe que esteve envolvida no trabalho de correção através do sistema IMPS era composta por treze técnicos. Em média eram corrigidos 600 questionários por dia. Os 67 lotes que foram criados e submetidos ao IMPS totalizavam 140.029 questionários com erros. Os lotes foram submetidos ao sistema de crítica a partir do dia 10/04/2002, tendo-se concluído toda a correção no dia 22/05/2002.

Após analisar cada mensagem de erro e solucionar o problema, questionário a questionário, o operador acionava um comando que submetia todo lote novamente ao sistema de crítica para certificar-se de que não apresentava mais erros.

Tendo em vista a crítica e imputação da variável "sexo", os totais desta variável – homens e mulheres – de cada domicílio, foram recalculados.

12.5.2.2 Crítica intra-registros

Antes da descrição dos diversos aplicativos utilizados para detecção e depuração das inconsistências dos dados do Questionário da Amostra, é necessário ter conhecimento de algumas questões.

Em primeiro lugar, devido a complexidade dos temas investigados no Questionário da Amostra e, consequentemente, do número de variáveis que o compõe, só foi possível implementar o processo de imputação utilizando o sistema DIA, a partir da criação de um conjunto bastante significativo de variáveis auxiliares, antes de se iniciar a crítica em cada aplicativo.

Considerando-se que essas variáveis auxiliares foram criadas, inicialmente, com base em arquivos ainda não corrigidos, houve a necessidade, assim que se concluía a imputação de um aplicativo, que os totais fossem gerados novamente, a partir dos arquivos depurados.

Por outro lado, a análise do resultado da imputação, realizada durante a execução dos testes dos aplicativos com os dados do censo, para o Espírito Santo e Rondônia, apontou algumas situações de inconsistência. Essas situações não decorreram de equívocos na execução do DIA. Ou seja, na sua essência, o processo de imputação funcionou de acordo com os pressupostos do sistema, gerando registros que obedeciam ao estabelecido pelas regras de crítica. O que se detectou foi uma perda na coerência entre algumas poucas variáveis. Por esse motivo, decidiu-se implementar alguns procedimentos prévios à execução dos aplicativos do DIA – pré-DIA, com vistas a preparar essas variáveis para serem imputadas.

Assim, no bloco de fecundidade, foram implementados procedimentos pré-DIA em alguns registros envolvendo as variáveis: filhos tidos nascidos vivos, filhos tidos nascidos vivos que estavam vivos e filhos tidos nascidos mortos. No bloco de instrução, foram previamente tratados alguns registros relativos às variáveis “conclusão do curso no qual estudou” e “código do curso mais elevado concluído”, bem como, no bloco de mão-de-obra, alguns dos que envolviam a variável “posição na ocupação”. A variável auxiliar “idade calculada do entrevistado”, obtida através de um algoritmo próprio, também passou por esse tipo de ajuste prévio. Todos os procedimentos pré-DIA podem ser consultados no documento Definições necessárias à implementação da apuração centralizada dos dados referentes ao questionário da amostra (CD 1.02) no censo demográfico de 2000 (2003).

A definição do conjunto de aplicativos para a imputação dos dados da amostra obedeceu, em primeiro lugar, à conveniência da composição em um só aplicativo das críticas relativas a blocos distintos do questionário, buscando-se estabelecer um processo de imputação mais criterioso. Em função disso, foi implementada a imputação conjunta de parte dos blocos de instrução e fecundidade.

O segundo aspecto levado em conta para o estabelecimento dos aplicativos, foi garantir ao DIA a possibilidade de geração do conjunto completo de críticas, tendo em conta o elevado número de variáveis a serem depuradas no Questionário da Amostra. Desse modo, os aplicativos foram definidos, levando-se em conta a quantidade de variáveis investigadas em cada parte do questionário, mas também o número de regras de crítica a serem implementadas. Isso determinou a partição dos aplicativos de instrução e fecundidade, de mão-de-obra, de migração e de fecundidade.

Além disso, também foi importante a definição da ordem de execução dos aplicativos. A experiência do passado e a análise dos resultados da imputação, durante a fase de testes, mostrou a importância de determinadas variáveis serem imputadas somente após a correção prévia de outras. Dentro dessa lógica, certas variáveis já corrigidas tornar-se-iam fixas durante a execução dos aplicativos seguintes.

Já no caso dos aplicativos de fecundidade, além de todas as razões apontadas, as divisões I, II e III ocorreram, tendo em vista a necessidade de se estabelecer uma ordenação na seqüência em que as variáveis deveriam ser imputadas, a partir do grau de confiança associado às variáveis envolvidas.

As variáveis do bloco de fecundidade foram imputadas, em parte, nos aplicativos Instrução e Fecundidade da Pessoa Responsável pela Família ou Individual em Domicílio Coletivo e Instrução e Fecundidade das Demais Pessoas da Família e, em parte, nos aplicativos Fecundidade I, II e III.

A ordem de execução dos aplicativos no sistema DIA foi a seguinte: Instrução e Fecundidade da Pessoa Responsável pela Família ou Individual em Domicílio Coletivo; Instrução e Fecundidade das Demais Pessoas da Família; Mão-de-Obra I e II; Migração I e II; Deficiência; e Fecundidade I, II e III.

Após os lotes terem sido imputados através desse aplicativos, alguns procedimentos de conclusão foram necessários – procedimentos pós-DIA. Como exemplo, temos o decorrente da estratégia utilizada durante a execução do sistema, ou seja, a necessidade de submeter os lotes a um tradutor para os códigos de ocupação e atividade, recompondo o banco original codificado a cinco dígitos, visto que o sistema de crítica foi executado com os códigos a três dígitos. Outros procedimentos pós-Dia, decorrentes do processo analítico da variável Rendimento, também foram implementados.

Os procedimentos desenvolvidos no pós-DIA, podem ser conhecidos, consultando-se as Referências, no final do capítulo.

a) Aplicativo Domicílio

As variáveis do questionário imputadas por este aplicativo foram:

- V0202 – Tipo;

- V0203 – Total de cômodos;

- V0204 – Cômodos servindo de dormitório;

- V0205 – Condição de ocupação do domicílio;

- V0206 – Condição de ocupação do terreno do domicílio;

- V0207 – Forma de abastecimento de água;

- V0208 – Canalização da água;

- V0209 – Número de banheiros;

- V0210 – Existência de sanitário;

- V0211 – Tipo de escoadouro;

- V0212 – Destino do lixo;

- V0213 – Iluminação elétrica;

- V0214 – Existência de rádio;

- V0215 – Existência de geladeira ou freezer;

- V0216 – Existência de videocassete;

- V0217 – Existência de máquina de lavar roupa;

- V0218 – Existência de forno de microondas;

- V0219 – Existência de linha telefônica instalada;

- V0220 – Existência de microcomputador;

- V0221 – Quantidade de televisores;

- V0222 – Quantidade de automóveis para uso particular; e

- V0223 – Quantidade de aparelhos de ar condicionado.

Durante a execução do aplicativo, as variáveis V0201 "espécie do domicílio" e V7100 "total de pessoas no domicílio foram declaradas fixas, servindo de referência para a correção das demais variáveis.

Todas as variáveis foram imputadas probabilisticamente, de acordo com as suas respectivas distribuições marginais dadas pelas freqüências dos registros não suspeitos e o método proporcional.

A variável V7100 foi utilizada, especificamente, para corrigir as incompatibilidades envolvendo a variável V0204. As variáveis V0205 e V0209 foram tratadas com pesos, respectivamente, 2 e 1, enquanto às demais foi atribuído o peso médio igual a 5.

b) Aplicativo Instrução e Fecundidade da Pessoa Responsável pela Família ou Individual em Domicílio Coletivo

Neste aplicativo foram corrigidas as inconsistências relativas às variáveis do CD 1.02, de todo o bloco de instrução e algumas relativas ao bloco de fecundidade, somente para a "pessoa responsável pela família" ou classificada como "individual em domicílio coletivo".

As variáveis criadas e imputadas, foram:

- V4620 – Total de filhos tidos nascidos vivos;

-V0463 – Total de filhos tidos nascidos vivos e que estavam vivos;

- V4640 – Indicadora do preenchimento da V0464;

- V4660 – Indicadora do preenchimento da V0466;

- V4670 – Total de filhos nascidos mortos;

- V4075 – Idade calculada do entrevistado; e

- V4654 – Idade calculada do último filho nascido vivo.

As seguintes variáveis do questionário foram imputadas:

- V0428 – Sabe ler e escrever;

- V0429 – Freqüência à escola ou creche;

-- V0430 – Curso que freqüenta;

- V0431 – Série que freqüenta;

- V0432 – Curso mais elevado que freqüentou no qual concluiu pelo menos uma série;

- V0433 – Última série concluída com aprovação;

- V0434 – Conclusão do Curso no Qual Estudou; e

- V4353 – Código do curso mais elevado concluído;

As variáveis auxiliares, criadas e consideradas fixas, foram as seguintes:

- V4060 – Grupo de idade do cônjuge;

- V4301 – Grupo de anos de estudo;

- V4403 – Indicadora do estado conjugal da mulher;

- V4567 – Grupo de idade qüinqüenal do entrevistado; e

- V4453 – Código, a três dígitos, da variável ocupação.

Foram as seguintes, as variáveis tratadas em etapas anteriores, portanto fixas:

- V4007 – Faixa de idade do entrevistado;

- V0401 – Sexo;

- V0402 – Relação com a pessoa responsável pelo domicílio; e

- V0403 – Relação com a pessoa responsável pela família.

Para o bloco de fecundidade, foram corrigidas as inconsistências relativas ao "total de filhos tidos nascidos vivos", ao "total de filhos tidos nascidos vivos e que estavam vivos" e ao "total de filhos nascidos mortos", sem a intenção de se corrigir as suas parcelas – homens e mulheres – que foram tratadas em aplicativo posterior.

O procedimento de imputação implementado para esse aplicativo, centrou-se na utilização de distribuição conjunta dos registros não suspeitos e no método proporcional, para as seguintes variáveis:

- V0428 conjunta com a variável V4075;

- V4075 conjunta com as variáveis V0403 e V4060;

- V4620 conjunta com as variáveis V0463 e V4567;

- V0463 conjunta com as variáveis V4301 e V4567;

- V4654 conjunta com as variáveis V4620 e V4567; e

- V4670 conjunta com as variáveis V4567 e V4301.

As variáveis V0430 e V0431 foram imputadas no DIA, de acordo com a situação, através de método determinístico ou probabilístico. A imputação determinística passou a ser uma estratégia em virtude da ocorrência de eventuais inconsistências entre os registros da série e do grau, envolvendo níveis de ensino diferentes: primário, ginásio e colegial, clássico, científico etc. – quando o correto seria primeiro grau ou segundo grau. Nos demais casos, em que a modalidade foi a imputação probabilística, utilizou-se a distribuição marginal dos registros não suspeitos e o método proporcional.

Para todas as demais variáveis tratadas nesse aplicativo, a imputação aconteceu de acordo com a distribuição marginal dos registros não suspeitos, através do método proporcional. A ponderação para a variável V4075, devido ao maior grau de confiabilidade em relação às demais variáveis, foi 1 e, para todas as outras, utilizou-se o peso médio 5.

c) Aplicativo Instrução e Fecundidade das Demais Pessoas da Família

Neste aplicativo, foram corrigidas as inconsistências relativas às variáveis do CD 1.02 de todo o bloco de instrução e algumas do bloco de fecundidade, das demais pessoas da família, exceto a "pessoa responsável" ou o "individual em domicílio coletivo".

Os procedimentos utilizados para a imputação foram, quase todos, os mesmos utilizados e descritos no aplicativo anterior. Desse modo, as variáveis tratadas em etapas anteriores e as imputadas, do questionário ou criadas, foram exatamente as mesmas.

Da lista de variáveis criadas e consideradas fixas do aplicativo anterior, foi excluída a V4060 e introduzida a V4061"grupo de idade da pessoa responsável pela família" ou "individual em domicílio coletivo". Assim, essa nova variável passou a ser considerada, através da distribuição conjunta, para fazer a imputação da variável V4075 "idade calculada do entrevistado".

d) Aplicativos de Mão-de-Obra

Nestes aplicativos, foram tratadas as inconsistências relativas ao bloco de mão-de-obra do CD 1.02. Houve a necessidade de implementar dois aplicativos, Mão-de-Obra I e Mão-de-Obra II.

Aplicativo Mão-de-Obra I

Neste aplicativo, as variáveis criadas e imputadas foram:

- V4453 – Código, a três dígitos, da variável ocupação;

- V4510 – Indicadora de rendimento no trabalho principal;

- V4520 – Indicadora de rendimento nos demais trabalhos;

- V4570 – Indicadora de rendimento de aposentadoria ou pensão;

- V4580 – Indicadora de rendimento de aluguel;

- V4590 – Indicadora de rendimento de pensão alimentícia, mesada ou doação recebida de não-morador;

- V4600 – Indicadora de rendimento de renda mínima, bolsa-escola, seguro-desemprego etc.; e

- V4610 – Indicadora de outros rendimentos.

As variáveis imputadas do questionário foram:

- V0439 – Trabalho remunerado na semana de referência;

-V0440 – Trabalho remunerado do qual estava temporariamente afastado na semana de referência;

- V0441 – Ajuda sem remuneração na semana de referência no trabalho exercido por pessoa moradora do domicílio ou trabalho sem remuneração como aprendiz ou estagiário;

- V0442 – Ajuda sem remuneração na semana de referência no trabalho exercido por pessoa moradora do domicílio empregada em atividade de cultivo, extração vegetal etc.;

- V0443 – Trabalho na semana de referência em atividade de cultivo, extração vegetal0 etc. destinados à alimentação de pessoas moradoras do domicílio;

- V0444 – Número de trabalhos na semana de referência;

- V0447 – Posição na ocupação;

- V0448 – Empregado pelo regime jurídico dos funcionários públicos ou como militar;

- V0449 – Número de empregados;

- V0450 – Contribuinte de instituto de previdência oficial;

- V0453 – Número de horas trabalhadas por semana no trabalho principal;

- V0454 – Número de horas trabalhadas por semana nos demais trabalhos;

- V0455 – Providência para conseguir trabalho; e

- V0456 – Aposentado de instituto de previdência oficial.

Foram as seguintes, as variáveis auxiliares criadas e consideradas fixas:

- V4301 – Grupo de anos de estudo;

- V4302 – Grupo de anos de estudo, por sexo;

- V4471 – Grupo de Posição na Ocupação;

- V4568 – Indicadora de grupos de idade a partir de 10 anos.

Foram consideradas fixas, as seguintes variáveis tratadas em etapas anteriores:

- V4075 – Idade calculada do entrevistado;

- V0428 – Sabe ler e escrever;

- V0429 – Freqüência à escola ou creche;

- V0430 – Curso que freqüenta;

- V4353 – Código do Curso Mais Elevado Concluído; e

- V0402 – Relação com a pessoa responsável pelo domicílio.

À exceção da variável V0447, que teve peso igual a 1, todas as demais do Aplicativo Mão-de-Obra I foram imputadas com peso médio 5.

Para a imputação das variáveis, V0439 a V0443 e V0455 foram utilizados os registros não suspeitos e o método proporcional da distribuição conjunta com a variável V4568. Já a estratégia de imputação das variáveis V4580, V4590, V4600 e V4610 considerou a distribuição conjunta com a variável V4301. Por fim, para a variável V4453 foi levada em conta a distribuição conjunta com as variáveis V4302 e V4471.

As demais variáveis do aplicativo foram imputadas de acordo com a distribuição marginal dos registros não suspeitos, através do método proporcional.

Aplicativo Mão-de-Obra II

Neste aplicativo, foram corrigidas as inconsistências na variável V4463 "Código a Três Dígitos da Variável Atividade". Para isso, as variáveis do Aplicativo Mão-de-Obra I, que serviram de referência, permanecendo fixas, foram a V4453 e V0448. Como ponderação, atribuiu-se o peso médio 5 e a variável V4463 foi imputada de acordo com a distribuição marginal dos registros não suspeitos, através do método proporcional.

e) Aplicativos de Migração

Nestes aplicativos, foram corrigidas as inconsistências do bloco de migração do CD 1.02 para todas as pessoas do domicílio.

Todas as variáveis, considerando-se os dois aplicativos implementados, foram imputadas de acordo com a distribuição marginal dos registros não suspeitos, através do método proporcional. Com exceção da variável V4254, que teve ponderação 3, todas as demais variáveis tiveram peso médio 5.

Aplicativo Migração I

Para este aplicativo, as variáveis criadas e imputadas foram as seguintes:

- V4201 – Indicadora de preenchimento da V0420.

- V4254 – Indicadora da relação de igualdade entre a V4251 e a V0102;

- V4264 – Recodificação da V4261;. e

- V4270 – Recodificação da V4276.

As variáveis imputadas, originais do Questionário da Amostra, foram:

- V0415 – Sempre morou neste município;

- V0416 – Tempo de moradia sem interrupção neste município;

- V0417 – Nasceu neste município;

- V0418 – Nasceu nesta Unidade da Federação;

- V0419 – Nacionalidade;

- V4210 – Código da Unidade da Federação ou país estrangeiro de nascimento;

- V0422 – Tempo de moradia sem interrupção na Unidade da Federação;

- V4230 – Código da Unidade da Federação ou país estrangeiro de residência anterior; e

- V0424 – Local de Residência em 31 de Julho de 1995.

As variáveis auxiliares criadas e consideradas fixas foram:

- V1023 – Indicadora de UF/município totalmente urbano, onde foi realizado o censo.

- V1030 – Indicadora de município de Brasília;

- V4454 – Indicadora de preenchimento da ocupação;

- V4231 – Indicadora da relação entre a V4230 e V0102; e

- V4252 – Indicadora de município totalmente urbano.

As variáveis tratadas em etapas anteriores e, portanto, consideradas fixas, foram:

- V4075 – Idade calculada do entrevistado; e

- V0429 – Freqüência à escola ou creche.

Aplicativo Migração II

No aplicativo Migração II, foram corrigidas todas as inconsistências do bloco de migração de todas as pessoas do domicílio.

Houve a criação da variável V4251 "Indicadora da UF/país estrangeiro", e as variáveis do questionário, objeto de imputação, foram as seguintes:

- V0420 – Ano em que fixou residência no Brasil; e

- V4260 – Código da UF ou país estrangeiro de residência em 31/7/95.

