TEXTOS PARA DISCUSSÃO, ISSN 0103-6661

APURAÇÃO DOS DADOS INVESTIGADOS PELO QUESTIONÁRIO DA AMOSTRA - CD 1.02 DO CENSO DEMOGRÁFICO DE 1991

NÚMERO 86 **DEZEMBRO DE 1996**

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

APURAÇÃO DOS DADOS INVESTIGADOS PELO QUESTIONÁRIO DA AMOSTRA - CD 1.02 DO CENSO DEMOGRÁFICO DE 1991

Luís Carlos de Souza Oliveira
Bacharel em Estatística
Laura Baridó Indá
Bacharel em Estatística
Mauro dos Santos Mendonça
Bacharel em Ciências Administrativas
Rita Luzia Aguiar Lima
Bacharel em Estatística

Rio de Janeiro
1996

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Av. Franklin Roosevelt, 166 - Centro

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

DIRETORIA DE PESQUISAS

LENILDO FERNANDES SILVA

DIRETORA-ADJUNTA DE PESQUISAS

MARIA MARTHA MALARD MAYER

CHEFE DA COORDENAÇÃO TÉCNICA DO CENSO DEMOGRÁFICO

VALÉRIA DA MOTTA LEITE

TEXTOS PARA DISCUSSÃO

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

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

Apuração dos dados investigados pelo questionário da amostra - CD 1.02 do censo demográfico de 1991 / Luis Carlos de Souza Oliveira ... [et al.].- Rio de Janeiro : IBGE, Diretoria de Pesquisas, 1996.

79 p . - (Textos para discussão / Diretoria de Pesquisas, ISSN 0103-6661 ; n. 86)

ISBN 85-240-0627-7

1. Brasil - Censo demográfico, 1991 - Questionários. 2. Crítica de dados -Metodologia. I. Oliveira, Luis Carlos de Souza. II. IBGE. Diretoria de Pesquisas. III. Série.

Informações: Biblioteca Setorial da Diretoria de Pesquisas - Rua Visconde de Niterói, 1246, Bloco B, sala 1211-B, Mangueira
Telefone: (021) 284-8938 / 567-5322 - ramal 303

APRESENTAÇÃO

Este documento tem o objetivo de apresentar os procedimentos adotados na apuração do questionário da amostra - CD 1.02 do Censo Demográfico de 1991.

A execução dos procedimentos abrange etapas correspondentes a recepção e empastamento, transcrição de dados para o meio magnético, pré-crítica, codificação assistida, crítica, correção automática, análise da correção automática, expansão e armazenamento dos dados para tabulação e análise.

Desse modo, analogamente à apuração do questionário básico - CD 1.01, a intenção de produzir este documento baseou-se na necessidade de registrar, não só os fatos relevantes ocorridos durante a apuração do CD 1.02 mas, principalmente, as estratégias adotadas nas etapas do trabalho de apuração, com vistas à obtenção de um trabalho sobre a pesquisa em suas várias etapas, tornando-a o mais transparente possível, facilitando seu entendimento, além de subsidiar o planejamento do próximo Censo Demográfico.

AGRADECIMENTOS

Antes mesmo de iniciar qualquer dissertação sobre o trabalho de apuração dos dados do CD 1.02, convém destacar a importância da integração entre os técnicos das Unidades da Federação, da Coordenação Técnica do Censo Demográfico, do Departamento de Metodologia e da Coordenação de Informática.

Não houvesse o esforço de cooperação entre as partes envolvidas, desde a coleta dos dados, treinamento, implantação e execução dos sistemas de pré-crítica, codificação assistida e crítica, culminando com o suporte operacional na apuração centralizada e no armazenamento dos dados, no sentido de realizar um trabalho sério, e ao mesmo tempo prazeroso, todas as fases desse imenso trabalho, que é a apuração de um censo, tenderiam a ser desastrosas.

Em especial destacamos a participação dos analistas: Ataíde José de Oliveira Venâncio (DI - DIDEM) no que diz respeito à implementação do suporte computacional; Vandeli dos Santos Guerra (DPE - DEREN) pela colaboração na definição de críticas; Sonia Albieri e Zélia Bianchini (DPE - DEMET) pela atenciosa leitura e comentários acerca deste documento. Possíveis erros ou omissões são de nossa inteira responsabilidade.

Os autores

SUMÁRIO

1 - INTRODUÇÃO Erro! Indicador não definido.