As variáveis tratadas em etapas anteriores e, portanto, fixas, foram:

- V4075 – Idade calculada do entrevistado;

- V0419 – Nacionalidade;

- V4201 – Indicadora de preenchimento da V0420;

- V0422 – Tempo de moradia sem interrupção na Unidade da Federação;

- V0102 – Código da UF onde foi realizado o censo;

- V4254 – Indicadora da relação de igualdade entre a V4251 e a V0102 ; e

- V4264 – Recodificação da V4261.

f) Aplicativo Deficiência

Neste aplicativo, foram corrigidas, para todos os moradores do domicílio, as inconsistências entre as variáveis do bloco de deficiência e as incompatibilidades entre essas variáveis e a variável "ocupação". As omissões de registro para as variáveis "cor ou raça" e "religião ou culto" foram corrigidas no DIA através de imputação determinística, atribuindo-lhes o código de informação ignorada.

As variáveis imputadas, pertencentes ao questionário, foram:

- V0408 – Cor ou raça;

- V4090 – Código da religião ou culto;

- V0410 – Deficiência mental permanente;

- V0411 – Autoavaliação da capacidade de enxergar;

- V0412 – Autoavaliação da capacidade de ouvir;

- V0413 – Autoavaliação da capacidade de caminhar / subir escadas; e

- V0414 – Existência de deficiências

As variáveis tratadas em etapas anteriores e, portanto, fixas, foram as seguintes:

- V0402 – Relação com a pessoa responsável pelo domicílio;

- V0403 – Relação com a pessoa responsável pela família;

- V4007 – Faixa de idade do entrevistado; e

- V4453 – Código, a três dígitos, da variável ocupação.

As variáveis tiveram o peso médio 5 e utilizou-se para a imputação a distribuição marginal dos registros não suspeitos através do método proporcional.

g) Aplicativos de Fecundidade

Através dos Aplicativos Fecundidade I, II e III, concluiu-se a correção de todo o bloco de fecundidade, iniciada pelos dois Aplicativos de Instrução e Fecundidade, descritos anteriormente. O peso médio 5 foi atribuído às variáveis nesses três aplicativos. Parte da imputação do aplicativo Fecundidade III - para mulheres com mais de 19 filhos - foi feita através de imputação manual, pois o DIA não conseguiu criar o conjunto completo de críticas

Aplicativo Fecundidade I

Neste aplicativo, foram corrigidas as inconsistências em relação às características da fecundidade das mulheres cuja variável V4620 "total de filhos tidos nascidos vivos" estava em branco ou preenchida com a informação 0 (zero) até 8. Foram também imputadas, as parcelas, homens e mulheres, da variável V4670 "total de filhos tidos nascidos mortos" para todas as mulheres.

Foram as seguintes as variáveis imputadas, investigadas no questionário:

- V4621 – Filhos tidos nascidos vivos, homens;

- V4622 – Filhos tidos nascidos vivos, mulheres;

- V4631 – Filhos tidos nascidos vivos que estavam vivos, homens;

- V4632 – Filhos tidos nascidos vivos que estavam vivos, mulheres;

- V0464 – Sexo do último filho nascido vivo;

- V4671 – Filhos tidos nascidos mortos, homens; e

- V4672 – Filhos tidos nascidos mortos, mulheres.

As variáveis tratadas em etapas anteriores, portanto fixas, foram:

- V4620 – Total de filhos tidos nascidos vivos;

- V0463 – Total de filhos tidos nascidos vivos que estavam vivos; e

- V4670 – Total de filhos tidos nascidos mortos.

Como estratégia, algumas variáveis foram imputadas de acordo com a distribuição conjunta, através dos registros não suspeitos, e o método proporcional, a saber:

- as variáveis V4671 e V4672, através da variável V4670;

- as variáveis V4621 e V4622, através da variável V4620; e

- as variáveis V4631 e V4632, através da variável V0463.

As demais variáveis desse aplicativo foram imputadas de acordo com a distribuição marginal dos registros não suspeitos e o método proporcional.

Aplicativo Fecundidade II

No Aplicativo Fecundidade II, foram corrigidas as inconsistências em relação às variáveis V4631 e V4632, das mulheres cuja variável V4620 assumiu um valor de 9 a 31.

As variáveis imputadas, presentes no questionário, foram:

- V4631 – Filhos tidos nascidos vivos que estavam vivos, homens; e

- V4632 – Filhos tidos nascidos vivos que estavam vivos, mulheres.

As variáveis tratadas em etapas anteriores, portanto fixas, foram:

- V4620 – Total de filhos tidos nascidos vivos; e

- V0463 – Total de filhos tidos nascidos vivos que estavam vivos.

Utilizou-se a distribuição conjunta dos registros não suspeitos e o método proporcional para imputar as variáveis V4631 e V4632, por meio da V0463

Aplicativo Fecundidade III

Neste aplicativo, foram corrigidas as inconsistências em relação às características da fecundidade relativas às variáveis V4621, V4622 e V0464, das mulheres cuja variável V4620 estava preenchida com valores de 9 a 19.

As variáveis V4621 e V4622 foram imputadas de acordo com a variável V4620, através da distribuição conjunta dos registros não suspeitos e o método proporcional. As demais variáveis deste aplicativo foram imputadas de acordo com a distribuição marginal dos registros não suspeitos e o método proporcional.

A imputação manual dessas variáveis, para as mulheres com mais de 19 filhos, foi feita, em microcomputador, por dois técnicos, trabalhando as informações reunidas em nível de Brasil. Quando havia informação para o total de filhos e para um dos sexos, o número de filhos para de outro sexo foi imputado pela diferença. Quando havia somente registro para uma das categorias - total de filhos, filhos (as) homens ou mulheres – as demais categorias eram imputadas tomando como base os dados de outra mulher – doador – que apresentasse informações corretas para as três categorias e, também, coincidência com a categoria declarada para o registro a ser imputado.

h) Críticas das Variáveis de Rendimento

Na fase de crítica de consistência dos dados de rendimento das pessoas pesquisadas na amostra do Censo Demográfico 2000, buscou-se verificar para cada tipo de rendimento (trabalho principal, demais trabalhos, de aposentadoria ou pensão, etc.) a ocorrência de registros que, por alguma incorreção, pudessem distorcer os resultados obtidos para cada um deles isoladamente e, consequentemente, para o rendimento total.

Para auxiliar nesta verificação, além da geração de alguns indicadores por município, foram selecionados, por unidade da federação, os valores extremos para cada tipo de rendimento e, para o do trabalho principal, por posição na ocupação e categoria do emprego, associados a outras características das pessoas e dos domicílios de residência, a fim de se ter uma primeira visão de inconsistências entre características. O exame mais aprofundado dos casos de inconsistências entre características por meio de consulta às imagens dos questionários revelou dois tipos de problemas. No primeiro, registros de rendimento de trabalho mostravam-se inconsistentes em decorrência de classificações incorretas das características do trabalho ocorridas na fase de coleta. Foram verificadas classificações de posição na ocupação incompatíveis com as ocupações e atividades econômicas. Para estas situações foram definidos procedimentos de critica automatizada restritos aos registros de trabalho.

Verificaram-se ainda, em muito menor quantidade que as observadas durante os trabalhos de crítica dos resultados do universo, principalmente, situações de reconhecimento como valores, de traços delimitadores das quadrículas desenhadas para registro dos algarismos e, ainda, de sombras decorrentes de registros feitos nos versos de algumas das folhas do questionário. Para estes casos, foram definidos procedimentos automatizados de eliminação dos registros indevidos do rendimento, considerando que, na maior parte dos casos, eram de valores formados por seqüência do dígito 1, combinado ou não com o dígito 7; e, para definir as situações em que deveria haver imputação do rendimento, foram considerados outros registros indicativos de sua existência ou não.

Concluída esta etapa, avaliaram-se, de forma global, os resultados e, de forma pontual, os valores extremos, inclusive com exame das imagens dos questionários. Os valores com fortes inconsistências foram ignorados para serem tratados na etapa de crítica e imputação dos rendimentos.

12.5.2.3 Análise do processo de crítica e imputação

No Censo Demográfico 2000, os procedimentos de crítica e imputação dos dados foram constantemente monitorados a fim de evitar alterações na estrutura da informação. Vários foram os instrumentos utilizados com esse objetivo, como as tabelas (conjunto de indicadores que são calculados na execu-

ção do programa de crítica e imputação), a análise demográfica, os estudos de população e o controle das alterações nas respostas originais constantes do questionário. Este texto trata sobre esse último instrumento.

a) análise dos relatórios do sistema DIA

Os relatórios gerados pelo sistema DIA permitem que sejam realizadas diversas análises, entre elas conhecer o nível de imputação de cada uma das variáveis tratadas no sistema. No caso do questionário da amostra, foram tratadas, através dos diversos aplicativos desenvolvidos no DIA, 22 variáveis do bloco de características do domicílio e 65 do bloco de características dos moradores, considerando-se os diversos temas envolvidos. No total, foram 20 274 412 registros de pessoas e 5 304 711 registros de domicílios.

No bloco de domicílios, a variável com maior índice de imputação (4,15% do total de registros, correspondentes a 220 240 ocorrências) foi a que se refere ao número de cômodos servindo de dormitório. Outras 7 variáveis tiveram imputação entre 1,16% e 1,89% do total de registros e as 14 restantes tiveram imputação em menos de 1,00% dos registros.

No bloco de características de moradores, observou-se a seguinte distribuição nos percentuais de imputação:

Tabela 12.33 - Número de variáveis imputadas, segundo o percentual de imputação

Percentual de imputação	Número de variáveis
Total	**65**
De 0,01 a 1,00	27
De 1,01 a 2,00	27
De 2,01 a 3,50	8
De 3,51 a 5,13	0
De 5,14 a 7,47	3

Fonte: IBGE, Censo Demográfico 2000.

As 3 variáveis com maiores percentuais de imputação (5,14%, 6,09% e 7,47%, correspondentes a 1 042 499, 1 234 566 e 1 515 441 ocorrências, respectivamente) foram as que se referem, pela ordem, à atividade principal em que a pessoa trabalhava na semana de referência, ao município onde trabalha ou estuda – para quem trabalhava ou estudava em município diferente do de residência – e ao curso mais elevado no qual concluiu pelo menos uma série – para quem não freqüentava escola mas já havia freqüentado.

Ao longo deste capítulo, serão detalhados os diversos procedimentos de análise e validação das imputações. Por outro lado, observa-se que 40,0% das variáveis tiveram menos de 1,00% dos registros imputados e outras 40,0% sofreram imputação entre 1,01% e 2,00% dos registros, significando uma boa qualidade geral de preenchimento dos questionários em campo.

Outra análise do processo de crítica e imputação dos Resultados da Amostra decorrentes da aplicação do sistema DIA foi realizada para cada um dos aplicativos da crítica intra-registros, tendo como orientação diretrizes bastante semelhantes às estabelecidas no item 12.3.2.6, que tratou da mesma tarefa para os Resultados do Universo. Os elementos para realização desse trabalho estão

detalhados no Plano de análise da correção automática e elementos de apoio para a análise da formação do Lote siatema DIA, CD 1.02 – questionário da amostra 2002; assim, apontaremos aqui somente os pontos distintos do processo.

No tocante à análise em nível do lote, foram considerados os mesmos indicadores E e F para detecção das situações suspeitas; no entanto, os limites de tolerância estabelecidos foram, respectivamente, 5% e 10%, para os aplicativos de fecundidade – "Fecundidade I, II e III"- e 12% e 50% para os demais aplicativos.

Da mesma forma, a análise em nível de município, considerou também os indicadores E, Ij, Maxi Dj(i), Maxi Aj(i), Tj e Bj , levando em conta, no entanto, os seguintes limites de tolerância, respectivamente:

- indicador E – 12%;

- indicador Ij, – 2% para fecundidade e 3% para os demais aplicativos; e

- indicadores Maxi Dj(i), Maxi Aj(i), Tj e Bj – 2% para os aplicativos de fecundidade e, respectivamente 3%, 3%, 10% e 5% para os demais aplicativos.

b) análise da consistência da imputação e validação dos resultados

É importante fazer uma primeira observação com respeito à Análise do processo de Crítica e Imputação e à dificuldade em separá-la do item validação dos resultados.

Para avaliar os efeitos do processo de imputação na informação obtida, muitas vezes, esta era comparada com a proveniente de outras fontes e com a série histórica com o intuito de descobrir se os indicadores eram compatíveis com o comportamento esperado. Este foi o mesmo recurso utilizado posteriormente na validação dos resultados. Houve grande dificuldade em tratar estes aspectos separadamente já que, muitas vezes, as divergências com a tendência esperada levaram a descobrir problemas no processo analisado.

Portanto, neste item serão abordadas tanto as análises dos processos de crítica e imputação quanto a validação dos resultados obtidos.

Foram realizadas análises para algumas Unidades da Federação e, em alguns casos, para alguns municípios onde era estudado o comportamento de cada quesito do questionário antes e depois da imputação automática. Para isso, foram utilizados a visualização de imagens, a listagem de registros completos das pessoas, contendo as variáveis a serem analisadas, e as matrizes de contingência, conforme modelo abaixo:

Figura 12.5 - Matriz de contingência com valores antes e após o processo de imputação

VPOSTERIOR / VANTERIOR	Total 1	%	1	2	3	...	j	...	n
Total2									
%									
1									
2									
3									
...									
j									
...									
m									

Onde:

VANTERIOR = código do quesito antes da correção automática

VPOSTERIOR = código do quesito após a correção automática

A acumulação de valores na diagonal indica ausência de modificações no processo, portanto, constituíram o alvo do estudo os casos em que houve significativo aumento, ou redução, de frequência de casos observados em alguma categoria fora da diagonal. As distorções, em geral, desapareceram com a revisão e alteração de algumas regras contidas no plano de crítica. Em alguns casos, elas estavam justificadas por se tratar de correção de erros sistemáticos.

As listagens dos registros e as matrizes de contingência foram obtidas através da utilização do REDATAM+G4 (REcuperação de DAdos para Áreas pequenas por Microcomputador, 4ª Geração), um programa computacional desenvolvido pelo Centro Latino-americano e Caribenho de Demografia – CELADE. Com este objetivo, foi feita a junção dos arquivos antes e depois da imputação.

Faixa de idade

A faixa de idade é uma variável auxiliar, criada para ajudar na imputação de dados e no controle da consistência entre as variáveis imputadas e as demais variáveis relacionadas.

Inicialmente, o cálculo da faixa de idade levou em consideração somente as respostas obtidas em alguns quesitos do bloco de nupcialidade ("Vive em companhia de cônjuge ou companheiro?", "Qual é (era) a natureza da última união?" e "Qual o seu estado civil?"); de mão-de-obra ("Na semana de 23 a 29 de julho trabalhou em atividade de cultivo, extração vegetal, criação de animais ou pesca, destinados à alimentação de pessoas moradoras no domicílio?", "Quantos trabalhos tinha na semana de 23 a 29 de julho?", "Nesse trabalho era:", "No período de 30 de junho a 29 de julho de 2000, tomou alguma providência para conseguir algum trabalho?" e "Em julho de 2000, era aposentado de instituto de previdência oficial?"); de fecundidade ("Quantos(as) filhos(as) nascidos(as) vivos(as) teve até 31 de julho de 2000?" e "Quantos(as) filhos(as) nascidos(as) mortos(as) teve até 31 de julho de 2000?"); de relação com a pessoa responsável pelo domicílio ou de relação com a pessoa responsável pela família.

Foi estabelecida faixa 1 (0 a 9 anos de idade), quando não havia qualquer resposta aos quesitos selecionados de nupcialidade, mão-de-obra e fecundidade, faixa 2 (10 anos ou mais de idade) para os casos em que havia resposta válida em todos os quesitos selecionados de nupcialidade, mão-de-obra e fecundidade, mas não se tratava de pai/mãe ou sogro/sogra do responsável pelo domicílio ou pela família, e faixa 3 (mais de 20 anos) quando, além de haver resposta válida em todos os quesitos selecionados de nupcialidade, mão-de-obra e fecundidade, tratava-se de pai/mãe ou sogro/sogra do responsável pela família ou pelo domicílio.

Embora as perguntas referentes aos quesitos de nupcialidade, fecundidade e mão-de-obra só se apliquem a pessoas com 10 anos ou mais de idade, em alguns casos havia respostas válidas para pessoas com idade entre 0 e 9 anos. Tais respostas ocorreram, principalmente, para o bloco de nupcialidade, sendo que estas eram coerentes com a idade das pessoas para as quais foram obser-

vadas. Em outras palavras, as crianças declararam nunca terem vivido em companhia de cônjuge ou companheiro, assim como declararam serem solteiras. Nestes casos, a pessoa foi alocada nas faixas 2 ou 3 e sua idade trocada posteriormente, quando compatibilizada com a "faixa de idade".

A troca de idade acarretou uma mudança na distribuição de idade, com diminuição do número de crianças de 0 a 9 anos e conseqüente aumento das pessoas com outras idades.

Para a correção do problema, a idade da pessoa foi introduzida no processo de construção da "faixa de idade". Assim, a faixa passou a ser calculada, levando em consideração a idade, sempre que a idade calculada a partir do mês e do ano de nascimento (V4075) fosse igual[5] à idade declarada (V4062). No caso de não haver informação de mês e ano de nascimento ou mesmo de haver diferença entre as idades declarada e calculada, a faixa foi determinada pelos quesitos de nupcialidade, mão-de-obra, fecundidade e relação com a pessoa responsável pelo domicílio ou pela família, conforme procedimento anteriormente adotado.

Declaração de Idade

Apesar da idade calculada através de mês e ano de nascimento ser a informação mais robusta entre as investigadas, houve preocupação em melhorar a sua qualidade. Para tanto, além das perguntas tradicionais do mês e ano de nascimento – "Qual é o mês e o ano do seu nascimento?", assim como da idade presumida – "Qual é a sua idade presumida?", no Censo Demográfico 2000 foi introduzida a idade declarada das pessoas, através da pergunta "Qual era a sua idade em 31 de julho de 2000?".

Inicialmente, o cálculo da idade foi feito a partir das informações já tradicionais, mês e ano de nascimento ou idade presumida. Contudo, a nova informação foi fundamental, pois possibilitou uma análise de compatibilidade entre a idade declarada pela pessoa e a idade calculada, utilizando-se mês e ano de nascimento ou idade presumida.

Algumas divergências foram encontradas, e uma análise mais detalhada foi feita para as pessoas com cem anos ou mais de idade, grupo em que o número de casos com diferença era mais significativo. Em todo o Brasil, foram encontrados na amostra, sem expansão, 1.416 casos com divergência entre os dois tipos de idade, no entanto, havia ao todo 2.544 pessoas com 100 anos ou mais de idade.

Através do sistema de consulta de imagens dos questionários, verificou-se ter havido, na maioria dos casos, erro sistemático na grafia da centena do ano de nascimento, no momento da coleta. Nestes casos, as diferenças entre a idade declarada e a idade calculada pelo mês e ano de nascimento eram, predominantemente, de 100 anos[6]. Verificou-se, ainda, que a década de nascimento das pessoas cujas idades apresentaram esta diferença era predominantemente a de 1980, assim ao invés de grafar 1989 o recenseador teria grafado 1889, por exemplo.

Para ter certeza de que a idade correta era a declarada e não a obtida pelo mês e o ano de nascimento, foram analisados o sexo, a relação com o responsável pelo domicílio e alfabetização. A razão de sexos encontrada(1,13) não é típica entre os mais idosos. Além disso, em 82% dos casos as pessoas eram filhas do responsável pelo domicílio e em 92% sabiam ler ou escrever. Portanto, indicadores mais compatíveis com jovens, conforme indicava a idade declarada.

---

[5] Considerou-se igual sempre que o módulo da diferença era 0 ou 1.

[6] Considerou-se diferença de 100 anos sempre que esta estava entre 99 e 101.

Para estes casos, prevaleceu a idade declarada, ou seja, a idade considerada passou a ser a idade declarada sempre que a idade calculada, a partir do mês e ano de nascimento, era maior que 100 anos e a diferença entre ambas era de 100 anos. Adicionalmente, a idade declarada passou a ser incorporada também para o cálculo da variável auxiliar faixa de idade.

Domicílio

Em relação aos dados preliminares do censo demográfico, foram realizadas análises dos resultados nas etapas de crítica referentes ao NIM, Pré-Dia e Pós-Dia. Inicialmente, as análises contemplaram unidades da federação selecionadas, como o Rio de Janeiro, Bahia e Espírito Santo. Nestes casos, os dados se mostraram coerentes na sua estrutura interna e em comparação aos resultados obtidos no Censo Demográfico 1991. Em um segundo momento, a análise foi ampliada para os demais estados do País com resultados semelhantes aos encontrados na etapa anterior.

Família

Foram analisadas as imputações das relações de parentesco realizadas através de procedimentos automáticos de imputação, utilizando o sistema New Imputation Methodology – NIM.

Os procedimentos utilizados para validação dos resultados foram obtidos a partir da comparação com os resultados de outras pesquisas domiciliares anteriores que continham as mesmas questões, como por exemplo PNADs da década de 1990 e o Censo Demográfico 1991, através de tabulações que apresentavam a distribuição dos tipos de família. Os resultados mostraram que as imputações não alteraram as distribuições das variáveis envolvidas.

Filtros

Tendo em vista que a captura dos dados foi feita através de processo de leitura e reconhecimento ótico de caracteres (ICR) e marcas, foi necessário criar procedimentos para minimizar os erros provenientes de sombras, sujeiras e outras alterações que pudessem modificar as informações advindas dos questionários. Com essa finalidade, foram estabelecidos limites de valores para as respostas obtidas em alguns quesitos, os quais foram considerados razoáveis, tomando-se em conta os valores esperados em cada uma das variáveis envolvidas. Para o controle dos valores que ultrapassassem os limites estabelecidos, foi introduzida uma rotina no programa de reconhecimento de caracteres e marcas cuja função foi recomendar verificação visual dos campos onde foram encontrados tais valores. O processo de verificação teve a finalidade de confirmar os valores lidos pelo *scanner* e não de criticá-los, ou seja, uma vez confirmado o valor através da imagem do questionário, mesmo sendo improvável, o verificador deveria confirmá-lo, pois a crítica e análise de consistência seriam feitas em etapa posterior.

Em razão de terem sido encontrados, já na fase de análise, valores improváveis em algumas variáveis, suspeitou-se que a rotina introduzida no programa de reconhecimento de caracteres e marcas havia falhado, não obstante tivesse funcionado em uma massa de teste feita na ocasião de sua introdução. Para testar o funcionamento da rotina, foi confeccionado um questionário contendo valores fora dos respectivos limites para os campos "Quantos banheiros

existem neste domicílio?", "Qual o mês e ano do seu nascimento?", "Qual era a sua idade em 31 de julho de 2000?", "Qual é a sua idade presumida?" e "Qual foi o seu rendimento bruto do mês de julho de 2000?". No teste, os campos com problemas não foram selecionados para a verificação visual, portanto, foi confirmado que a rotina não estava funcionando.

Considerando-se que este problema afetou apenas a etapa de reconhecimento dos caracteres e marcas e que as ferramentas utilizadas permitiam o reprocessamento com alguma agilidade, foi possível executar novamente a etapa reconhecimento de caracteres e fazer a verificação visual nos casos previstos na rotina.

Como já havia sido feita uma análise exploratória dos dados da amostra, aproveitou-se este momento para solicitar a inclusão de novos valores limites assim como novas variáveis com respectivos limites de valores a serem aceitos sem verificação visual.

Migração

A maioria dos problemas detectados na informação sobre movimentos migratórios, proveniente do campo, foi solucionada a partir de mudanças realizadas no processo de imputação pelo sistema DIA, com exceção do problema referente ao entendimento do quesito 4.23, no qual se investigava "a Unidade da Federação ou país estrangeiro de residência anterior". No processo de análise da informação proveniente do campo, observou-se, em alguns casos, um não entendimento deste quesito. Na realidade, o objeto investigado era a Unidade da Federação na qual o indivíduo residia antes de mudar-se para a Unidade da Federação em que foi recenseado. Esta formulação propiciou um entendimento equivocado do objeto em estudo, na medida em que muitos indivíduos declararam morar há menos de dez anos na Unidade da Federação em que foram recenseados e responderam a própria unidade como residência anterior. Uma das possíveis explicações para esta observação, tem como origem uma interpretação onde confundiu-se "Unidade da Federação de residência anterior" - conforme o texto da pergunta no questionário – com "Unidade da Federação da residência anterior". Assim, muitos dos indivíduos que efetuaram movimentos entre municípios do estado em que foram recenseados, depois de terem ali chegado, provenientes de outra Unidade da Federação, consideraram este último movimento e declararam como Unidade da Federação de residência anterior a própria unidade. Estes casos foram incluídos na categoria "ignorado".

Em virtude da expressiva quantidade de pessoas, cujo tempo ininterrupto de residência na Unidade da Federação era menor que a idade declarada e, consequentemente, teriam que ter declarado outra unidade diferente daquela em que foram investigados, mas não o fizeram, ficou decidido que a informação desses indivíduos seria considerada "ignorada". A justificativa para este fato é proveniente da forma do processo de imputação deste quesito, onde todas as Unidades da Federação e países estrangeiros tinham possibilidade de serem selecionados, segundo a distribuição dentro do lote, o que podia distorcer os fluxos migratórios.

Este fato foi observado em todas as Unidades da Federação, como mostra a Tabela 12.34. Com relação à população não-natural do estado, dos 8.691.756 que responderam ao quesito, 313.590 foram considerados ignorados, por terem respondido o mesmo estado em que foram recenseados, representando 3,6% do total. Os maiores percentuais foram encontrados nos estados do Mato

Grosso (7,6%), Pará (5,2%), São Paulo (4,6%) e Rondônia (4,6%). O menor percentual foi encontrado no Distrito Federal (0,1%), fato esperado, pois só existe um município, o próprio Distrito Federal.

Tabela 12.34 - Pessoas com menos de 10 anos ininterruptos de residência que responderam ao quesito de residência anterior cujas declarações foram consideradas ignoradas - 2000

Unidades da Federação atual	Não naturais com menos de 10 anos ininterruptos de residência		
	Total	Ignorado	% de ignorado
Brasil	**8 691 756**	**313 590**	**3,6**
Rondônia	173 263	7 981	4,6
Acre	23 967	332	1,4
Amazonas	144 991	2 696	1,9
Roraima	83 765	624	0,7
Pará	355 198	18 557	5,2
Amapá	89 055	1 487	1,7
Tocantins	170 058	6 114	3,6
Maranhão	128 687	3 835	3,0
Piauí	95 809	1 258	1,3
Ceará	165 289	2 907	1,8
Rio Grande do Norte	94 392	1 494	1,6
Paraíba	108 909	2 196	2,0
Pernambuco	194 921	5 183	2,7
Alagoas	74 896	2 287	3,1
Sergipe	79 868	1 649	2,1
Bahia	296 706	8 657	2,9
Minas Gerais	548 244	13 649	2,5
Espírito Santo	221 429	4 808	2,2
Rio de Janeiro	561 315	22 891	4,1
São Paulo	2 638 297	122 676	4,6
Paraná	426 257	16 185	3,8
Santa Catarina	327 143	7 499	2,3
Rio Grande do Sul	149 593	2 988	2,0
Mato Grosso do Sul	176 171	5 824	3,3
Mato Grosso	362 108	27 385	7,6
Goiás	598 356	22 149	3,7
Distrito Federal	403 070	282	0,1

Fonte: IBGE, Censo Demográfico 2000.

Educação

Em primeiro lugar, serão apresentados os limites etários estabelecidos para a imputação de dados relativos à freqüência escolar para, em seguida, serem relatadas as inconsistências verificadas após a aplicação do sistema DIA, assim como os procedimentos adotados para correção.

Limites etários estabelecidos para a imputação de dados nos quesitos "Qual é o curso que frequenta?" e "Qual é a série que frequenta?".

Os limites inferiores e superiores de idade para a freqüência escolar em cada nível ou modalidade de ensino foram definidos, levando-se em conta o disposto na Lei de Diretrizes e Bases da Educação – LDB – além de discutidos com técnicos do Ministério da Educação e Cultura – MEC, responsáveis pela elaboração do Censo Escolar. Sendo assim, as idades consideradas adequadas para cursar os diversos níveis e modalidades de ensino serviram de referência para a construção de tais limites, que foram ampliados, levando-se em consideração a realidade observada no campo.

Educação Infantil (Creche e Pré-escola)

Foram consideradas as idades pontuais entre 0 e 11 anos, sendo de 0 a 4 anos para freqüência à creche e 3 a 11 anos para pré-escola.

Classe de alfabetização

Para o curso regular, foram consideradas as idades de 4 a 11 anos. Para a classe de Alfabetização de Adultos, foram consideradas as idades a partir de 13 anos.

Ensino Fundamental

Para a freqüência neste nível de ensino, foram considerados os seguintes limites etários: regular seriado ou não seriado, mínimo 5 anos de idade; e supletivo de 1º grau, mínimo 13 anos de idade.

Ensino Médio

Para a freqüência neste nível de ensino, foram considerados os seguintes limites etários: regular seriado ou não seriado, mínimo 13 anos de idade; e supletivo de 2º grau, mínimo 17 anos de idade.

Educação Superior

Foram considerados os seguintes limites etários: graduação, mínimo 16 anos de idade e mestrado ou doutorado, mínimo 20 anos de idade.

Análise dos resultados

Inicialmente, foi feita uma análise de caráter exploratório a partir de tabelas de contingência e da construção de indicadores. Nesta etapa, observou-se que as informações sobre o analfabetismo estavam compatíveis com o esperado. Paralelamente, foi realizado um estudo da estrutura etária, através do qual detectou-se que a aplicação do sistema DIA acarretou alterações nas freqüências das idades de 0 a 2, 5 e 6 anos, pois promoveu alteração nas idades para que estas ficassem compatíveis com os níveis de ensino freqüentados.

Para corrigir as mudanças na estrutura etária, foram elaboradas algumas regras de imputação determinística. Na elaboração, foram consideradas as informações adicionais obtidas com a análise detalhada dos dados e também com a troca de experiência com os técnicos do MEC. As regras aplicadas foram as seguintes:

primeira mudança no nível freqüentado – pessoas com 5 ou 6 anos de idade, com declaração de freqüência à creche, passaram a freqüência da pré-escola; e

segunda mudança no nível freqüentado – crianças com 0, 1 e 2 anos de idade, com declaração de freqüência à pré-escola, passaram a freqüência da creche.

A análise das variáveis "Qual é o curso mais elevado que freqüentou, no qual concluiu pelo menos uma série?", "Qual é a última série concluída com aprovação?" e "Concluiu o curso no qual estudou?" foi feita, inicialmente, com os dados do Espírito Santo, Rio de Janeiro e Bahia.

Detectou-se uma alteração sistemática na distribuição de freqüência da variável "Qual é o curso mais elevado que freqüentou, no qual concluiu pelo menos uma série?", após a aplicação do sistema DIA. Esta alteração foi verificada sempre que a pessoa, com idade entre 17 e 37 anos e que concluiu o curso no qual estudou, possuía declaração de conclusão do antigo primário, ensino fundamental ou 1ª grau e a última série concluída com aprovação era 4ª e, em alguns casos, 3ª. Nesses casos, a informação do curso mais elevado que freqüentou, no qual concluiu pelo menos uma série, foi alterada para ensino médio ou 2º grau, mais compatível com as séries informadas e o curso concluído.

Em razão desta alteração, aumentou o número de pessoas com 11 anos de estudo e, conseqüentemente, diminuiu o número das que tinham 4 anos. No entanto, a média de anos de estudo para a população de 10 anos ou mais de idade apresentou uma alteração pequena, em torno de 0,2 pontos percentuais.

Em consulta feita ao MEC, foi obtida a informação de que, até recentemente, as escolas forneciam certificado de conclusão do primeiro segmento do ensino fundamental às pessoas que concluíam este nível através de curso supletivo. Essa informação nos levou a considerar que os recenseadores podiam ter entendido que as pessoas concluíram o curso fundamental e não apenas o primeiro segmento do fundamental, como seria o correto.

Para corrigir as alterações detectadas, criaram-se as seguintes críticas:

1. se a pessoa tinha mais de 16 e menos de 37 anos e concluiu o antigo primário em 3 ou 4 séries, então o curso mais elevado que freqüentou, no qual concluiu pelo menos uma série, passava para ensino fundamental ou 1º grau incompleto. Este procedimento também foi aplicado aos casos em que o quesito "Concluiu o curso no qual estudou?" estava em branco; e.

2. se a pessoa tinha mais de 16 anos e concluiu o ensino fundamental ou 1º grau em 3 ou 4 séries, então se declarou sim no quesito "Concluiu o curso no qual estudou?" passava para não concluiu. Este procedimento também foi aplicado ao caso em que o quesito "Concluiu o curso no qual estudou?" estava em branco.

Trabalho e Rendimento

Durante a fase de crítica e imputação dos dados da parte de trabalho do Questionário da Amostra do Censo Demográfico 2000, foram realizadas análises de consistência para cada característica investigada, considerando os resultados de cada quesito e sua associação com outras informações do próprio questionário. Estas análises revelaram a existência de algumas falhas sistemáticas, oriundas da fase de coleta, que o procedimento geral de crítica e imputação não poderia solucionar. Assim sendo, foram adotados procedimentos específicos e de imputação determinística para resolver estas situações. Constatou-se, ainda, que para determinados segmentos da população, a imputação de quesitos omitidos causava distorção nos resultados de algumas características. Também para solucionar estas situações, foram adotados procedimentos de imputação determinística.

Ao final de todo o processo de apuração, foi desenvolvido o estudo, apresentado a seguir, sobre a imputação das informações do tema trabalho e rendimento.

A análise da imputação das variáveis foi iniciada com a identificação das informações dos indivíduos que, originalmente, não tinham qualquer resposta aos quesitos do tema trabalho e rendimento (todos eles em branco). Para todo o Brasil, ocorreram 24 268 desses casos, que representam 0,1% do total de pessoas com 10 anos ou mais de idade da amostra (16 194 309).

Nesse grupo de casos que estavam originalmente em branco, 50 foram imputados como ocupados, dos quais 46 foram classificados como trabalhadores domésticos (44 deles com carteira de trabalho assinada e 2 sem carteira assinada) e 4 como empregados com carteira de trabalho assinada. Outros 2 117 desses casos foram imputados como pessoas desocupadas. Os demais (22 101 casos) foram imputados, portanto, como pessoas não-economicamente ativas.

No grupo de pessoas imputadas como "não-economicamente ativas" (22 101 casos), a maior parte estava na faixa de 10 a 14 anos de idade (71,2%), sendo que 15 402 pessoas não sofreram imputação na variável V4752 "idade calculada em anos" (69,7%), conforme tabela a seguir.

Tabela 12.35 - Pessoas que tinham as variáveis de trabalho e rendimento originalmente em branco e foram imputadas como não economicamente ativas, por condição de imputação na variável V4752, segundo os grupos de idade

Grupos de idade	Pessoas que tinham as variáveis de trabalho e rendimento originalmente em branco e foram imputadas como não economicamente ativas		
	Total	Condição de imputação na variável V4752	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**22 101**	**21 306**	**795**
10 a 14 anos	15 732	15 402	330
15 a 24 anos	3 137	2 956	181
25 anos ou mais	3 232	2 948	284
	Valores relativos		
Total	**100,0**	**96,4**	**3,6**
10 a 14 anos	71,2	69,7	1,5
15 a 24 anos	14,2	13,4	0,8
25 anos ou mais	14,6	13,3	1,3

Fonte: IBGE, Censo Demográfico 2000.

Posição na ocupação

O estudo dos temas trabalho e rendimento foi iniciado com a variável V0447, que define a posição na ocupação. A partir do levantamento das proporções de informações imputadas no processo de apuração, identificou-se a necessidade de aprofundar a caracterização de algumas posições tais como trabalhadores domésticos e aprendizes ou estagiários.

Trabalhadores Domésticos

No estudo da variável "posição na ocupação", verificou-se que as proporções de imputações de informações nos códigos 1 (trabalhador doméstico com carteira de trabalho assinada) e 2 (trabalhador doméstico sem carteira de trabalho assinada) da variável V0447, ao final do processo de apuração, foram extremamente elevadas, como pode ser constatado na Tabela 12.36 a seguir:

Tabela 12.36 - Pessoas registradas na variável V0447, com e sem imputação, segundo as categorias da variável

Categorias da variável	Pessoas registradas na variável V0447					
	Valores absolutos			Valores relativos (%)		
	Total	Sem imputação	Com imputação	Total	Sem imputação	Com imputação
Total	**20 274 412**	**19 910 122**	**364 290**	**100,0**	**98,2**	**1,8**
Trabalhador doméstico com carteira de trabalho assinada	164 674	100 895	63 779	100,0	61,3	38,7
Trabalhador doméstico sem carteira de trabalho assinada	414 090	286 083	128 007	100,0	69,1	30,9
Empregado com carteira de trabalho assinada	2 470 074	2 411850	58 224	100,0	97,6	2,4
Empregado sem carteira de trabalho assinada	1 930 353	1 901 589	28 764	100,0	98,5	1,5
Empregador	209 478	207 903	1 575	100,0	99,2	0,8
Conta própria	1 858 789	1 806 211	52 578	100,0	97,2	2,8
Aprendiz ou estagiário sem remuneração	32 256	30 446	1 810	100,0	94,4	5,6
Não remunerado em ajuda a membro do domicílio	393 503	391 112	2 391	100,0	99,4	0,6
Trabalhador na produção para o próprio consumo	309 682	308 196	1 486	100,0	99,5	0,5
Branco	12 491 513	12 465 837	25 676	100,0	99,8	0,2

Fonte: IBGE, Censo Demográfico 2000.