Falar sobre a realização de uma pesquisa estatística, como o Censo Demográfico, é uma tarefa bastante complicada, dada sua importância como base para a realização de outras pesquisas mas, também, devido à complexidade na execução das suas várias etapas de trabalho relacionadas entre si, necessitando assim de um bom planejamento para que seja possível a obtenção de dados consistentes e fidedignos. Dessa forma, sua realização é considerada um grande projeto investigatório envolvendo etapas que devem ser cumpridas em conjunto pelo corpo técnico da Diretoria de Informática (DI) e da Diretoria de Pesquisas (DPE).

O questionário da amostra - CD 1.02 foi o responsável não só pela investigação das características básicas, aquelas investigadas para 100% das unidades domiciliares, como também pela investigação mais abrangente de características de domicílios e pessoas tais como: raça/cor, religião, nupcialidade, características de migração, instrução, fecundidade, mão-de-obra e rendimento.

A apuração dos dados do questionário da amostra que antecedeu a tabulação e divulgação dos resultados compreendeu duas etapas de trabalho:

- a primeira etapa subdividiu-se em 5 fases: recepção e empastamento, transcrição (digitação), pré-crítica, codificação assistida por computador[1] e crítica de incompatibilidades. Todas essas fases foram devidamente executadas, descentralizadamente, em 20 Unidades da Federação que, naquele momento, transformaram-se em pólos de apuração dos dados de algumas Unidades.

- a segunda etapa do trabalho foi realizada de modo centralizado, a saber: formação de lotes de apuração, execução do sistema de crítica e correção automática através do sistema DIA[2] - *Detección e Imputación Automática de errores para datos cualitativos*, análise da correção automática e expansão da amostra (determinação dos pesos a serem associados a cada unidade da amostra).

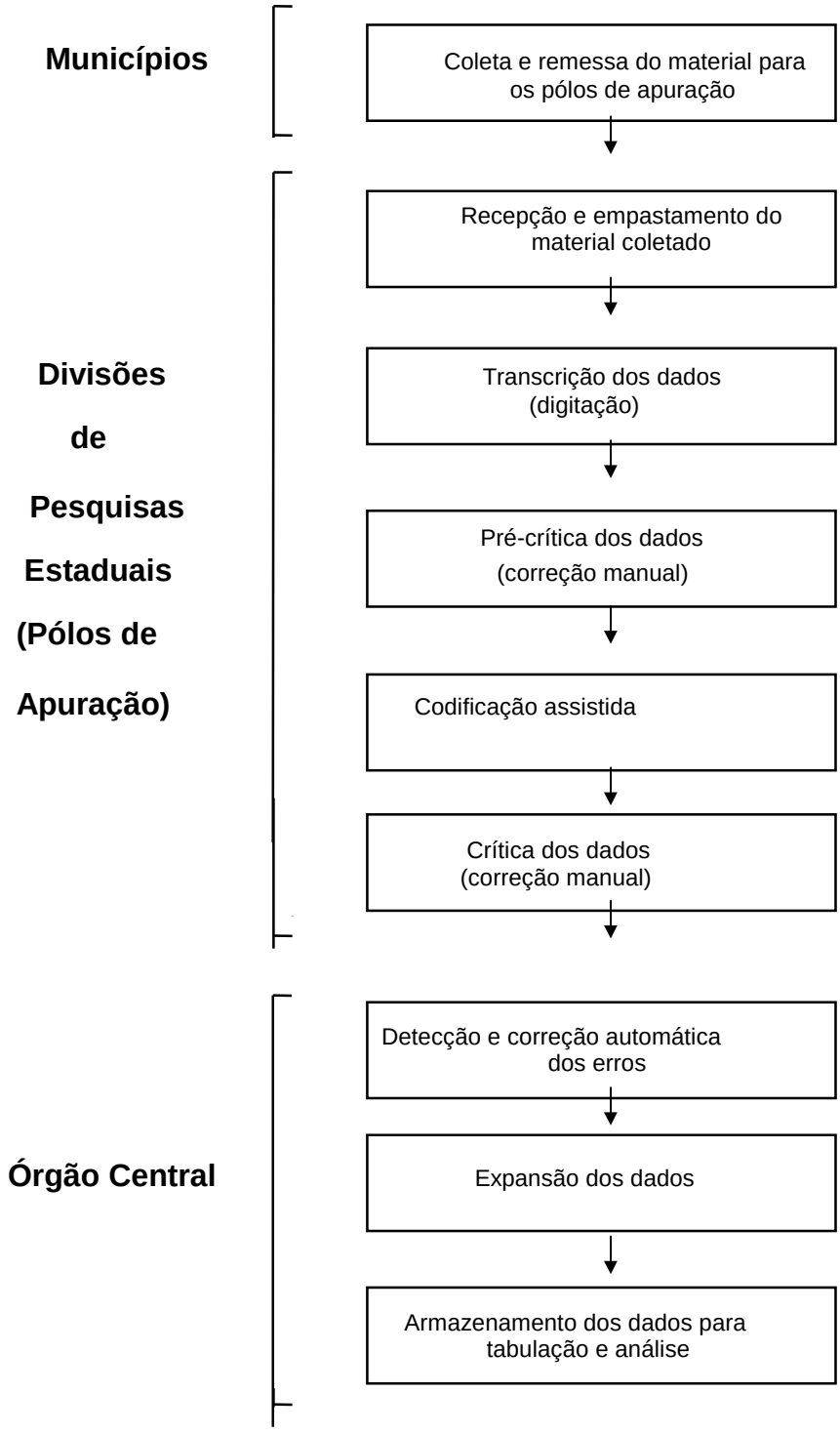
O Censo Demográfico de 1991 foi inovador em vários pontos importantes para a sua realização: descentralização de uma parte da apuração executada em 20 Unidades da Federação, utilização da codificação assistida, utilização do sistema Dia e a expansão dos dados da amostra do censo através de um método de estimação utilizado no Censo Canadense que baseia-se na metodologia GLSEP[3] - *Generalized Least Square*