A fim de permitir um melhor entendimento e uma avaliação mais adequada dos fatores que levaram ao elevado número de imputações nos itens 1 e 2 dessa variável, foram feitos, inicialmente, outros cruzamentos de variáveis. Para orientar esses novos cruzamentos, levou-se em consideração a informação sobre ocupação das pessoas selecionadas e a atividade econômica do empreendimento em que trabalhavam.

Para as pessoas classificadas como trabalhadoras domésticas com imputação na variável V0447, foi verificada a ocorrência de imputações nas variáveis que identificavam a atividade do empreendimento em que trabalhavam (V4462) e a ocupação da pessoa (V4452). O resultado desse cruzamento revelou que:

1 - Para o conjunto de 63 779 pessoas que, ao final do processo de apuração, foram imputadas como trabalhadoras domésticas com carteira de trabalho assinada, 52 555 não apresentavam marca de imputação na variável identificadora da ocupação (V4452) e nem na variável identificadora de atividade econômica do empreendimento (V4462). Assim sendo, 82,4% das pessoas com imputação no item 1 da variável V0447 foram assim classificadas em função da sua ocupação e da atividade econômica; portanto, estavam claramente identificadas como trabalhadoras domésticas. Constataram-se 6 casos de pessoas sem imputação na variável de atividade (que identifica a atividade econômica do serviço doméstico em um código especifico) e com imputação na variável de ocupação e, ainda, 9 242 casos de pessoas sem imputação na variável de ocupação (em que, do conjunto de ocupações admissíveis para a atividade econômica do serviço doméstico, uma parcela é exclusiva dessa atividade) e com imputação na variável de atividade econômica. Em 1 976 casos (3,1% dos casos em que houve imputação no item 1 da variável V0447), as variáveis de ocupação e atividade também foram imputadas.

Tabela 12.37 - Pessoas registradas com imputação no item 1 - 'trabalhador doméstico com carteira de trabalho assinada' da variável V0447, por condição de imputação na variável V4462 - atividade, segundo a condição de imputação na variável V4452 - ocupação

Condição de imputação na variável V4452 - ocupação	Pessoas registradas com imputação no item 1 - trabalhor doméstico com carteira de trabalho assinada da variável V0447		
	Total	Condição de imputação na variável V4462 - atividade	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**63 779**	**52 561**	**11 218**
Sem imputação	61 797	52 555	9 242
Com imputação	1 982	6	1 976
	Valores relativos		
Total	**100,0**	**82,4**	**17,6**
Sem imputação	96,9	82,4	14,5
Com imputação	3,1	0,0	3,1

Fonte: IBGE, Censo Demográfico 2000.

2 - Para o conjunto de 128 007 pessoas que, ao final do processo de apuração, foram imputadas como trabalhadoras domésticas sem carteira de trabalho assinada, 112 675 não apresentavam marca de imputação na variável identificadora da ocupação (V4452) e nem na variável identificadora da atividade econômica (V4462). Assim sendo, 88,0% das pessoas com imputação no item 2 da variável V0447 em função da sua ocupação e atividade estavam claramente identificadas como trabalhadoras domésticas. Constataram-se, também, 20 casos de pessoas sem imputação na variável de atividade (que identifica a atividade econômica do serviço doméstico em um código especifico) e com imputação na variável de ocupação e, ainda, 14 324 casos de pessoas sem imputação na variável de ocupação (em que o conjunto de ocupações admissíveis para a atividade econômica do serviço doméstico, uma parcela é específica unicamente dessa atividade). Em 988 casos (0,8% dos casos em que houve imputação no item 2 da variável V0447), as variáveis de ocupação e atividade também foram imputadas.

Tabela 12.38 - Pessoas registradas com imputação no item 2 - 'trabalhador doméstico sem carteira de trabalho assinada' da variável V0447, por condição de imputação na variável V4462 - atividade, segundo a condição de imputação na variável V4452 - ocupação

Condição de imputação na variável V4452 - ocupação	Pessoas registradas com imputação no item 2 - trabalhor doméstico sem carteira de trabalho assinada da variável V0447		
	Total	Condição de imputação na variável V4462 - atividade	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**128 007**	**112 695**	**15 312**
Sem imputação	126 999	112 675	14 324
Com imputação	1 008	20	988
	Valores relativos		
Total	**100,0**	**88,0**	**12,0**
Sem imputação	99,2	88,0	11,2
Com imputação	0,8	0,0	0,8

Fonte: IBGE, Censo Demográfico 2000.

Com esses primeiros cruzamentos, justifica-se a maior parte das informações imputadas para pessoas classificadas nas posições na ocupação como trabalhadoras domésticas com e sem carteira de trabalho assinada (respectivamente, itens 1 e 2 da variável V0447) a partir das informações sobre ocupação e atividade econômica correspondentes.

Entre outras investigações, ainda foi verificada a parcela em que se têm indicativos sobre carteira de trabalho assinada. Do conjunto de pessoas classificadas como trabalhadoras domésticas com carteira de trabalho assinada, com imputação dessa característica, 78,1% (49 812) tinham sido originalmente classificadas como empregadas com carteira de trabalho assinada, sendo que em 66,5% dos casos (42 405) não houve imputação de ocupação nem de atividade. Assim, a identificação original de pessoa com registro em carteira de trabalho foi mantida após a imputação de posição na ocupação na categoria de trabalhadores domésticos.

Tabela 12.39 - Pessoas registradas com imputação no item 1 - 'trabalhador doméstico com carteira de trabalho assinada' da variável V0447, por condição de imputação na variável V4462 - atividade, segundo a condição de imputação na variável V4452 - ocupação

Condição de imputação na variável V4452 - ocupação	Pessoas registradas com imputação no item 1 - trabalhor domético com carteira de trabalho assinada da variável V0447, originalmente classificadas como empregador com carteira trabalho assinada		
	Total	Condição de imputação na variável V4462 - atividade	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**49 812**	**42 406**	**7 406**
Sem imputação	49 792	42 405	7 387
Com imputação	20	1	19
	Valores relativos		
Total	**100,0**	**85,1**	**14,9**
Sem imputação	100,0	85,1	14,8
Com imputação	0	0,0	0,0

Fonte: IBGE, Censo Demográfico 2000.

Essa informação foi também importante na imputação das pessoas classificadas como trabalhadoras domésticas sem carteira de trabalho assinada. Dessas pessoas, 81,8% (104 726) tinham sido originalmente classificadas como empregadas sem carteira de trabalho assinada, sendo que para 73,0% delas (93 397) não houve imputação de ocupação nem de atividade.

Tabela 12.40 - Pessoas registradas com imputação no item 2 - 'trabalhador doméstico sem carteira de trabalho assinada' da variável V0447, por condição de imputação na variável V4462 - atividade, segundo a condição de imputação na variável V4452 - ocupação

Condição de imputação na variável V4452 - ocupação	Pessoas registradas com imputação no item 2 - trabalhor domético sem carteira de trabalho assinada da variável V0447 originalmente classificadas como empregador sem carteira trabalho assinada		
	Total	Condição de imputação na variável V4462 - atividade	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**104 726**	**93 398**	**11 328**
Sem imputação	104 704	93 397	11 307
Com imputação	22	1	21
	Valores relativos		
Total	**100,0**	**89,2**	**10,8**
Sem imputação	100,0	89,2	10,8
Com imputação	0,0	0,0	0,0

Fonte: IBGE, Censo Demográfico 2000.

Portanto, as informações sobre a ocupação da pessoa e a atividade econômica do empreendimento em que trabalhava, associadas a informações sobre registro em carteira de trabalho, explicam a maior parte das imputações da posição na ocupação de trabalhadores domésticos com e sem carteira de trabalho assinada.

Por outro lado, no grupo de pessoas com imputação da posição na ocupação como trabalhadoras domésticas com carteira de trabalho assinada, 1 886 não registraram originalmente informações nas variáveis de posição na ocupação, ocupação nem atividade do empreendimento em que trabalhavam, o que corresponde a 3,0% desse grupo ao final do processo de apuração e crítica. Ou seja, as variáveis V0447, V4452 e V4462 estavam originalmente em branco. Para as pessoas com imputação da posição na ocupação como trabalhadoras domésticas sem carteira de trabalho assinada, 199 delas (ou 0,2%) não possuíam informações originais nessas variáveis. Nesses casos, outras informações foram consideradas no processo de crítica para a definição da posição na ocupação.

Aprendizes ou estagiários sem remuneração

Para as pessoas classificadas como aprendizes ou estagiárias sem remuneração, a proporção de imputação do código de posição na ocupação foi de 5,6%, que corresponde a 1 810 indivíduos.

Considerando esse grupo de pessoas com informação imputada de posição na ocupação "aprendiz ou estagiário sem remuneração", 1 696 pessoas declararam que ajudaram, sem remuneração, no trabalho exercido por pessoa conta própria ou empregadora, moradora do domicílio, ou como aprendiz ou estagiário (V0441), conforme descrito na tabela a seguir.

Tabela 12.41 - Pessoas registradas com imputação no item 7 - Aprendizes ou estagiários sem remuneração da variável V0447, por condição de imputação na variável V0441,segundo a variável V0441

Variável V0441	Pessoas registradas com imputação no item 7 - Aprendizes ou estagiários sem remuneração da variável V0447		
	Total	Condição de imputação na variável V0441	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**1 810**	**1 729**	**81**
Sem imputação	1 776	1 696	80
Com imputação	34	33	1
	Valores relativos		
Total	**100,0**	**95,5**	**4,5**
Sem imputação	98,1	93,7	4,4
Com imputação	1,9	1,8	0,1

Fonte: IBGE, Censo Demográfico 2000.

O registro do código 1 da variável V0441 poderia levar à imputação como "aprendiz ou estagiário sem remuneração" ou como "trabalhador não remunerado em ajuda a membro do domicílio"; entretanto, proporcionalmente, houve mais imputação no código 7 (aprendiz ou estagiário sem remuneração) do que no código 8 (trabalhador não remunerado em ajuda a membro do domicílio).

Dois ou mais trabalhos

No que diz respeito aos indivíduos com mais de um trabalho, ocorreram 15 180 casos de imputação de informações, que representaram 5,7% do total de indivíduos nesta situação. Os indivíduos que possuíam informações originais, sem imputação, em pelo menos uma das variáveis horas trabalhadas nos demais trabalhos (V0454), recebiam apenas em benefícios nos demais trabalhos (V4521 igual à opção 1) ou receberam algum valor de rendimento dos demais trabalhos (V4522) eram 11 477, correspondendo a 75,6% dos casos de imputação de dois ou mais trabalhos (V0444).

Observando cada um desses grupos separadamente, 10 425 pessoas informaram horas trabalhadas nos demais trabalhos, 7 325 informaram valor de rendimento nos demais trabalhos e 841 informaram receber somente em benefícios nos demais trabalhos.

Por outro lado, no caso da variável V0454 "horas trabalhadas nos demais trabalhos na semana", 11,6% dos casos (correspondentes a 31 068 pessoas) com horas diferentes de "zero" foram imputados. Desse grupo, 26 313 pessoas (84,7%) declararam originalmente possuir dois ou mais trabalhos na semana de referência (variável V0444 sem imputação).

Rendimentos

A variável indicadora de imputação de que a pessoa não possuía rendimento do trabalho principal (V4511) teve proporção significativa de valores imputados (23,8%, que representam 101 236 pessoas). Nesse grupo, 89,2% das pessoas informaram originalmente que eram trabalhadores não remunerados em ajuda a membro da unidade familiar e outros 10,1% que eram aprendizes ou estagiários sem remuneração. Portanto, 99,2% (ou 100 465 pessoas) tinham informações originais que justificavam a imputação de que não tinham rendimento do trabalho principal.

Tabela 12.42 - Pessoas registradas com imputação no item 0 da variável V4511, por condição de imputação na variável V0447, segundo a variável V0447

Variável V0447	Pessoas registradas com imputação no item 0 (1)		
	Total	Condição de imputação na variável V0447	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**101 236**	**100 465**	**771**
Aprendiz ou estagiário sem remuneração	10 618	10 207	411
Não remunerado em ajuda a membro do domicílio	90 618	90 258	360
	Valores relativos		
Total	**100,0**	**99,2**	**0,8**
Aprendiz ou estagiário sem remuneração	10,5	10,1	0,4
Não remunerado em ajuda a membro do domicílio	89,5	89,2	0,4

Fonte: IBGE, Censo Demográfico 2000.
(1) Não tem rendimento no trabalho principal.

Já a imputação de informações na variável V4521 "não tem rendimento nos demais trabalhos", apesar de pequena em relação ao total de informantes (3,9%), representa 280 193 pessoas. Desse grupo, 89,7% afirmaram originalmente que tinham apenas um trabalho na semana de referência (V0444), o que justifica, portanto, a maior parte da imputação na V4521 (251 313 pessoas).

Tabela 12.43 - Pessoas registradas com imputação no item 0 da variável V4521, por condição de imputação na variável V0444, segundo a variável V0444

Variável V0444	Pessoas registradas com imputação no item 0 (1)		
	Total	Condição de imputação na variável V0444	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**280 193**	**255 747**	**24 446**
Aprendiz ou estagiário sem remuneração	274 895	251 313	23 582
Não remunerado em ajuda a membro do domicílio	5 298	4 434	864
	Valores relativos		
Total	**100,0**	**91,3**	**8,7**
Aprendiz ou estagiário sem remuneração	98,1	89,7	8,4
Não remunerado em ajuda a membro do domicílio	1,9	1,6	0,3

Fonte: IBGE, Censo Demográfico 2000.
(1) Não tem rendimento nos demais trabalhos.

As proporções de pessoas com informações imputadas de rendimentos não oriundos de trabalho foram baixas, sempre inferiores a 0,8% do total, conforme tabela a seguir:

Tabela 12.44 - Pessoas registradas nas variáveis de rendimentos não oriundos de trabalho, por condição de imputação, segundo as variáveis indicadoras de imputação dos rendimentos

Variáveis indicadoras de imputação dos rendimentos	Pessoas registradas nas variáveis de rendimentos não oriundos de trabalho		
	Total	Condição de imputação	
		Sem imputação	Com imputação
	Valores absolutos		
M4573 - Aposentadoria e pensão	16 194 309	16 081 787	112 522
M4583 - Aluguel	16 194 309	16 099 246	95 063
M4593 - Pensão alimentícia, mesada ou doação de não morador	16 194 309	16 097 014	97 295
M4603 - Renda mínima, bolsa-escola (programas oficias de auxílio)	16 194 309	16 089 242	105 067
M4613 - Outros rendimentos	16 194 309	16 080 151	114 158
	Valores relativos		
M4573 - Aposentadoria e pensão	100,0	99,3	0,7
M4583 - Aluguel	100,0	99,4	0,6
M4593 - Pensão alimentícia, mesada ou doação de não morador	100,0	99,4	0,6
M4603 - Renda mínima, bolsa-escola (programas oficias de auxílio)	100,0	99,4	0,6
M4613 - Outros rendimentos	100,0	99,3	0,7

Fonte: IBGE, Censo Demográfico 2000.

Ocupação e Atividade Econômica

No questionário do Censo Demográfico 2000, as informações sobre ocupação e atividade econômica são descritas pelo informante e registradas por extenso. Na apuração, estas descrições foram associadas a categorias definidas nas classificações de ocupações e de atividades econômicas, por meio de códigos numéricos. A imputação de informações nessas variáveis é posterior ao processo de codificação.

As críticas referentes à parte de trabalho e rendimento estão divididas em dois aplicativos do sistema DIA. Concluída a execução do primeiro aplicativo, as variáveis nele tratadas não podem mais ser alteradas no segundo aplicativo.

As variáveis V0447 “posição na ocupação” e V4452 “ocupação” foram tratadas no primeiro aplicativo, enquanto que a variável V4462 (atividade econômica) fez parte do segundo aplicativo. Assim sendo, no caso de ter havido inconsistência entre o registro da variável V4462 e os das variáveis V0447 e V4452, o procedimento adotado foi o de ajustar o registro da variável V4462.

Tabela 12.45 - Pessoas registradas nas variáveis V4452 e V4462, por condição de imputação, segundo as variáveis

Variáveis	Pessoas registradas nas variáveis V4452 e V4462		
	Total	Condição de imputação	
		Sem imputação	Com imputação
	Valores absolutos		
V4452 - ocupação	7 782 899	7 446 237	336 332
V4462 - atividade	7 782 899	6 746 051	1 036 848
	Valores relativos		
V4452 - ocupação	100,0	95,7	4,3
V4462 - atividade	100,0	86,7	13,3

Fonte: IBGE, Censo Demográfico 2000.

A proporção de informações imputadas na variável V4452 “ocupação” foi de 4,3%. Desse grupo, 14,6% (ou 49 081 pessoas) foram classificadas no processo de codificação como tendo ocupação "ignorada" e passaram a ter a informação de que sua ocupação estava mal definida ou era ignorada (código "zero"). Outras 7,7% (25 959 pessoas) tinham originalmente código de ocupação em branco. As demais pessoas sofreram imputação no código de ocupação como resultado do processo de crítica.

Já os registros da variável de atividade econômica (V4462) tiveram maior proporção de informações imputadas (13,3% dos códigos ou 1 036 848 pessoas). Desses, 8,5% (87 997 pessoas) estavam originalmente em branco. Um grupo majoritário de 76,3% (ou 791 307 pessoas) não tinha marca de imputação nem na variável V0447 "posição na ocupação", nem na variável V4452 "ocupação", ou seja, possuía alguma resposta nessas questões. Outros 17,2% sofreram imputação no código de ocupação, mas declararam sua posição na ocupação.

Tabela 12.46 - Pessoas registradas com imputação na variável V4462, por condição de imputação na variável V4452, segundo a condição de imputação na variável V0447

Condição de imputação na variável V0447	Pessoas registradas com imputação na variável V4462		
	Total	Condição de imputação	
		Sem imputação	Com imputação
	Valores absolutos		
Total	**1 036 848**	**849 427**	**187 421**
Sem imputação	969 692	791 307	178 385
Com imputação	67 156	58 120	9 036
	Valores relativos		
Total	**100,0**	**81,9**	**18,1**
Sem imputação	93,5	76,3	17,2
Com imputação	6,5	5,6	0,9

Fonte: IBGE, Censo Demográfico 2000.

Com a finalidade de explicar essa imputação de 13,3% na variável atividade econômica, foi construída a matriz de contingência para o Brasil. Nela, foram identificados os maiores fluxos de imputação entre grupos de atividade econômica. Em uma análise integrada por especialistas de todas as áreas envolvidas, considerou-se razoável a imputação feita através do sistema DIA, tendo em vista que as atividades imputadas passaram a ter coerência com a ocupação e posição na ocupação declaradas pelo informante. Declarações estas que também foram confirmadas, quando comparadas com as imagens digitalizadas dos questionários. O mesmo procedimento foi utilizado para confirmar as imputações de ocupação.

Outras variáveis relacionadas com os temas "Trabalho e rendimento"

Nas demais variáveis do grupo temático "Trabalho e rendimento", as proporções de informações imputadas ficaram, majoritariamente, em torno de 1%.

Os valores absolutos e as proporções de informações imputadas nas demais variáveis do tema estão na Tabela 12.46 que segue.

Tabela 12.47 - Pessoas registradas em outras variáveis relacionadas com o tema "Trabalho e rendimento", por condição de imputação, segundo as variáveis

Variáveis	Pessoas registradas nas variáveis V4452 e V4462		
	Total	Condição de imputação	
		Sem imputação	Com imputação
Valores absolutos			
V0439 - Trabalho remunerado na semana	16 194 309	16 056 004	138 305
V0440 - Estava afastado temporariamente do trabalho	9 463 867	9 192 840	271 027
V0441 - Trabalho não remunerado em ajuda a conta própria ou empregador	9 143 596	9 034 042	109 554
V0442 - Trabalho não remunerado em ajuda empregado	8 880 062	8 727 384	152 678
V0443 - Trabalhador produção próprio consumo	8 721 092	8 613 347	107 745
V0448 - Empregados pelo Regime Jurídico dos Funcionários Públicos ou como Militares	1 930 353	1 890 701	39 652
V0449 - Quantidade de empregados	209 478	207 857	1 621
V0450 - Constribuição para previdência	3 980 423	3 945 532	34 891
V0453 - Horas trab. principal na semana	7 782 899	7 718 116	64 783
V0455 - Providência para conseguir trabalho	8 411 410	8 333 781	77 629
V0456 - Aposentado de inst. previdência oficial	16 194 309	15 896 974	297 335
Valores relativos			
V0439 - Trabalho remunerado na semana	100,0	99,1	0,9
V0440 - Estava afastado temporariamente do trabalho	100,0	97,1	2,9
V0441 - Trabalho não remunerado em ajuda a conta própria ou empregador	100,0	98,8	1,2
V0442 - Trabalho não remunerado em ajuda a empregado	100,0	98,3	1,7
V0443 - Trabalhador produção próprio consumo	100,0	98,8	1,2
V0448 - Empregados pelo Regime Jurídico dos Funcionários Públicos ou como Militares	100,0	97,9	2,1
V0449 - Quantidade de empregados	100,0	99,2	0,8
V0450 - Constribuição para previdência	100,0	99,1	0,9
V0453 - Horas trab. principal na semana	100,0	99,2	0,8
V0455 - Providência para conseguir trabalho	100,0	99,1	0,9
V0456 - Aposentado de inst. previdência oficial	100,0	98,2	1,8

Fonte: IBGE, Censo Demográfico 2000.

Fecundidade e Mortalidade Infantil

As informações referentes à Fecundidade e Mortalidade Infantil constituem um único bloco no questionário do Censo Demográfico 2000, portanto, os procedimentos para a análise das variáveis relacionadas a esses temas foram semelhantes.

Em um primeiro momento, foram descobertos erros sistemáticos no campo, tais como informações deixadas em branco e valores fora dos limites, e ensaiaram-se diversos procedimentos de imputação determinística para corrigi-los. Detectaram-se, também, problemas na leitura realizada pelo *scanner* que afetavam os níveis de mortalidade infantil, corrigidos mediante a aplicação dos filtros descritos no anexo de CD-ROM deste texto.

Quando do cálculo das taxas de mortalidade infantil, utilizando informações preliminares da amostra do Censo 2000 sobre filhos tidos nascidos vivos e filhos sobreviventes, ocorreram divergências ao comparar os resultados das estimativas obtidas com as esperadas. Essa comparação tomou como parâmetros as estimativas derivadas do Projeto de Projeções de População por sexo e idade para o Brasil, Grandes Regiões e Unidades da Federação, do IBGE, realizado em convênio com o Fundo da População das Nações Unidas – FNUAP, que utilizou a metodologia dos Métodos das Componentes e as decorrentes da incorporação das séries da Pesquisa Nacional por Amostra de Domicílios – PNAD – da década de 1990, à série histórica obtida via Censos Demográficos 1940 a 1991.

Em decorrência da extrema sensibilidade das estimativas de mortalidade às variações nos dados originais e das diferenças observadas nos dados básicos sem crítica, criou-se dentro do âmbito da Diretoria de Pesquisas um grupo especial de trabalho para analisar o tratamento das informações sobre fecundidade das mulheres e mortalidade de seus respectivos filhos. Esse grupo foi integrado por demógrafos deste Instituto, por estatísticos e especialistas na área de metodologia. As primeiras estimativas já apontavam para uma diferença nas taxas de mortalidade infantil, que apresentava aumento em relação às séries observadas a partir das PNADs realizadas na década de 1990. Visando aprofundar a investigação desse problema, foi solicitado ao Departamento de Metodologia um estudo sobre a confiabilidade e a variabilidade das estimativas da mortalidade infantil, baseado nos dados da PNAD.

Esse estudo (SILVA; PESSOA, 2002) esclareceu o porquê das diferenças observadas entre os dados do Censo e da PNAD. Em relação a esta última, o fato de a amostra ser sempre escolhida a partir dos mesmos setores e a estimativa analisada ser muito sensível a pequenas variações nos dados básicos, causava diferenças nos valores obtidos. Embora essa tendência seja correta para essa amostra, não se ajusta completamente ao comportamento da mortalidade no Brasil.

Com base nas conclusões do grupo de trabalho, decidiu-se que as informações de mortalidade e fecundidade seriam divulgadas em volume separado, com os cálculos dos indicadores efetuado a partir da amostra completa do Censo Demográfico 2000.

Com respeito ao tratamento da informação para a divulgação definitiva dos resultados, os exercícios realizados apontaram para uma relativa aproximação entre os valores das taxas e os valores esperados, porém mostraram de forma igual que o processo de crítica e imputação poderia não ser suficiente para corrigir os problemas detectados, especialmente considerando que as correções se restringiam aos dados que apresentavam inconsistências entre si, isto é, que as informações mencionadas já estavam dentro do intervalo definido para cada variável ou conjunto de variáveis. Nesse sentido, os filtros e procedimentos de verificação solicitados para a etapa da entrada de dados foram efetivamente aplicados

antes que o conjunto das informações fosse submetido aos procedimentos de crítica através do DIA. Posteriormente, aplicaram-se as imputações determinísticas para resolver os problemas mais freqüentes na informação básica.

Ao final do processo de crítica e imputação, concluiu-se sobre as estimativas da Fecundidade e da Mortalidade Infantil que: as distribuições das variáveis que permitiam estimar os citados parâmetros demográficos não sofreram alterações com o processo de crítica e imputação; e as estimativas propriamente ditas não foram modificadas substancialmente com a passagem do sistema DIA.

Resumindo, na divulgação dos resultados definitivos, foram utilizados procedimentos normais de crítica e imputação já com os dados depurados, chegando-se a resultados finais não muito distintos daqueles divulgados inicialmente, na ocasião do lançamento da *Tabulação avançada do censo demográfico 2000: resultados preliminares da amostra (2002).*

Ainda analisando os dados do bloco de Fecundidade, observou-se que os níveis da natimortalidade, expressos pelos percentuais de "filhos tidos nascidos mortos no total de filhos tidos", permaneceram elevados ao longo de todo o processo de crítica e imputação. Tal fenômeno foi observado no Censo Demográfico 1980 e foi constatado que o mesmo não fora provocado pelo processo de imputação.

Relativamente às razões de sexo dos "filhos tidos nascidos vivos e nascidos mortos", segundo os grupos de idade das mulheres, os Gráficos 12.16 e 12.17 mostram de que modo a imputação através do sistema DIA, com as conseqüentes correções das inconsistências e imputação dos dados faltantes, modificou o comportamento desses indicadores.

No caso dos "filhos tidos nascidos vivos" houve, de modo geral, pequenas reduções nos valores correspondentes às razões de sexo, quando comparados os dados criticados com os não criticados. Estes valores oscilaram em torno de 105 homens para cada 100 mulheres, o que se aproxima do padrão observado nas Estatísticas Vitais. A série das razões de sexo no nascimento de 2000 mostrou-se semelhante à de 1980, tendo 1991 um comportamento atípico. Entretanto, na variável "filhos tidos nascidos mortos", as razões finais ficaram bastante próximas das iniciais, oscilando próximo a 150 homens para cada 100 mulheres. A exceção foi do grupo de 10 a 14 anos de idade que, por ser um grupo muito rarefeito, apresentou grande variação. Na comparação com 1980 e 1991, observou-se novamente uma semelhança com os dados do Censo Demográfico 1980, sendo que para 1991, essas razões de sexo eram mais elevadas que aquelas apresentadas internacionalmente sobre natimortalidade.

Gráfico 12.16 - Razões de sexo dos filhos tidos nascidos vivos Brasil - 1980-2000

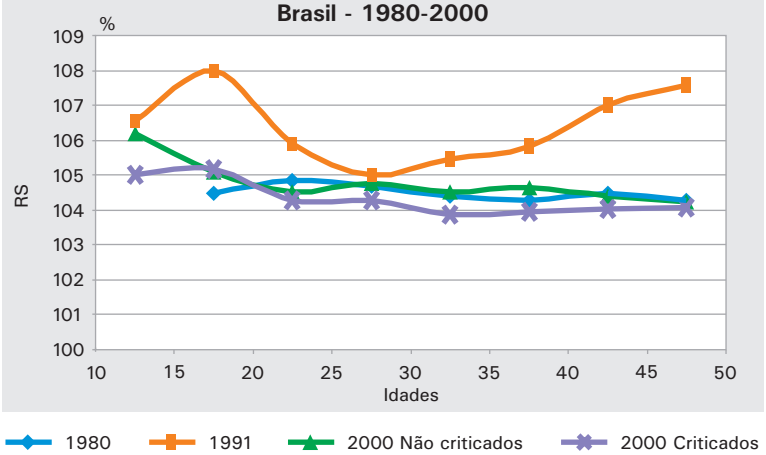
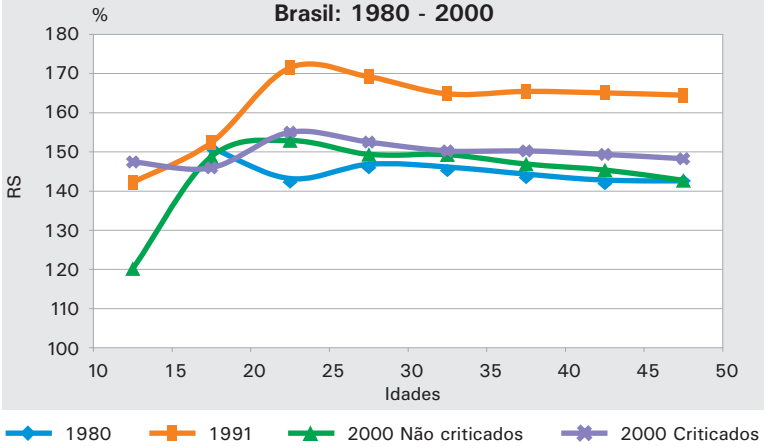


Gráfico 12.17 - Razões de sexo dos filhos tidos nascidos mortos
Brasil: 1980 - 2000



12.5.2.4 Imputação das variáveis de rendimento

Este item descreve o processo de imputação das variáveis de rendimentos dos indivíduos recenseados no Censo Demográfico 2000 através do Questionário da Amostra. As motivações que levaram ao desenvolvimento de tal processo e uma breve descrição da metodologia já estão apresentadas no item 12.3.2.7, que trata da imputação da variável de rendimento do responsável pelo domicílio, ou individual em domicílio coletivo, do Conjunto Universo. Aqui, são apresentados os aspectos específicos dessa aplicação e os resultados obtidos.

Vale lembrar que o Questionário da Amostra foi aplicado a 20% ou 10% dos domicílios de cada um dos municípios brasileiros, respectivamente, com até 15 000 e com mais de 15 000 habitantes. Nesse questionário, havia uma série de perguntas cujo objetivo era caracterizar os moradores do domicílio em 31 de Julho de 2000, data de referência do Censo Demográfico 2000. Entre essas perguntas, havia um grupo referente aos valores e às origens dos rendimentos dos moradores com 10 anos ou mais de idade, na data de referência do censo. Desse grupo de perguntas, derivou-se um conjunto de variáveis que descrevem o perfil dos rendimentos, no mês de julho de 2000, de cada um dos moradores recenseados. Essas variáveis são as seguintes:

Rendimento bruto no trabalho principal;

Rendimento bruto nos demais trabalhos;

Rendimentos provenientes de aposentadoria, pensão;

Rendimentos provenientes de aluguel;

Rendimentos provenientes de pensão alimentícia, mesada, doação recebida de não-morador;

Rendimentos provenientes de renda mínima/bolsa-escola, seguro-desemprego, etc. (programas oficiais de auxílio);

Outros rendimentos recebidos; e

Rendimento total (soma de todos os rendimentos obtidos).

A existência de valores faltantes nessas variáveis (valores não declarados) pode trazer diversos efeitos sobre análises de rendimentos a serem feitas. Daí, fez-se necessária a imputação de valores de rendimentos dos não respondentes.

Aspectos gerais

Algumas colocações a respeito dos efeitos da não-resposta em pesquisas já foram tratadas no item 12.3.2.7. Uma questão ali colocada refere-se aos possíveis efeitos de ignorar a não resposta ao se fazer inferências sobre uma variável estudada. Como resposta a essa questão, é visto que, em caso de não-resposta não diferencial, isto é, ao acaso, o efeito existente é o de aumento da variância das estimativas. Já para o caso da não-resposta diferencial, há o efeito de um impacto que se dá sob a forma de vício nas estimativas obtidas, com esse vício crescendo com a taxa de não-resposta e com a diferença entre respondentes e não respondentes.

Análises realizadas com dados do Censo Demográfico 1991 mostraram que a não-resposta nos rendimentos dos chefes de domicílio ocorria de forma diferencial em relação a algumas das variáveis presentes no questionário, brevemente descritas em 12.3.2.7.

Estudo semelhante foi feito para o projeto de Imputação de Rendimentos no Questionário da Amostra, onde a partir dos microdados utilizados para a Tabulação Avançada do Censo Demográfico 2000, verificou-se a existência de não-resposta diferencial nos quesitos de rendimento. Um exemplo disso é apresentado na Tabela 12.48, onde são observadas taxas de não-resposta do rendimento total em algumas unidades da federação, segundo algumas categorias da variável relação com o responsável pelo domicílio. Conforme pode-se verificar nessa tabela, no contigente de não respondentes do rendimento total (moradores que não responderam a uma ou mais das categorias de rendimento), há uma alta concentração na categoria "filho(a), enteado(a)", caracterizando-se assim a não resposta diferencial para o rendimento total com respeito à relação com o responsável pelo domicílio.

Tabela 12.48 – Distribuição da não resposta da variável de rendimento total, para as pessoas de 10 anos ou mais de idade, em algumas das categorias da variável relação de parentesco com o responsável pelo domicílio, para algumas Unidades da Federação

Unidades da Federação	Distribuição da não resposta da variavel rendimento total, por categoria da variável relação com o responsável pelo domicílio (%)				
	Todas as categorias	Pessoa responsável	Cônjuge, Companheiro(a)	Filho(a), Enteado(a)	Demais categorias
Pará	100	4	2	79	15
Tocantins	100	16	16	60	8
Bahia	100	1	9	74	16
São Paulo	100	5	8	75	12
Rio Grande do Sul	100	0	8	75	17
Distrito Federal	100	0	0	75	25

Fonte: IBGE, Censo Demográfico 2000, Tabulação Avançada.

São duas as alternativas para lidar com o problema da não resposta diferencial: uso de estimadores adequados para dados faltantes (LITTLE; RUBIN, 1987) e de métodos de imputação (substituição por valores estimados em cada

caso individual). No caso de não-resposta parcial de uma variável, a preferência das agências de estatísticas oficiais é geralmente por métodos baseados em imputação das variáveis não informadas. Tal preferência se deve à maior simplicidade dessa alternativa no processamento posterior dos dados, particularmente quando estes precisam ser publicados na forma de arquivos de microdados com as informações de cada respondente individual. Albieri (1992) investigou a aplicação de vários métodos para imputação da renda na Pesquisa Mensal de Emprego do IBGE.

O método desenvolvido para a imputação de rendimentos no Questionário da Amostra trabalha com a idéia de estabelecer uma relação entre rendimentos declarados pelos moradores e um grupo de variáveis do questionário da amostra, cujos valores são conhecidos para todos os moradores dos domicílios pesquisados, e, a partir dessa relação, imputar valores de rendimento para os não respondentes.

As variáveis existentes no Questionário da Amostra foram estudadas para a seleção de variáveis explicativas dos rendimentos dos moradores, em suas categorias e em seu total, buscando-se um conjunto que fosse diverso o suficiente para descrever de forma satisfatória as diferentes relações existentes ao longo do País. Infelizmente, verificou-se que relações satisfatórias entre rendimentos e variáveis do Questionário da Amostra só eram encontradas para o rendimento do trabalho principal e o rendimento total, isto é, não foram verificadas relações entre as variáveis do questionário e as demais categorias de rendimento. Com isso, para a imputação das categorias de rendimento, que não a do trabalho principal e a de total, foi adotado um procedimento baseado na relação entre o rendimento total e as variáveis selecionadas do Questionário da Amostra.

Ao contrário do ocorrido na imputação de renda dos responsáveis por domicílios ou individuais em domicílios coletivos no Conjunto Universo do Censo Demográfico 2000, na imputação de rendimentos no Questionário da Amostra não foi realizada a imputação de rendimentos nulos. Isto porque há variáveis no Questionário da Amostra que poderiam ser utilizadas para fins de predição de quem teria ou não rendimento nulo nas categorias de rendimento existentes, o que não ocorria no caso do Conjunto Universo.