[1] Ver Silva, Hanono e Barbosa (1993).
[2] Ver Rubio e Criado (1988).
[3] Ver Albieri e Dias (1994).

Estimation Procedure.

A seguir, é apresentada a figura 1 que procura enfatizar as grandes fases que envolveram a apuração do questionário da amostra CD 1.02.

Figura 1 - Visão Geral do processo de apuração do CD 1.02



2 - APURAÇÃO DESCENTRALIZADAErro! Indicador não definido.

2.1 - A Recepção e o Empastamento do Material ColetadoErro! Indicador não definido.

Esta fase da apuração compreendeu os seguintes procedimentos:

- Recepção e separação dos vários tipos de formulários: CD 1.01 (questionário básico), CD 1.02 (questionário da amostra)[4], CD 1.03 (lista de domicílio coletivo), CD 1.06 (caderneta do setor) e CD 1.07 (folha de coleta).
- Conferência de todos os setores relacionados no CD 1.10 (boletim-resumo) e da quantidade dos CD 1.01 e CD 1.02, com o correspondente registro no resumo do CD 1.06. Durante a conferência dos questionários poderiam ocorrer alguns problemas como a falta de questionário e a existência de questionário não relacionado na folha de coleta. Nesse momento eram tomadas providências necessárias de acordo com os procedimentos definidos pela Coordenação Técnica do Censo Demográfico (CTD), com a finalidade de sanar os mesmos.
- Empastamento do material que baseou-se nos arquivos gerados pela digitação do CD 1.10, cujas informações serviram de base para a emissão de uma listagem, "relação de empastamento", dos SLIPs[5] e da folha de acompanhamento de serviço. A esse tipo de empastamento é que se denominou "empastamento lógico", funcionando como uma fase preparatória do empastamento físico, propriamente dito.

Na verdade o empastamento físico foi responsável pela colocação dos questionários nas pastas, bem como a colocação das etiquetas nas mesmas, com indicação do material nelas contido. Somente após todos os questionários de uma mesma pasta terem sido devidamente empastados, eram preparadas as unidades de trabalho da digitação, denominadas "tarefas de digitação", contendo cerca de 10 questionários no caso do CD 1.02.

Todo o trabalho envolvendo a recepção dos formulários até o respectivo acondicionamento nas próprias pastas foi desenvolvido segundo alguns critérios explicativos contidos em "X Recenseamento Geral do Brasil - 1991 - Manuais das Atividades da Apuração Descentralizada do Censo Demográfico de 1991 - Instrução de Empastamento".

A tabela1 apresenta os quantitativos de pastas e questionários da amostra para o

[4] Ver cópia do questionário no anexo.
[5] SLIP é o nome dado ao documento utilizado para definir, separar e controlar os questionários de cada tarefa de digitação. O SLIP era impresso com vários campos de controle da digitação: modelo de questionário, município, nº da pasta, distrito, subdistrito, setor e números dos questionários inicial e final que compunham a tarefa.