A seguir, é apresentado o conjunto de variáveis selecionadas para utilização no processo de imputação, com a descrição de cada variável precedida da respectiva nomenclatura adotada:

1. COND.TRAB – associada à condição do morador no seu trabalho principal na semana de 23 a 29 de julho de 2000, possuindo as seguintes categorias: a – trabalhador doméstico com carteira de trabalho assinada; b – trabalhador doméstico sem carteira de trabalho assinada; c – empregado com carteira de trabalho assinada; d – empregado sem carteira de trabalho assinada, que não militar ou funcionário público estatutário; e – militar ou funcionário público estatutário; f – empregador; g – conta-própria; h – aprendiz ou estagiário sem remuneração; i – não remunerado em ajuda a membro do domicílio; j – trabalhador na produção para o próprio consumo;

2. GRUP.ATIV – resultante de agregação das categorias referentes à atividade principal do negócio, firma, empresa, instituição ou entidade em que o morador trabalhava na semana de 23 a 29 de julho de 2000;

3. TOT.BAN – resultante da uma combinação entre a quantidade de banheiros existentes no domicílio e a existência ou não de sanitário em domicílios sem banheiro. Assume os seguintes valores: -1, em domicílios sem banheiro e sanitário; 0, em domicílios sem banheiro e com sanitário; total de banheiros, em domicílios com a existência de 1 ou mais banheiros;

4. UTENS – resultante da existência ou não de determinados utensílios domésticos no domicílio, assumindo os seguintes valores: 0, em domicílios sem nenhum dos seguintes utensílios: videocassete, máquina de lavar roupa, forno de microondas e computador; 1, em domicílios com pelo menos um dos utensílios citados;

5. QTD.TVS – associada à quantidade de televisores existentes no domicílio, assumindo os seguintes valores: 0, em domicílios sem televisor; 1, em domicílios com 1 televisor; 2, em domicílios com 2 televisores; 3, em domicílios com 3 ou mais televisores;

6. QTD.AUTO – associada à quantidade de automóveis para uso particular existentes no domicílio, assumindo os seguintes valores: 0, em domicílios sem automóvel para uso particular; 1, em domicílios com 1 automóvel para uso particular; 2, em domicílios com 2 automóveis para uso particular; 3, em domicílios com 3 ou mais automóveis para uso particular;

7. QTD.AR – associada à quantidade de aparelhos de ar-condicionado, assumindo os seguintes valores: 0, em domicílios sem aparelho de ar-condicionado; 1, em domicílios com 1 aparelho de ar-condicionado; 2, em domicílios com 2 aparelhos de ar-condicionado; 3, em domicílios com 3 ou mais aparelhos de ar-condicionado;

8. SEXO – associada ao sexo do morador, possuindo as seguintes categorias: a - morador do sexo masculino; b - morador do sexo feminino;

9. REL.RESP.DOM – associada à relação do morador com o responsável pelo domicílio, possuindo as seguintes categorias: a – pessoa responsável; b – cônjuge, companheiro(a); c – filho(a), enteado(a); d – pai, mãe, sogro(a); e – neto(a), bisneto(a); f – irmão, irmã; g – outro parente; h – agregado(a); i – pensionista; j – empregado(a) doméstico(a); k – parente do(a) empregado(a) doméstico(a); l – individual em domicílio coletivo;

10. IDADE – idade do morador, em anos completos, em 31 de julho de 2000;

11. ANOS.EST – anos de estudo do morador;

12. IND.TRAB.PRINC – indicativa da condição do morador em relação ao rendimento proveniente de trabalho principal, possuindo as seguintes categorias: a – o rendimento proveniente de trabalho principal é zero; b – somente possui benefícios; c – possui rendimento proveniente de trabalho principal; d – não possui trabalho principal ou trabalha na produção para o próprio consumo;

13. IND.APOSENT – indicativa da condição do morador em relação ao rendimento proveniente de aposentadoria/pensão, possuindo as seguintes categorias: a – não possui rendimento proveniente de aposentadoria/pensão; b – possui rendimento proveniente de aposentadoria/pensão.

Metodologia

Conforme dito anteriormente, apenas para o rendimento total e para o rendimento do trabalho principal foi possível obter outras variáveis - covariáveis - que tivessem poder de descrição do comportamento dessas variáveis. Para as demais categorias de rendimento, o procedimento de imputação foi realizado com a utilização da árvore de regressão definida com a variável rendimento total. Assim, cada indivíduo era localizado em um estrato da árvore do rendimento total e, nesse estrato, selecionava-se o doador do rendimento não declarado. Em caso de não declaração de mais de um rendimento, as informações não eram selecionados de um único doador, pois havia o risco, considerável, deste não ser encontrado com tais rendimentos não nulos. Em caso de não haver doador para alguma categoria de rendimento em algum estrato, o doador era selecionado no subgrupo de registros que originou o estrato.

Aplicação e conclusões

Para o processo de imputação de rendimentos, utilizaram-se os mesmos lotes de registros definidos para a crítica e imputação de dados nos questionários de domicílios selecionados na amostra do Censo Demográfico 2000, realizadas com o uso do sistema DIA. Esses lotes correspondem a uma partição do conjunto de moradores dos domicílios selecionados na amostra, obedecendo aos domínios das unidades da federação (UF), isto é, um mesmo lote não contém registros de mais de uma UF. Para o processo de imputação de rendimentos, foram excluídos de cada lote os registros cujo rendimento total encontrava-se fora das cercas construídas para detectar outliers (valores atípicos). Nos 215 lotes utilizados na imputação de rendimentos, havia o total de 16.130.468 registros, com o menor deles possuindo 4.696 registros e o maior 295.074 registros. A distribuição da quantidade de lotes por UF pode ser vista na Tabela 12.16, deste capítulo.

Tal como no procedimento de imputação da variável de rendimento apurada no Conjunto Universo, para cada um dos 215 lotes de apuração dos questionários da amostra foi aplicado o procedimento de imputação baseado em árvores de regressão, descrito anteriormente. Tal como no procedimento de imputação da variável de rendimento do Conjunto Universo descrito no item 12.3, esse procedimento foi implementado com o software S-Plus e executado em ambiente operacional Windows 98. Como os lotes de registros residiam em arquivos do ambiente operacional OS/390 (mainframe IBM), foi desenvolvida, utilizando o software SAS e seus recursos para a conexão desses dois ambientes operacionais, uma rotina computacional para automatizar todo o processo de produção dessa imputação, constituído das seguintes etapas: a) preparação do arquivo de entrada para o S-Plus; b) ativação do S-Plus para a imputação propriamente dita; c) transferência dos resultados para o ambiente OS/390; e d) atualização dos registros nos lotes originais com os valores imputados.

A regra de parada na construção das árvores de regressão baseou-se no número máximo de nós terminais permitido nas árvores e no contingente populacional mínimo exigido em cada nó terminal. Visto que seria impraticável a análise dos gráficos de queda da deviance para cada uma das 215 árvores, tal como descrito no item 12.3.2.7, uma das regras de parada adotada foi a da partição de cada lote em no máximo 25 estratos. Análises preliminares com dados da Tabulação Avançada do Censo 2000, anteriormente citadas, indicaram ser este um número de nós terminais para o qual, em geral, não haveria "ganhos consideráveis" com novas partições. Quanto ao contingente populacional em cada estrato, foi estipulado que este deveria ser de no mínimo 100 pessoas.

Estatísticas referentes às taxas de não-resposta nos lotes para cada uma das categorias de rendimento e para o rendimento total são apresentadas nas Tabelas 12.49 a 12.56. As taxas de não-resposta para cada categoria de rendimento foram calculadas, levando-se em conta apenas os moradores que possuíam cada um dos rendimentos; somente para o rendimento total, foram considerados todos os moradores presentes na amostra.

Tabela 12.49 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação para rendimento no trabalho principal

Taxas de não-resposta - rendimento no trabalho principal					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,68%	1,91%	3,05%	4,15%	5,42%	18,01%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.50 - Estatísticas descritivas das daxas de não-resposta nos lotes de imputação para rendimento nos demais trabalhos

Taxas de não-resposta - rendimento nos demais trabalhos					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0%	%	%	1,65%	0,16%	5,77%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.51 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação para rendimento proveniente de aposentadoria, pensão

Taxas de não-resposta - rendimento proveniente de aposentadoria, pensão					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,87%	1,74%	2,33%	2,49%	2,96%	7,03%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.52 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação para rendimento proveniente de aluguel

Taxas de não-resposta - rendimento proveniente de aluguel					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0%	0,16%	0,43%	0,48%	0,70%	2,81%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.53 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação para rendimento proveniente de pensão alimentícia, mesada, doação recebida de não-morador

Taxas de não-resposta - rendimento proveniente de pensão alimentícia, mesada, doação recebida de não-morador					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0%	0,11%	0,38%	0,45%	0,67%	2,77%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.54 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação para rendimento proveniente de programas oficiais de auxílio

Taxas de não-resposta - rendimento proveniente de programas oficiais de auxílio					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0%	1,66%	3,03%	4,69%	5,63%	41,68%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.55 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação para outros rendimentos

Taxas de não-resposta – outros rendimentos					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0%	0,31%	0,57%	0,62%	0,82%	2,56%

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.56 - Estatísticas descritivas das taxas de não-resposta nos lotes de imputação para rendimento total

Taxas de não-resposta - rendimento total					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,59%	1,25%	1,63%	1,93%	2,43%	5,65%

Fonte: IBGE, Censo Demográfico 2000.

Os gráficos 12.18 e 12.19 apresentam os resultados da participação de cada variável explicativa nas árvores de regressão construídas para o rendimento total e para o rendimento do trabalho principal, respectivamente. Essa participação é vista sob o ponto de vista da ordem em que a variável gerou uma partição na árvore pela primeira vez. As variáveis são identificadas nos gráficos de acordo com a numeração atribuída, quando foram apresentadas no início deste item.

Gráfico 12.18 - Posições de entrada das variáveis auxiliares nas Árvores de Rendimento Total

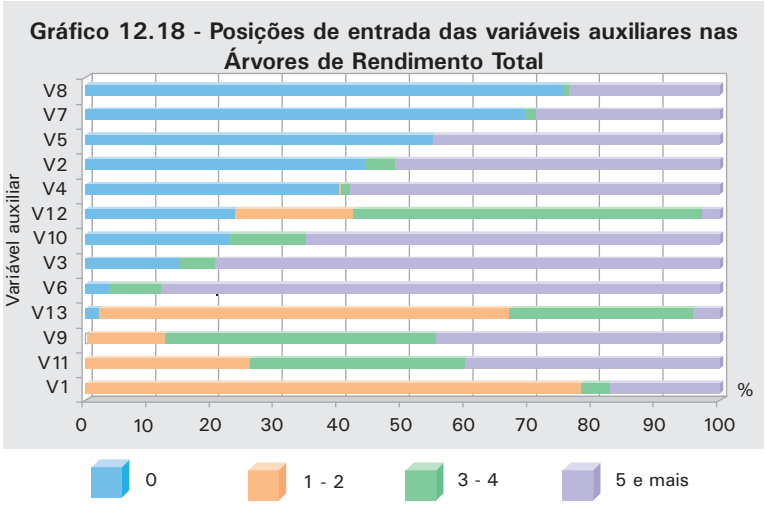
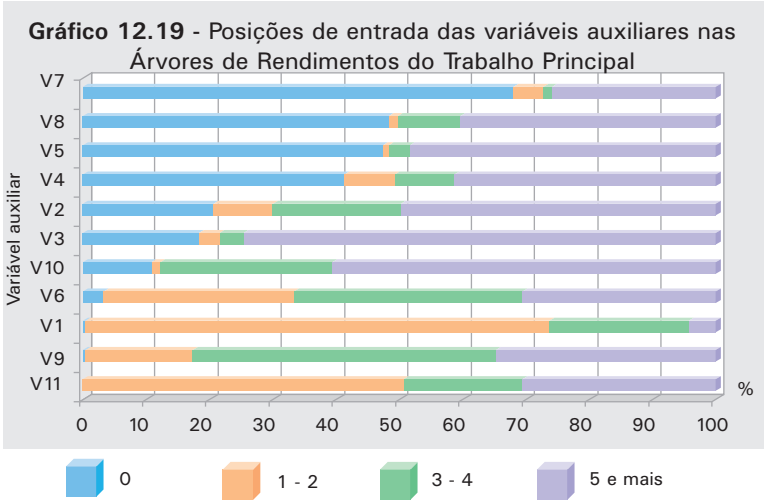


Gráfico 12.19 - Posições de entrada das variáveis auxiliares nas Árvores de Rendimentos do Trabalho Principal



Como forma de avaliar a qualidade do resultado da imputação em cada lote, foi aplicado o teste estatístico de Kolmogorov-Smirnov (LEHMANN, c1975). Este teste visa verificar se duas amostras de dados provém de uma mesma população. O teste foi aplicado para se comparar os vetores de rendimentos totais em cada subgrupo antes e depois da execução do procedimento de imputação. Cada lote só teve seu respectivo processo de imputação aprovado se o teste de Kolmogorov-Smirnov indicasse que as rendas antes e depois da imputação, em cada estrato formado, apresentavam a "mesma distribuição".

Finalizando, são apresentadas na Tabela 12.56 algumas estatísticas referentes às distribuições das taxas de imputação nos estratos obtidos nas árvores de regressão construídas para o rendimento total.

Tabela 12.57 - Estatísticas descritivas das taxas de não-resposta do rendimento total nos estratos das árvores de regressão dos lotes de imputação de rendimento no questionário da amostra

Taxas de não resposta nos estratos das árvores de regressão do rendimento total					
Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0%	1,25%	2,48%	4,72%	4,60%	100%

Fonte: IBGE, Censo Demográfico 2000.

12.5.3 Expansão da amostra

Para expansão dos dados coletados no Questionários da Amostra do Censo Demográfico 2000, foram calculados pesos para as unidades domiciliares pesquisadas, sendo tais pesos atribuídos também a cada um de seus moradores. Por unidades domiciliares pesquisadas, entende-se os domicílios particulares ocupados e as famílias e pessoas sós moradoras em domicílio coletivo.

12.5.3.1 Método para obtenção dos Pesos

O método utilizado para obtenção dos pesos foi um processo de calibração em relação a um conjunto de variáveis auxiliares (restrições) para as quais se conhecem os totais populacionais, já que tais variáveis auxiliares foram levantadas pelo Questionário Básico. A calibração buscou ajustar os pesos iniciais (inverso da

fração amostral de domicílios) de maneira que, dentro de uma determinada área geográfica, denominada área de ponderação, ao se aplicar os pesos calibrados às variáveis auxiliares, fossem obtidos os totais já conhecidos para todas as unidades da população que constituem o universo da pesquisa. Dessa maneira, além de melhorar a precisão dos estimadores, obtêm-se estimativas mais consistentes para as variáveis pesquisadas somente pelo questionário da amostra.

O cálculo dos pesos calibrados foi baseado no método dos Mínimos Quadrados Generalizados – MQG, porém com a imposição de limites nos pesos finais, para evitar pesos muito pequenos ou muito grandes. O limite mínimo utilizado foi 1, de maneira que um domicílio representasse pelo menos ele próprio. O limite máximo foi definido como cinco vezes o peso médio esperado, ou seja, 25 no caso de municípios com fração amostral planejada de 20% (caso em que o peso médio esperado era 5) e 50 no caso de municípios com fração amostral planejada de 10% (caso em que o peso médio esperado era 10). Sem a utilização desses limites, o método MQG pode gerar pesos negativos ou muito grandes, o que não teria sentido prático.

A metodologia para utilização do método MQG baseou-se em proposta de Bankier (1990) e, para sua implementação, um sistema em linguagem SAS foi desenvolvido por técnicos do IBGE.

O produto final da aplicação dessa metodologia é um peso ajustado para cada unidade domiciliar da amostra, ou seja, cada um dos questionários da amostra, que é repetido nos registros de cada pessoa moradora na unidade domiciliar.

a) Definição das variáveis auxiliares

A escolha das variáveis auxiliares cujos valores são utilizados como restrições no processo de ajustamento do qual decorrem os pesos é um aspecto importante do método aplicado. A forma ou prioridade de tratamento dessas variáveis, sobretudo, quando não existe uma solução que atenda simultaneamente a todas as restrições, é outro ponto sensível do método.

As variáveis auxiliares constituem um subconjunto das variáveis comuns à amostra e ao universo e são referentes a características de domicílios ou de pessoas, apesar do ajustamento ser realizado de forma a fornecer pesos para cada uma das unidades domiciliares.

A metodologia de ajuste de um modelo linear generalizado multivariado envolve cálculos com matrizes, inclusive inversão. Por essa razão, as restrições definidas que, por sua vez, dão origem a uma dessas matrizes, devem satisfazer algumas condições essenciais, sendo a principal delas a de não serem linearmente dependentes (redundantes). Além disso, é também considerado o conceito de restrições quase linearmente dependentes (e, portanto, quase redundantes), que afetam a estabilidade da solução do modelo.

Outras duas condições impostas para a aplicação dessa metodologia referem-se à sua significância estatística. O tamanho da restrição, medido como o número de domicílios aos quais a restrição se aplica em uma dada área de ponderação, não deve ser muito pequeno sob pena de tornar instável o processo de estimação. Quando uma restrição não atinge um número mínimo de unidades domiciliares, fixado em função da fração de amostragem, essa restrição é considerada rara.

Além disso, uma restrição definida pode causar a obtenção de um peso muito grande ou muito pequeno, quando comparado com o peso médio esperado (5 ou 10) em função da fração amostral adotada na área de ponderação ou até um peso negativo, constituindo-se em restrição geradora de peso extremo.

Dessa forma, o programa de ajuste do modelo incorpora procedimentos de eliminação de restrições que se enquadrem nas condições acima, observando a ordem que segue: restrições raras, restrições redundantes, restrições quase redundantes e restrições responsáveis por pesos extremos.

Convém ressaltar que a eliminação de restrições pode implicar diretamente no fato de não se ter a garantia da calibração desejada para as variáveis eliminadas para a presente área de ponderação.

As restrições inicialmente definidas para a aplicação da metodologia MQG, para cada uma das áreas de ponderação, encontram-se na relação abaixo. Elas constituem o conjunto denominado conjunto 1 de restrições para calibração.