Brasil e as Unidades da Federação.

Tabela 1 - Número de pastas e questionários segundo as Unidades da Federação

Unidades da Federação	Pastas	Questionários (CD 1.02)
BRASIL	27.819	4.032.256
Rondônia	188	26.872
Acre	73	9.834
Amazonas	373	45.768
Roraima	39	5.514
Pará	825	107.073
Amapá	52	6.079
Tocantins	244	29.808
Maranhão	831	106.517
Piauí	541	66.490
Ceará	1.122	151.238
Rio Grande do Norte	569	72.066
Paraíba	714	89.699
Pernambuco	1.249	172.833
Alagoas	496	61.572
Sergipe	328	42.139
Bahia	2.405	306.892
Minas Gerais	3.319	462.429
Espírito Santo	478	70.539
Rio de Janeiro	2.093	357.326
São Paulo	5.528	881.517
Paraná	1.673	249.356
Santa Catarina	966	141.046
Rio Grande do Sul	1.832	292.787
Mato Grosso do Sul	358	52.972
Mato Grosso	431	60.908
Goiás	855	124.532
Distrito Federal	237	38.450

Fonte: Listagem de março de 1996 contendo informações sobre pastas, nº de setores, etc.

2.2 - Transcrição dos Dados Erro! Indicador não definido.

De posse do material coletado, adequadamente empastado, foi iniciado o processo de digitação dos dados para o meio magnético.

2.2.1 - Execução da transcrição dos dados Erro! Indicador não definido.

O processo de transcrição dos dados do CD 1.02 foi executado de tal modo que, cada digitador ficou responsável por uma tarefa enquanto que a verificação da mesma foi feita por um outro digitador, distinto do primeiro.

Na primeira digitação foram digitadas as variáveis do quadro 1, V101 (município), V102 (pasta), V103 (nº na pasta), V104 (distrito), V105 (subdistrito) e V106 (setor). Os dados destas variáveis foram confrontados com os dados constantes do SLIP, sendo que na hipótese de ocorrer alguma divergência entre os dados dos questionários e os do SLIP, a tarefa era suspensa até o supervisor responsável solucionar o problema.

A verificação dos dados digitados foi feita para 100% dos questionários através das tarefas, do mesmo modo que o executado na apuração do CD 1.01[6], acarretando a dupla digitação, enquanto que os SLIPs foram digitados apenas uma vez, no primeiro ciclo.

Nesse sentido, quando o dado digitado pela segunda vez coincidia com o primeiro, o mesmo era armazenado no arquivo; em caso contrário, havia a necessidade de uma digitação que confirmasse qualquer dos dados já digitados.

2.3 - Pré-Crítica dos Dados Erro! Indicador não definido.

A operação de crítica e correção dos erros detectados nos dados investigados no CD 1.02 envolveu duas fases de trabalho, uma manual e outra automática.

O processo de crítica manual foi desenvolvido em 2 fases, a saber: a pré-crítica dos dados e a crítica de incompatibilidades. Anteriormente à execução, propriamente dita, das fases de pré-crítica e codificação assistida, foi realizado um treinamento centralizado dividido em duas turmas abrangendo os períodos de 08 a 12/02/93 e 15 a 19/02/93. Nessa oportunidade foram apresentadas as partes teórica e prática do funcionamento dos sistemas dessas duas fases, bem como a análise dos relatórios e acertos *on line.*

As críticas definidas na fase de pré-crítica compreenderam as de possibilidades das variáveis criticadas naquele momento, onde eram verificados se os valores digitados

[6] Ver Oliveira, Indá, Lima e Bianchini(1994).

correspondiam a códigos válidos para cada um desses quesitos investigados, incluindo o preenchimento de quesitos obrigatórios; conferência da seqüência da identificação, do número de questionários dentro de cada pasta, do número de pessoas dentro dos questionários e dos totais de pessoas; e algumas críticas de incompatibilidades cuja correção dependia do acesso aos instrumentos de coleta.

2.3.1 - Variáveis tratadas na pré-crítica Erro! Indicador não definido.

V111 (total de homens), V112 (total de mulheres) e V301 (sexo)

Os dados destas variáveis foram confrontados e corrigidos manualmente de forma a ficarem consistentes entre si, tendo em vista a possibilidade de que algum registro de pessoa não tivesse sido digitado. Assim sendo, existia a chance de incluir os dados não digitados antes mesmo da execução da fase de codificação.

V201 (espécie do domicílio)

A correção desta variável foi feita através das críticas com a V110 (nº no CD 1.03) que, dependendo do preenchimento desta última, o domicílio era considerado particular ou coletivo.

V209 (aluguel) e V356, V357, V360, V361 (rendimentos)

Nessa fase estas variáveis foram criticadas através da detecção de possíveis inconsistências entre o valor do aluguel/rendimento e o número de dígitos.

As instruções relativas ao acerto destas variáveis encontram-se, detalhadamente, em “X Recenseamento Geral do Brasil de 1991 - Manuais das Atividades da Apuração Descentralizada do Censo Demográfico de 1991 - Plano de Pré-Crítica - Questionário CD 1.02”.

2.3.2 - Criação de questionários faltosos Erro! Indicador não definido.

De maneira análoga ao questionário básico - CD 1.01, existiram também os questionários faltosos do CD 1.02, isto é, aqueles desaparecidos durante a fase do empastamento.

O processo de criação do questionário faltoso da amostra foi semelhante ao do questionário básico (CD 1.01), colocando-se código 9999 na V109 (n° no CD 1.07) durante o empastamento, quando existia registro de domicílio ocupado na folha de coleta e não existia nenhum questionário, pois se existisse o questionário básico correspondente ao registro, todas as informações disponíveis seriam transmitidas para o questionário da amostra e, desse modo, este questionário não seria mais tratado como faltoso.

Para os questionários faltosos (V109 = 9999) o processo era retirar, também, das folhas de coleta as informações disponíveis como: o n° no CD 1.03, espécie do domicílio e o total de pessoas por sexo, preenchendo a V301 (sexo) de todas as pessoas do questionário.

No entanto, é importante ressaltar a diferença existente entre o tratamento dado ao questionário faltoso referente ao questionário básico e o questionário faltoso da amostra. De acordo com Oliveira e outros (1994), o tratamento dado ao questionário faltoso (CD 1.01) foi recuperar as informações disponíveis nas folhas de coleta como: o nº no CD 1.03, total de pessoas por sexo e espécie de domicílio, enquanto as demais variáveis foram imputadas após a execução da correção automática pelo sistema DIA, segundo um processo seqüencial de procura de questionários corretos que exerceram a função de doadores dos dados de domicílios e pessoas.

O questionário faltoso da amostra não seguiu os mesmos passos dos demais, isto é, não passou pela fase de codificação nem da crítica de incompatibilidades, e a ausência desses dados foi contornada pelo processo de expansão da amostra.

Isso é razoável visto que diferentemente do questionário básico, o questionário da amostra é um questionário com um maior nível de detalhamento e não seria correto adotar procedimento idêntico ao anterior, de procurar um doador que tivesse os mesmos totais de pessoas por sexo e mesmo valor na V201 e copiar as informações referentes às demais variáveis.

2.4 - Codificação Assistida por Computador Erro! Indicador não definido.

Posteriormente à execução da pré-crítica, os questionários foram submetidos a uma outra fase de trabalho que, em relação aos censos anteriores, foi considerada por

todos um grande avanço no processo de codificação dos quesitos em aberto: religião, migração, curso concluído, ocupação e atividade.

A fase de codificação do Censo de 1980 tinha como objetivo a aplicação de códigos aos quesitos em aberto, de forma manual, utilizando alguns manuais de codificação. Todos os quesitos de um mesmo questionário eram examinados e codificados por um único operador. Do modo como era realizado esse tipo de trabalho, fazia-se necessário lançar mão de grandes controles operacionais e, por mais que houvessem técnicos atuando como orientadores desses codificadores, era mais difícil manter uma uniformidade nos procedimentos[7].

Com base no sucesso obtido em outras pesquisas que já utilizavam a codificação assistida por computador e também no teste realizado no Censo Experimental de Limeira (1988), o IBGE optou por introduzir no Censo Demográfico de 1991, a codificação assistida por computador.

O funcionamento do processo de codificação consiste de duas aplicações: codificação automática e codificação assistida[8].

Em todo o processo de codificação é utilizado um procedimento de fonetização que partindo de uma palavra com mais de duas letras, realiza eliminação de plural, de masculino e feminino assim como tratamento de fonemas de maneira padronizada (por exemplo: substituição de C K Q S X Z por C, retirada do H, etc.).