Em unidades domiciliares (domicílios particulares ocupados + famílias ou pessoas sós em domicílios coletivos)

1. Número total de pessoas
2. Número total de unidades domiciliares
3. Número de pessoas do sexo masculino
4. Número de pessoas na faixa de idade de 0 a 4 anos
5. Número de pessoas na faixa de idade de 5 a 9 anos
6. Número de pessoas na faixa de idade de 10 a 14 anos
7. Número de pessoas na faixa de idade de 15 e 19 anos
8. Número de pessoas na faixa de idade de 20 a 24 anos
9. Número de pessoas na faixa de idade de 25 a 29 anos
10. Número de pessoas na faixa de idade de 30 a 34 anos
11. Número de pessoas na faixa de idade de 35 a 39 anos
12. Número de pessoas na faixa de idade de 40 a 44 anos
13. Número de pessoas na faixa de idade de 45 a 49 anos
14. Número de pessoas na faixa de idade de 50 a 59 anos
15. Número de pessoas na faixa de idade de 60 a 69 anos
16. Número de pessoas na faixa de idade de 70 anos ou mais
17. Número de pessoas do sexo masculino na faixa de idade de 0 a 4 anos
18. Número de pessoas do sexo masculino na faixa de idade de 5 a 9 anos
19. Número de pessoas do sexo masculino na faixa de idade de 10 a 14 anos
20. Número de pessoas do sexo masculino na faixa de idade de 15 a 19 anos
21. Número de pessoas do sexo masculino na faixa de idade de 20 a 24 anos

22. Número de pessoas do sexo masculino na faixa de idade de 25 a 29 anos

23. Número de pessoas do sexo masculino na faixa de idade de 30 a 34 anos

24. Número de pessoas do sexo masculino na faixa de idade de 35 a 39 anos

25. Número de pessoas do sexo masculino na faixa de idade de 40 a 44 anos

26. Número de pessoas do sexo masculino na faixa de idade de 45 a 49 anos

27. Número de pessoas do sexo masculino na faixa de idade de 50 a 59 anos

28. Número de pessoas do sexo masculino na faixa de idade de 60 anos ou mais

29. Número de pessoas moradoras na situação urbana

30. Número de pessoas do sexo feminino moradoras na situação urbana

31. Número de pessoas do sexo feminino moradoras na situação rural

Em domicílios particulares permanentes ocupados

32. Número de pessoas do sexo masculino que são chefes ou individuais

33. Número total de pessoas

34. Número total de domicílios

35. Número de domicílios urbanos

36. Número de domicílios com 1 ou 2 moradores

37. Número de domicílios com 3 moradores

38. Número de domicílios com 4 moradores

39. Número de domicílios com 5 moradores

40. Número de domicílios com 6 ou mais moradores

b) Análise da qualidade da calibração

As restrições, acima apresentadas, foram agrupadas em dez conjuntos alternativos que foram utilizados em ordem de prioridade. Esse procedimento foi adotado para garantir que alguma calibração fosse feita, mesmo que em um conjunto menor de características. O primeiro conjunto foi formado por todas as restrições, como listadas, e os demais formados pela agregação de faixas etárias, agregação de faixas de moradores por domicílio ou mesmo a retirada de grupos de restrições.

No cálculo dos pesos calibrados, para cada área de ponderação, foi utilizado inicialmente o conjunto de restrições número 1. Quando não se obteve uma solução satisfatória, a área foi processada novamente, utilizando o conjunto 2 e assim sucessivamente até o conjunto 10, caso anteriormente não tenha sido atingida a qualidade de ajuste adequada. A composição de cada um dos 9 conjuntos alternativos de restrições está apresentada em anexo no CD-ROM.

A análise da qualidade do ajuste (calibração) era feita automaticamente pelo sistema através das diferenças entre os valores populacionais conhecidos para as restrições e os valores estimados utilizando-se os pesos calculados. Para cada grupo de restrições, foram definidos limites específicos tolerados para essas diferenças.

Para as áreas de ponderação onde não ocorreu o ajuste para nenhum dos dez conjuntos de restrições, o sistema automaticamente escolheu o conjunto que proporcionou o melhor ajuste, no sentido de minimizar a soma dos quadrados das diferenças entre o valor conhecido das restrições e o valor estimado para essas mesmas restrições. Essa estatística foi calculada com base nas variáveis do conjunto 1 de restrições, para todos os dez conjuntos avaliados.

Deve-se ressaltar que para algumas áreas de ponderação onde houve um desequilíbrio forte entre a fração amostral de domicílios e a fração amostral de pessoas, ou seja, em áreas onde o número médio de pessoas por domicílio no universo e na amostra diferiram muito, pode ter ocorrido falta de ajuste na variável total de pessoas. Portanto, é possível a ocorrência, para alguns municípios, de divergência entre o valor do número de pessoas calculado através da expansão da amostra e o valor verificado na investigação do universo dos domicílios, que são os números oficiais do censo.

Do total de 9 336 áreas de ponderação definidas para o Brasil, mais de 91% teve solução para o conjunto 1 de restrições.

No final do processo, foi garantido que pelo menos a restrição "número total de domicílios" fosse respeitada para todas as áreas de ponderação. Nesse contexto, "número total de domicílios" iguala o número total de questionários e engloba os domicílios particulares ocupados mais as famílias e pessoas sós moradoras em domicílios coletivos.

12.5.3.2 Áreas de ponderação

Define-se Área de Ponderação como sendo uma unidade geográfica, formada por um agrupamento mutuamente exclusivo de setores censitários, para a aplicação dos procedimentos de calibração das estimativas com as informações conhecidas para a população como um todo.

Foram definidas, para todo o Brasil, 9 336 áreas de ponderação e, tal como nos censos anteriores, a metodologia de expansão da amostra foi aplicada independentemente para cada uma delas.

O tamanho dessas áreas, em termos de número de domicílios e de população, não pode ser muito reduzido, sob pena de perda de precisão de suas estimativas. As áreas de ponderação foram definidas considerando essa condição e, também, os níveis geográficos mais detalhados da base operacional, como forma de atender a demandas por informações em níveis geográficos menores que os municípios.

Os livros técnicos de amostragem definem procedimentos para a determinação de tamanhos de amostra considerando os requisitos de precisão estabelecidos para uma pesquisa. Para tanto, define-se a margem de erro aceitável para o estimador amostral, supondo que a amostra seria selecionada sob Amostragem Aleatória Simples (AAS). Considerando o objetivo de estimar uma média com um erro máximo relativo de kr ao nível de confiança de 95%, tem-se uma equação que relaciona o tamanho total da amostra desejada com os requisitos de precisão especificados.

A definição do tamanho das áreas de ponderação para o cálculo dos pesos de expansão da amostra do Censo demográfico 2000 foi feita considerando questões técnicas estreitamente relacionadas com as acima descritas. O tamanho mínimo definido para uma área de ponderação foi estabelecido em 400 domicílios particulares ocupados na amostra, por ser um valor aproximado ao encontrado nos cálculos de tamanho de amostra aleatória simples, quando se considera a intenção de estimar uma proporção (pequena) de 5%, com uma precisão relativa máxima fixada não muito exigente (40%) em uma população considerada grande, para os efeitos de aproximação nas fórmulas, e considerando um nível de confiança estabelecido em 95%, para a construção de intervalos de confiança.

A decisão de fixar o tamanho da amostra e não o tamanho da população da área de ponderação foi tomada com base no fato que a precisão de estimativas provenientes de pesquisas por amostragem está diretamente relacionada com o tamanho absoluto da amostra e não com a fração amostral (relação entre tamanho de amostra e tamanho da população). Assim, nos municípios onde foi decidido que seriam definidas áreas de ponderação em nível geográfico mais desagregado que o próprio município, foi considerada essa restrição de tamanho, com o objetivo de preservar a precisão de estimativas. Nos municípios onde foi considerada apenas uma área, o próprio município, a restrição de tamanho não pode ser aplicada, pois o tamanho da amostra foi uma decorrência da fração amostral definida antes da realização do censo. Nos municípios pequenos em que, em função da definição da fração amostral, o tamanho da amostra de domicílios resultou em valores menores que 400 unidades, é possível que um número significativo de estimativas tenha baixa precisão, medida em termos de erro amostral. Albieri (2003) apresenta mais considerações sobre essa definição e suas implicações.

Para o Censo 2000, foram usados métodos e sistemas automáticos de formação de áreas de ponderação que conjugam critérios tais como tamanho (para permitir estimativas com qualidade estatística em áreas pequenas), contigüidade (no sentido de serem constituídas por conjuntos de setores limítrofes com sentido geográfico) e homogeneidade em relação a um conjunto de características populacionais e de infra-estrutura conhecidas.

As áreas de ponderação foram criadas, considerando os seguintes critérios:

- maior nível geográfico utilizado é o município; isto significa que uma área de ponderação é composta por setores censitários dentro de um único município, podendo ser o próprio município;
- menor tamanho de uma área de ponderação não municipal é de 400 domicílios particulares ocupados na amostra;
- em alguns municípios, as áreas de ponderação foram definidas considerando suas divisões administrativas, sempre respeitando o critério de tamanho mínimo; alguns municípios tiveram apenas 2 áreas definidas: uma considerando todos os setores do distrito-sede e outra considerando todos os setores dos demais distritos; em outros municípios, cujos distritos possuem tamanho que feriam o critério de tamanho mínimo, também foram definidas duas áreas: uma constituída por todos os seus setores urbanos e outra por todos os seus setores rurais, mesmo que isso significasse setores não contíguos;

- para um conjunto de municípios grandes em termos de população, foi feita uma consulta aos órgãos de planejamento municipal para que as áreas de ponderação fossem definidas em conjunto. Nesses municípios, também foram considerados os critérios de tamanho mínimo e de contiguidade do conjunto de setores para a definição das áreas de ponderação; e
- os municípios que não se enquadraram nas 4 situações acima tiveram suas áreas de ponderação definidas automaticamente, usando uma metodologia de agregação de setores, implementada por meio de um sistema computacional especialmente desenvolvido, que faz uso de informações georreferenciadas; essa metodologia considera os critérios de tamanho mínimo, vizinhança entre os setores e a homogeneidade dos setores em relação a um conjunto de características conhecidas para o universo no nível dos setores. Entre as 15 variáveis utilizadas constava, por exemplo: rendimento médio dos responsáveis pelos domicílios no setor, número médio de pessoas por domicílio particular permanente, proporção de domicílios particulares permanentes ligados à rede geral de água, média de anos de estudo dos responsáveis por domicílios. Para uma descrição detalhada do procedimento, ver Silva, Matzenbacher e Cortez (2002).

A divulgação dos resultados da amostra nos diversos formatos, publicações de tabelas, CD-ROM, microdados, dados agregados em nível de área de ponderação e em nível de município, contém em sua documentação as informações para a compreensão de como resultaram as áreas de ponderação consideradas durante o processo de expansão da amostra. Essa documentação inclui uma relação dos 484 municípios que tiveram mais de uma área de ponderação com informações sobre o número de suas áreas. Os demais 5 023 municípios tiveram apenas uma área de ponderação. Além disso, inclui um arquivo com informações básicas sobre cada uma das 9.336 áreas de ponderação, a saber: código da área de ponderação; tipo da área; número de setores; número de domicílios particulares ocupados na amostra; número de pessoas no universo; e uma descrição da composição geográfica da área de ponderação.

Outro arquivo que faz parte da documentação relacionada com a definição das áreas de ponderação é o que indica o código da área de ponderação (13 posições) para cada setor censitário da base geográfica do Censo 2000, identificado também pelo seu código (15 posições).

A tabela 12.59 ao final deste item apresenta o número de municípios total e por tipo de área que contém, por Unidade da Federação.

12.5.3.3 Estimação de totais, médias e razões

As estimações de totais para domínios de interesse, como por exemplo, as células de uma tabela, são feitas, utilizando-se, para cada unidade (pessoa, família ou domicílio), o peso correspondente, que foi determinado para cada domicílio da amostra. Esse mesmo peso foi atribuído a cada pessoa moradora e a cada família do domicílio. Assim, para estimar o total de uma característica utiliza-se o estimador $\hat{Y}$ definido por:

$$\hat{Y} = \sum_{i=1}^{n} p_i \, y_i$$

onde:

p_i é o peso associado à i-ésima unidade da amostra no domínio em questão;

y_i é o valor de y associado à i-ésima unidade da amostra no domínio;

n é o número de unidades na amostra do domínio em questão.

Dessa forma, é possível calcular estimativas para quaisquer variáveis investigadas no censo, independente de serem de pessoas, famílias ou domicílios.

Os pesos calculados com a metodologia adotada não são necessariamente inteiros e não devem ser substituídos por pesos inteiros para não provocar a quebra na consistência das restrições efetivamente utilizadas no ajuste no modelo. O uso de pesos fracionários preserva o método de expansão da amostra, produz resultados mais precisos do ponto de vista estatístico. Assim, para o cálculo das estimativas das tabelas de divulgação do censo foi utilizado o peso fracionário com 8 casas decimais, sendo, então, arredondadas as estimativas resultantes.

Para obter consistência com as tabelas de divulgação, é necessário que as estimativas sejam calculadas em cada célula básica da tabela e as linhas e colunas de totais e subtotais sejam obtidas por soma das estimativas básicas correspondentes, após terem sido arredondadas. Uma conseqüência desse procedimento é que os totais de uma mesma característica podem diferir ligeiramente de uma tabela para outra, em função do arredondamento das parcelas em cada tabela.

12.5.3.4 Estimação de erros amostrais ou avaliação da orecisão das estimativas

As conclusões de uma pesquisa por amostra devem ser apoiadas nas estimativas produzidas. Essas, por sua vez, embutem um erro amostral que deve situar-se dentro de um nível de confiança fixado. Assim, a avaliação dos erros amostrais é um ponto fundamental, pois dele decorre o grau de confiança nas conclusões analíticas obtidas. Para cada estimativa derivada da pesquisa, é possível obter uma medida de precisão que auxilia na análise e interpretação dos dados resultantes da pesquisa.

Os erros amostrais podem ser avaliados através das estimativas dos coeficientes de variação ou dos erros padrão calculadas a partir das estimativas das variâncias.

Embora seja possível estimar os erros amostrais de acordo com a metodologia usada na obtenção dos pesos, o método direto é bastante complexo (SÄRNDAL; SWENSSON; WRETMAN, c1992). Sugere-se, então, um método simples e rápido para obtenção de uma aproximação do erro padrão da estimativa, que pode ser usado para a construção de intervalos com níveis de confiança fixados. Como a amostra usada no Censo Demográfico 2000 é bastante grande e os domicílios se distribuem de forma aleatória dentro de cada setor censitário, pode-se aproximar o cálculo do erro padrão, segundo Cochran (1977), supondo que o esquema de seleção da amostra foi de amostragem aleatória simples sem reposição. Dessa maneira, um estimador do erro padrão de um estimador de total de uma característica y, representado por $\hat{Y}$, é dado por:

$$ep(\hat{Y}) = \sqrt{\frac{(1-f)}{f} N s^2(y)}$$

onde:

$ep(\hat{Y})$ é o erro padrão do estimador de total, $\hat{Y}$, para o domínio em questão;

f é a fração efetiva de amostragem observada no domínio em questão;

N é o total de unidades da população no domínio em questão;

$s^2(y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ é a variância amostral para o domínio em questão;

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ é a média amostral no domínio em questão;

y_i é o valor da característica y na i-ésima unidade da amostra no domínio;

n é o total de unidades da amostra no domínio em questão.

A divulgação dos resultados da amostra nos diversos formatos, publicações de tabelas, CD-ROM, microdados, dados agregados em nível de área de ponderação e em nível de município, contém em sua documentação um arquivo ou tabela com as frações amostrais, em porcentagem, efetivamente observadas para os domínios Brasil, Grandes Regiões, Unidades da Federação, Mesorregiões, Microrregiões e Municípios. A Tabela 12.57, a seguir, mostra o número de municípios com cada uma das duas frações efetiva aplicadas, por tamanho do município, medido em população recenseada no Censo 2000.

Tabela 12.58 - Número de municípios por fração amostral efetiva de domicílios, segundo a classe de tamanho populacional

Classes de população em 2000	Total	Fração amostral efetiva de domicílios	
		até 15%	+ de 15%
Total	5 507	2 020	3 487
Até 15 000 habitantes	3 540	114	3 426
Mais de 15 000 habitantes	1 967	1 906	61

Fonte: IBGE, Censo Demográfico 2000.

Como a maior parte das estimativas derivadas das informações coletadas na amostra do Censo Demográfico 2000 é proveniente de variáveis categóricas, para as quais y_i assume somente os valores 0 (se a unidade não pertence à categoria em questão), ou 1 (se a unidade pertence à categoria em questão), a expressão do estimador $ep(\hat{Y})$ reduz-se a:

$$ep(\hat{Y}) = \sqrt{\frac{(1-f)\hat{Y}(N-\hat{Y})}{Nf-1}}$$

Na Tabela 12.58, são apresentados valores de erros padrão calculados para alguns valores de estimativas de características de pessoas e domicílios para o Brasil.

O erro padrão é utilizado para construir intervalos de confiança que conterão o valor do total populacional[7], y_i com uma certa probabilidade decorrente do nível de confiança desejado na tomada de decisão, ou seja,

$$P[\hat{Y} - z_{\alpha/2}ep(\hat{Y}) < Y < \hat{Y} + z_{\alpha/2}ep(\hat{Y})] = 1-\alpha$$

onde:

α é o nível de significância e (1 - a) é o nível de confiança;

$z_{\alpha/2}$ é a abscissa da distribuição Normal padrão com área $\alpha/2$ à sua direita.

Assim, para um nível de confiança de 95%, tem-se $z_{\alpha/2} = 1{,}96$ e o intervalo de confiança é dado por:

$$[\hat{Y} - 1{,}96ep(\hat{Y});\ \hat{Y} + 1{,}96ep(\hat{Y})]$$

Pela Tabela 12.58, caso haja interesse em estimar um total de uma característica relativa às pessoas e essa estimativa para Brasil seja da ordem de 10 000 000, vê-se que seu erro padrão seria da ordem de 8 445. Portanto, de acordo com as fórmulas anteriores, um intervalo de 95% de confiança para o total da característica de interesse será dado por [9 983 448; 10 016 552]. Em termos percentuais, pode-se dizer que a estimativa da característica desejada é 10 000 000, com uma margem de erro relativo de 0,17% para cima ou para baixo.

Na prática, um intervalo de confiança de 95%, por exemplo, indica que, em cada 100 amostras selecionadas com o mesmo desenho, 95 produzirão estimativas cujo intervalo de confiança conterá o valor verdadeiro da população e em apenas 5 amostras este valor estará fora do intervalo de confiança.

Naturalmente, quanto maior o nível de confiança, maior será a amplitude do intervalo de confiança. A decisão sobre o nível de confiança decorre do grau de precisão que o usuário necessita em seu trabalho analítico.