Como a codificação automática é uma transformação de descrições (texto) em códigos, pode ocorrer que a uma descrição corresponda um único código, mais de um código ou nenhum. Na realidade a codificação automática é acionada quando a descrição ou texto de um determinado quesito é encontrado no banco de códigos, previamente construído. Nesse caso, o sistema de codificação aplica, automaticamente, o código correspondente ao quesito sem a interferência do codificador. Há de se convir que qualquer que seja o número de casos passíveis de serem codificados automaticamente, já representa um ganho considerável quando comparado ao processo de codificação manual utilizado em censos anteriores. Quando não foi possível a codificação automática, isto é, no banco de códigos existia mais de um código para uma determinada descrição ou não era encontrado nenhum código referente àquela descrição, o sistema emitia listagens com informações sobre o número da pasta, questionário, pessoa, quesito e a situação de excesso ou de falta de código para cada descrição.

[7] Ver Metodologia do Censo Demográfico de 1980.
[8] Ver Silva e outros (1993).

Com base na listagem e em outros instrumentos como o próprio questionário e a relação de códigos, o codificador entrava em ação a fim de sanar o problema surgido na codificação, escrevendo na própria listagem a descrição mais acertada para o quesito em questão. Feito isso o codificador registrava, *on line*, a nova descrição para que o sistema pudesse codificar o novo texto através de um novo ciclo de trabalho. A esse procedimento dá-se o nome de "codificação assistida".

É bom que se diga que o sucesso dessa fase do trabalho está intimamente relacionado à definição do banco de códigos e descritores, ou seja, quanto mais completo for esse banco, menores serão as chances dos códigos não serem encontrados, não precisando então, passar para a fase de codificação assistida.

2.5 - Crítica de Incompatibilidades Erro! Indicador não definido.

Para a execução da crítica, semelhantemente à realização das fases de pré-crítica e codificação assistida por computador, foi também realizado um treinamento centralizado em duas turmas no período de novembro a dezembro de 1993. A apuração dos dados foi preparada de modo que cada conjunto de questionários, acondicionados em pastas na fase da recepção e empastamento do material coletado, representasse um lote de trabalho para efeito de organização e controle do material coletado e criticado.

Essa fase do trabalho de apuração consistiu em corrigir as inconsistências que, de acordo com a arquitetura montada para a utilização do sistema DIA na correção automática dos dados, envolviam variáveis que deveriam estar "consistentes", no momento da correção automática, incluindo as variáveis "chave" para o salto de bloco, as críticas de possibilidades onde eram verificados se os códigos digitados eram válidos, além de críticas entre-registros[9], que não são tratadas pelo sistema DIA, visto que esse sistema só executa críticas que envolvem variáveis de um mesmo registro.

Vale ressaltar que, apesar de nem todas as críticas previstas para a apuração do questionário da amostra poderem ser trabalhadas em meio automático, qualquer que seja o método de detecção e correção, os analistas responsáveis pela elaboração e arquitetura do sistema de críticas tiveram sempre em mente a idéia de corrigir visualmente o menor número de críticas possível, procurando transferir o máximo de críticas para o sistema automático de modo a homogeneizar as correções e eliminar a possibilidade de existirem tendenciosidades.

Com o intuito de evitar todo um trabalho de renumeração das pessoas durante a

[9] Críticas entre-registros são aquelas que apresentam uma ligação existente entre dados referentes a dois ou mais registros.

crítica, foram introduzidos alguns critérios pré-estabelecidos com vistas à ordenação lógica das pessoas em contraposição à ordem física no questionário. Desse modo, as críticas envolvidas foram trabalhadas, não pela ordem em que as pessoas foram recenseadas no questionário, mas pela ordem em que deveriam estar registradas, isto é: em primeiro o chefe do domicílio e, em seguida, o cônjuge, os filhos (em ordem decrescente de idade), etc., no caso de família única. Essa ordem deveria ser respeitada dentro de cada família, nos domicílios particulares habitados por famílias conviventes, além de ordená-las a partir da primeira convivente, seguindo-se a segunda convivente até a quinta, se houvesse.

O trabalho de execução da crítica[10] foi realizado por críticos que tinham acesso a uma listagem emitida pelo sistema com o número do questionário, número da pessoa com erro, o número do erro e as variáveis envolvidas. De posse dessa listagem, dos instrumentos de coleta (folhas de coleta e questionários) e das instruções de crítica, esses críticos tinham condições de corrigir as falhas existentes.

É importante salientar que a partir da fixação de determinadas variáveis consideradas "consistentes" nessa fase, é que se desenvolveu toda a fase posterior de correção. Desse modo, são apresentadas a seguir as variáveis que foram realmente tratadas na etapa descentralizada, além de outras cujo tratamento ocorreu na fase da correção automática, mas que foram incluídas como forma de subsidiar a correção daquelas variáveis e de preparar os blocos do questionário para a fase de imputação.

- **Quadro 1 :**

V111 - total de homens

V112 - total de mulheres

Apareciam também nessa fase devido à necessidade de excluir pessoas nascidas após a data do censo (recenseadas indevidamente). Nesse caso, as variáveis contadoras também deveriam ser alteradas.

- **Quadro 3 :**

V301 - sexo

V302 - parentesco com o chefe do domicílio

V303 - parentesco com o chefe da família

V304 - família a que pertence

[10] Ver Manuais das atividades da apuração descentralizada do Censo Demográfico de 1991.

Estas variáveis foram criticadas com a V332 (se vive em companhia do cônjuge) para a organização da estrutura familiar, como já foi dito anteriormente, cujas correções só poderiam ser feitas analisando as informações do questionário.

V305 - número de ordem da mãe

Seu preenchimento dependia tão somente do registro das pessoas no questionário e encontrava-se envolvida em críticas entre-registros.

V306/V307 - mês/ano de nascimento ou idade presumida

Através delas foram obtidas as idades calculadas a fim de comporem as estruturas etárias a nível de Unidade da Federação, Município, etc., que juntamente com outras variáveis fornecem tabulações que permitem, dentre outras coisas, realizar estudos ao longo dos anos. Levando-se em conta a relevância destas variáveis para a obtenção das idades calculadas, optou-se pelo tratamento nessa fase pelos seguintes motivos:

- caso houvesse erro de digitação ou de preenchimento, a correção seria imediata, já que naquele momento o crítico tinha acesso aos questionários;

- caso fosse necessário excluir registros de pessoas recenseadas indevidamente (nascidas após a data de referência).

Desse modo, depreende-se que essa fase foi a mais adequada para a correção dessas variáveis, a fim de que o seu produto, que é a idade calculada, pudesse ser criticado na outra fase de correção, garantindo o mínimo de consistência entre as fases de crítica e a correção automática.

V308 - faixa de idade

Sua correção dependia das declarações da variável mês/ano e do preenchimento dos blocos do questionário e, posteriormente, foi utilizada na correção automática como apoio para a imputação de outras variáveis.

V309 - raça ou cor

V311 **-** deficiência física ou mental

Não existia motivo algum para não tratá-las nessa fase, uma vez que a imputação dessas variáveis seria realizada de forma desconexa das outras variáveis não oferecendo, então, ganho algum.

V314 - nasceu neste município
V315 - ano em que fixou residência no País
V3016 - variável indicadora da UF/País de nascimento
V317 - há quantos anos mora nesta Unidade da Federação
V318 - há quantos anos mora neste município
V3193 - variável indicadora do Município/País que residia antes de mudar para este município
V3214 - variável indicadora do Município/País que residia em 01/09/86

Nessa fase, as variáveis do bloco de migração foram incluídas a fim de subsidiar a correção da V314, indicadora de salto, de modo que na fase de correção automática esta permanecesse fixa. Por outro lado, o tratamento das variáveis indicadoras de Município/País, nessa fase, deveu-se ao simples fato de que somente a visualização do questionário daria respaldo a uma correção adequada. Associado a isso, existia a impossibilidade de se trabalhar, no sistema DIA, com variáveis cujo número de dígitos fosse maior que quatro.

V323 - sabe ler e escrever
V324 **-** série que freqüenta
V325 **-** grau da série que freqüenta
V326 **-** freqüência a curso não seriado
V327 **-** última série que concluiu com aprovação
V328 **-** grau da última série que concluiu com aprovação
V3291 **-** variável indicadora da espécie do curso mais elevado concluído com aprovação

A inclusão das variáveis do bloco de instrução nessa fase do trabalho se fez necessária para subsidiar o preenchimento adequado da V308 (faixa de idade) e, conseqüentemente, preparar o bloco para a fase de imputação.