A divulgação dos resultados da amostra nos diversos formatos, publicações de tabelas, CD-ROM, microdados, dados agregados em nível de área de ponderação e em nível de município, contém em sua documentação tabelas equivalentes à Tabela 12.58, para outros níveis geográficos, a saber, as 5 Grandes Regiões e as 27 Unidades da Federação:

[7] O valor da população é, de um modo geral, desconhecido, exceto para as características investigadas censitariamente.

Tabela 12.59 - Brasil - Erro padrão aproximado para alguns tamanhos de estimativas para características de pessoas e domicílios

Características de pessoas		Características de domicílios	
Tamanho da estimativa	Erro padrão aproximado	Tamanho da estimativa	Erro padrão aproximado
100	28	100	28
500	62	500	62
1 000	87	1 000	87
2 000	123	2 000	123
5 000	195	5 000	195
10 000	275	10 000	275
20 000	389	20 000	389
50 000	615	50 000	615
100 000	870	100 000	870
150 000	1 066	150 000	1 064
200 000	1 230	200 000	1 228
500 000	1 944	250 000	1 373
1 000 000	2 745	500 000	1 936
2 000 000	3 870	1 000 000	2 722
3 000 000	4 726	2 000 000	3 807
4 000 000	5 440	3 000 000	4 608
5 000 000	6 064	4 000 000	5 258
6 000 000	6 623	5 000 000	5 808
7 000 000	7 132	6 000 000	6 283
8 000 000	7 601	7 000 000	6 700
9 000 000	8 037	8 000 000	7 069
10 000 000	8 445	9 000 000	7 397
15 000 000	10 180	10 000 000	7 690
20 000 000	11 563	15 000 000	8 730
30 000 000	13 681	20 000 000	9 217
40 000 000	15 222	25 000 000	9 240
50 000 000	16 350	30 000 000	8 802
100 000 000	17 650	35 000 000	7 826
120 000 000	16 331	40 000 000	6 057
130 000 000	15 196	45 507 516	0
140 000 000	13 645		
150 000 000	11 513		
160 000 000	8 365		
169 799 170	0		

Fonte: IBGE, Censo Demográfico 2000.

Tabela 12.60 - Número de municípios total e por tipo de área que contém, por Unidade da Federação

Unidade da Federação	Quesito							
	Número de municípios existentes	Município	usuário	Distrito	Distrito-sede + Ag. Distritos[1]	Urbana + rural	Subdistrito + Ag. Subdistritos	Municípios feitos no skater
Total	**5507**	**5023**	**69**	**9**	**79**	**128**	**1**	**199**
Rondônia	52	46	1	1	0	4	0	0
Acre	22	19	0	0	0	2	0	1
Amazonas	62	57	1	0	0	4	0	0
Roraima	15	14	1	0	0	0	0	0
Pará	143	126	3	0	3	10	0	1
Amapá	16	15	1	0	0	0	0	0
Tocantins	139	137	1	0	0	0	0	1
Maranhão	217	201	3	1	0	12	0	0
Piauí	221	217	1	0	0	2	0	1
Ceará	184	147	3	0	25	8	0	1
Rio Grande do Norte	166	159	2	0	1	3	0	1
Paraíba	223	216	2	0	0	2	0	3
Pernambuco	185	155	5	0	6	13	0	6
Alagoas	101	94	1	0	0	6	0	0
Sergipe	75	69	1	0	0	4	0	1
Bahia	415	375	5	1	8	21	0	5
Minas Gerais	853	806	1	2	9	6	0	29
Espírito Santo	77	61	3	0	7	4	0	2
Rio de Janeiro	91	54	8	0	11	2	1	16
São Paulo	645	553	10	2	3	7	0	70
Paraná	399	370	3	0	2	6	0	18
Santa Catarina	293	275	3	2	1	4	0	8
Rio Grande do Sul	467	432	3	0	3	5	0	24
Mato Grande do Sul	77	72	1	0	0	2	0	2
Mato Grosso	126	122	2	0	0	1	0	1
Goiás	242	231	3	0	0	0	0	8
Distrito Federal	1	0	1	0	0	0	0	0

Fonte: IBGE, Censo Demográfico 2000.

(1) De fato, são 78 municípios nessa situação e um município (Queimados no Rio de Janeiro) que teve suas áreas definidas ou como subdistritos inteiros ou como agregados de subdistritos. Esse município foi classificado nessa categoria por ter sido o único caso nesse formato.

12.6 Tabulação dos dados

Os procedimentos de tabulação dos dados constituem a última etapa do processo de apuração das informações. No Censo Demográfico 2000, a tabulação teve início logo após o encerramento da coleta, com os Resultados Preliminares, continuou com a Sinopse Preliminar, alcançando, mais tarde, os resultados referentes ao Conjunto Universo e ao Questionário da Amostra.

Com exceção das tabulações dos Resultados Preliminares e da Sinopse Preliminar, que tiveram como base os dados do SIGC, ainda não submetidos ao processo de crítica, as demais tiveram seu início imediatamente após a realização desse trabalho de depuração de erros.

Independentemente do tipo de publicação, o processo de tabulação tem como referência uma proposta de plano tabular, discutida no âmbito do Comitê do Censo Demográfico 2000 e com os principais usuários externos das informações censitárias.

O início efetivo do processo de tabulação aconteceu com a confecção das molduras das tabelas, que corresponde à elaboração dos textos referentes aos títulos e rodapés, bem como os que constituem as indicações das colunas de dados e coluna indicadora de cada tabela. Esse trabalho foi desenvolvido pelo Centro de Documentação e Disseminação de Informações – CDDI, obedecendo-se às *Normas de apresentação tabular* (1993), para, mais tarde, receber a aprovação da Coordenação de Planejamento e Organização do Censo Demográfico 2000 – CPO.

O passo seguinte foi a elaboração da seleção de variáveis, ocasião em que técnicos da Diretoria de Pesquisas - DPE organizaram, para cada uma das tabelas do plano, as formas de obtenção dos seus dados, identificando o manejo dos códigos ou dos valores das variáveis coletadas nos questionários ou criadas durante o processo de crítica das informações.

O trabalho continuou com a entrega das molduras e respectivas seleções de variáveis a dois grupos de técnicos, um da Diretoria de Informática - DI e outro da DPE, para a obtenção dos dados. Para esse trabalho de dupla programação, a DI utilizou o Sistema Pegasus e a DPE o Sistema SAS.

Sistema Pegasus de tabulação

O Pegasus, sistema de tabulação utilizado no Censo 2000, foi desenvolvido em Visual Basic- VB-v6, FrontPage e Opus, sendo ativado via Internet Explorer 4.0. Assim, os diálogos foram implementados em VB-v6, as páginas *Web* desenvolvidas no FrontPage, as rotinas de tabulação em Opus e a visualização das tabelas em Excel 7.0. O sistema é composto por:

- banco de dados em um servidor RISC, que armazena a definição dos planos de tabulação, os dicionários e também as matrizes com os valores tabulados, que são visualizados, via browser, de qualquer microcomputador ligado à rede do IBGE com a configuração necessária para acesso ao Pegasus – Windows 98, Internet Explorer 4.0, MS Office 97; e

- rotinas com os diálogos, residentes em uma máquina Windows NT com um servidor *Web*, e que são instaladas na máquina do operador automaticamente pelo Internet Explorer, quando conectado ao sistema pela primeira vez;

O funcionamento do Pegasus se inicia quando a página *Web* de apresentação é chamada por um operador. Nela existem os links para funções de auxílio, para os manuais e para o sistema

O processo de tabulação dos microdados é bastante rápido, feito por um programa OPUS, residente em um servidor Unix que, para cada plano tabular, faz a leitura dos dados uma única vez.

Sistema de conferência

Um sistema de conferência, operando *on-line*, baseado em macros do Excel, implementava a comparação das tabelas geradas pelos Sistemas Pegasus e SAS. Nessa oportunidade, todos os elementos das tabelas eram cotejados, sendo qualquer diferença, entre seus textos ou valores, apontada pela rotina ao técnico operador.

Elegia-se para teste uma Unidade da Federação, para a qual todas as tabelas do plano passavam pelo processo de conferência. Somente após sanados as divergências, era gerado o plano tabular para todas as outras unidades e demais níveis de divulgação.

Deve-se ressaltar a importância que o sistema de conferência implementado representou para a velocidade e segurança de todo o processo de tabulação dos dados do censo, sendo que a rotina foi aperfeiçoada ao longo dos trabalhos para as diversas divulgações realizadas, tendo o início desse processo acontecido com a Sinopse Preliminar.

Referências

ALBIERI, S. *Apresentação da precisão de estimativas nas tabelas de pesquisas por amostragem do IBGE*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1999a.

_____. *A ausência de resposta em pesquisas:* uma aplicação de métodos de imputação. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, 1992. 138 p. (Informes de matemática. Série D-048/92). Dissertação de mestrado apresentada em 1989.

_____. *Considerações preliminares para o planejamento da amostra para a tabulação avançada do censo demográfico 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1999b.

_____. *Nota técnica sobre a definição do tamanho das áreas de ponderação do censo demográfico 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 7 p.

_____; DIAS, A. J. R.; MENEZES, A. C. F.; GREEN, A. P. L. *Controle de qualidade da captura de dados do censo 2000:* especificações para o planejamento das amostras: (2ª versão). Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2001.

_____; MARTELOTTE, M. C.; DUARTE, R. P. N. *Estudos para subsidiar o planejamento da amostra para a tabulação avançada do censo demográfico 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2000.

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

BANKIER, M. D. et al. Imputing numeric and qualitative variables simultaneously. *Proceedings of the Survey Research Methods Section*, Baltimore, 1996, p.90-99, [1996?].

BARBOSA, D. M. R.; SILVA, A. do N. *NIM - new imputation methodology.* Rio de Janeiro: IBGE, Diretoria de Informática, 2002. (Nota Técnica, 02/02).

BRASS, W. *Seminário sobre métodos para medir variables demográficas:* fecundidad y mortalidad, 1971, São José. San José: CELADE, 1973. 146 p. (Serie DS CELADE, n. 9).

_____ et al. *The demography of tropical Africa.* Princeton: Princeton University Press, 1968.

_____. Estimating mortality from deficient registration data. In: _____. *Methods for estimating fertility and mortality from limited and defective data*. Chapel Hill: University of North Carolina, 1975. p.117-123. (Laboratories for Population Statistics. An occasional publication).

BRAVO, P. C. Elementos de controle estatístico de qualidade. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 6., 1984, Rio de Janeiro. *Atas...* Rio de Janeiro: UFRJ, 1984. 585 p.

BREIMAN, L. et al. *Classification and regression trees*. Belmont, Calif.: Wadsworth International Group, c1984. (The Wadsworth statistics/probability series).

CAMISA, Z. Fecundidad y nupcialidad. In: ENCUESTA demográfica nacional de Honduras. [Tegucigalpa]: Dirección General de Estadística y Censos de Honduras; Santiago de Chile: CELADE, 1975. v. 3. (Série A. CELADE, n. 129).

CENSO DEMOGRÁFICO 2000: características da população e dos domicílios: resultados do universo. Rio de Janeiro: IBGE, 2001.

_____: fecundidade e mortalidade infantil: resultados preliminares da amostra. Rio de Janeiro: IBGE, 2002. 21 p.

_____: resultados preliminares. Rio de Janeiro: IBGE, 2000. 172 p.

CENSO demográfico 2000: manual do recenseador CD 1.09. Rio de Janeiro: IBGE, 2000. 151 p.

CENTRY user's guide: IMPS version 3.1. Washington, D.C.: Bureau of the Census International Systems Team, 1995.

COALE, A. J.; DEMENY, P. *Regional model life tables and stable populations*. Princeton, N.J.: Princeton University Press, 1966. 871 p.

_____; TRUSSELL, J. Estimating the time to which Brass estimates Apply, annex 1 to Samuel H. Preston and Alberto Palloni, fine-tuning Brass-type mortality estimates with data on ages of surviving children. *Population Bulletin of the United Nations*, New York, n. 10, p. 87-89, 1977.

_____. Model fertility schedules: variations in the age structure of childbearing in human populations. *Population Index*, Princeton, v. 40, n. 2, p. 185-257, 1974.

COCHRAN, W. G. *Sampling techniques*. 3rd ed. New York: Wiley, c1977. 428 p.

CONCOR user's guide: IMPS version 3.1. Washington, D.C.: Bureau of the Census International Systems Team, 1995.

CORREÇÃO de declaração de dados de fecundidade e mortalidade. [Rio de Janeiro: IBGE, 2002]. 4 p.

DEFINIÇÕES necessárias à implementação da apuração centralizada dos dados referentes ao questionário da amostra (CD 1.02) no censo demográfico de 2000. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 78 p.

DIAS, A. J. R. *Aspectos de amostragem do censo demográfico de 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2002. 17 p.

ESQUEMA de ordenação lógica das pessoas no questionário básico - censo 2000. Rio de Janeiro: IBGE, 2001.

FEENEY, G. Estimating infant mortality rates from child survivorship data by age of mother. *Asian and Pacific Census Newsletter*, v. 3, n. 2, p. 12-16, Nov. 1976.

_____. Estimating infant mortality trends from child survivorship data. *Population Studies*, London, v. 34, n. 1, p. 109-128, Mar. 1980.

FRIAS, L. A. de M.; RODRIGUES, P. *Brasil*: tábuas-modelo de mortalidade e populações estáveis. Rio de Janeiro: IBGE, 1981. 149 p. (Estudos e pesquisas, n. 10).

GENERALIZED *estimation system*: version 4.0: help guide. Ottawa: Statistics Canada, 1998.

GUIMARÃES, N. R. *Controle de qualidade do censo 2000:* análise final dos processos de reconhecimento e verificação. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2001. 54 p.

_____; DIAS, A.; ALBIERI, S. *Censo demográfico 2000:* controle de qualidade do processo de captura dos dados. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2001.

HRADESKY, J. L. *Productivity and quality improvement*: a practical guide to implementing statistical process control. New York: McGraw-Hill, 1988. 243 p.

LEHMANN, E. L. *Nonparametrics:* statistical methods based on ranks. San Francisco: Holden-Day, c1975. 457 p.

LITTLE, R. J. A.; RUBIN, D. B. *Statistical analysis with missing data*. New York: Wiley, 1987.

MANSOLDO, H. M.; SILVA, A. C. C. M. *Codificação automática/assistida*: uma proposta para o censo demográfico 2000. Rio de Janeiro: IBGE, Diretoria de Informática, 1997.

MANUAL X: indirect techniques for demographic estimation. New York: United Nations, Department of International Economic and Social Affairs, 1983. 304 p. (Population studies, n. 81).

MANUAL de crítica: sistema IMPS. Rio de Janeiro: IBGE, 2002.

MONTGOMERY, D. C. *Introduction to statistical quality control*. 3rd ed. New York: Wiley, c1996. Várias paginações.

NORMAS de apresentação tabular. 3. ed. Rio de Janeiro: IBGE, 1993. 62 p.

OLIVEIRA, J. de C. *Fecundidade e nupcialidade no Brasil e nos estados de São Paulo e Rio Grande do Norte:* tendências passadas e perspectivas. Rio de Janeiro: IBGE, 1991. 133 p.

PESSOA, D. G. C.; MOREIRA, G. G.; SANTOS, A. R. *Imputação de rendimentos no questionário da amostra do censo demográfico 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 17 p.

_____; SANTOS, A. R. *Imputação de rendimento dos responsáveis por domicílios – conjunto universo do censo demográfico 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 15 p.

_____; SILVA, P. L. N. *Análise de dados amostrais complexos*. São Paulo: Associação Brasileira de Estatística, 1998.

PLANO de análise da correção automática e elementos de apoio para a análise da formação do lote, sistema DIA, CD 1.02 - questionário da amostra. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2002. 37 p.

PROCEDIMENTOS a serem implementados para a execução da crítica (NIM e IMPS), questionário da amostra/censo 2000. Rio de Janeiro: IBGE, 2001.

RELATÓRIO sobre a experiência de tratamento automático de críticas entre registros, com vistas ao censo demográfico do ano 2000. [Rio de Janeiro]: IBGE, Diretoria de Pesquisas, 1999. 4 p.

RUBIO, E.; CRIADO, I. V. *Sistema DIA:* sistema de detección e imputación automática de errores para datos cualitativos. Madrid: Instituto Nacional de Estadística, 1988. v. 1: DIA : descripción del sistema.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. *Model assisted survey sampling.* New York: Springer-Verlag, c1992. 694 p.

SAS procedures guide: version 6. 3rd ed. Cary, NC: SAS Institute, c1990. 705 p.

SHRYOCK, H. S. et al. *The methods and materials of demography*. [Washington, D.C.]: Bureau of the Census, 1971. v.1.

SILVA, A. N. *Algumas considerações sobre o uso do NIM no censo demográfico 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 18 p.

_____; CORTEZ, B. F. *Censo demográfico 2000:* formação de lotes para a crítica de estrutura dos questionários da amostra – CD 1.02. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2000. 11 p.

_____; MATZENBACHER, L. A.; CORTEZ, B. F. *Processamento das áreas de expansão e disseminação da amostra no censo demográfico 2000*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2002. 28 p.

SILVA, L. F.; 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, n. 33).

SILVA, P. L. do N.; BIANCHINI, Z. M.; ALBIERI, S. *Uma proposta de metodologia para a expansão da amostra do censo demográfico de 1991*. Rio de Janeiro: IBGE, 1993. 106 p. (Textos para discussão, n. 62).

_____; PESSOA, D. G. C. *Estimando a precisão das estimativas das taxas de mortalidade obtidas a partir da PNAD*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2002. 24 p.

SINOPSE preliminar do censo demográfico 2000. Rio de Janeiro: IBGE, v. 7, 2000. Acompanha 1 CD-ROM.

TABULAÇÃO avançada do censo demográfico 2000: resultados preliminares da amostra. Rio de Janeiro: IBGE, 2002.

TRUSSELL, T. J. A re-estimation of the multiplying factors of the Brass technique for determining childhood survi-vors-hip rates. *Population Studies*, London, v. 19, n. 3, p. 97-107, 1975.

VENABLES, W. N.; RIPLEY, B. D. *Modern applied statistics with S-Plus.* New York: Springer, 1994.

WERKEMA, M. C. C. *Avaliação da qualidade de medidas*. Belo Horizonte: UFMG, Escola de Engenharia: Fundação Christiano Ottoni, 1996. 101 p. (Ferramentas da qualidade, 13).

_____. *Ferramentas estatísticas básicas para o gerenciamento de processos*. Belo Horizonte: UFMG, Escola de Engenharia: Fundação Christiano Ottoni, 1995. 384 p.