V330 - vive ou viveu em companhia de cônjuge
V331 - mês/ano em que contraiu a 1ª união
V332 - se vive em companhia do cônjuge, indique a união
V333 - se não vive em companhia do cônjuge, indique a condição
V334 - mês/ano em que passou a viver com o cônjuge atual

O bloco de nupcialidade foi tratado nessa fase, visto que além das críticas intra-

registros[11] existiam também as críticas entre-registros que não são tratadas pelo sistema DIA. Além disso, o possível preenchimento da V330 também auxiliou na definição da V308.

V335/V344 **-** As variáveis de fecundidade também fizeram parte dessa fase com vistas à preparação do bloco para as críticas da fase automática, corrigindo as V335/V336(filhos(as) tidos(as) que moram no domicílio) e V344 (mês/ano de nascimento do último(a) filho(a) nascido(a) vivo(a)) cuja correção dependia da visualização do questionário, além do fato desta última estar relacionada com a V305, cujo tratamento foi realizado somente nessa fase.

V3443 - idade do último(a) filho(a) nascido(a) vivo(a)

Esta variável participou de críticas entre-registros envolvendo as V305 e V344 que foram tratadas na etapa descentralizada.

V345 - trabalhou em todos ou em parte dos últimos 12 meses

V3461 - variável indicadora da ocupação principal

V3471 - variável indicadora do setor ou ramo de atividade

V349 **-** relação de trabalho ou a posição da ocupação principal

V350 - tem carteira de trabalho assinada

V351 - quantas pessoas trabalham no negócio, instituição, etc.

V352 **-** local de trabalho onde exerceu a ocupação principal

V353 **-** é contribuinte de Instituto de Previdência Pública

V354/V355 **-** horas habitualmente trabalhadas por semana

V356 , V357, V360 e V361 **-** tipos de rendimento

V358 **-** se não trabalhou indique a situação ou ocupação que tem

V359 **-** indique se é aposentado e/ou pensionista

Estas variáveis foram incluídas nesse momento com o intuito de auxiliar a correção da V345, que é a variável indicadora de salto do bloco de mão-de-obra, de tal forma que na fase subseqüente pudesse ser fixada e, as inconsistências dentro do bloco, corrigidas.

[11] São aquelas que apresentam uma ligação existente entre dados referentes a duas ou mais variáveis de um mesmo registro.

Nesse momento, são apresentadas as variáveis já tratadas e que permaneceram fixas durante a execução da fase de correção automática:

- **Quadro 1 -** V111 e V112
- **Quadro 2 -** V201
- **Quadro 3 -** V301, V302, V303, V304, V305, V306/V307, V308, V309, V311, V314, V3016, V3193, V3214, V330, V331, V332, V333, V334, V335, V336, V344 e V345.

3 - APURAÇÃO CENTRALIZADAErro! Indicador não definido.

3.1 - Formação e Seleção de LotesErro! Indicador não definido.

O início da apuração centralizada foi condicionado a um processo de validação exercido sobre os conjuntos de dados transmitidos pelos pólos de apuração, através da revisão da quantidade de questionários por setor e da reexecução centralizada do plano de críticas utilizado nos ambientes descentralizados.

Ocorre que para se levar a cabo as tarefas referentes à detecção e correção dos erros encontrados nos questionários, seja manual ou automática, está intrínseca a necessidade de se organizar, de alguma forma, em meio magnético, os questionários. Essa fase foi executada após a transmissão dos dados pelas Unidades responsáveis pela apuração da pré-crítica, codificação e crítica de incompatibilidades.

Diferentemente das fases anteriores de crítica que consideraram a pasta como lote de trabalho, a execução da correção automática foi realizada através dos chamados "lotes de apuração", os quais foram submetidos ao programa de detecção e correção dos erros.

Objetivando um mínimo de coerência durante a correção automática, fez-se necessário formar os lotes de apuração mediante alguns critérios considerados importantes na obtenção dos resultados finais.

Os critérios foram os seguintes[12]:

- considerar o tamanho máximo de um lote em cerca de 70000 questionários;
- formar os lotes levando em conta não só a situação do domicílio, mas também a divisão geográfica em cada Unidade da Federação, ordenando os setores segundo a mesorregião, microrregião, município, distrito e subdistrito, com o intuito de preservar, na medida do possível, as características regionais. Essa preocupação em preservar a homogeneidade em cada região tem a ver também com a definição das áreas de ponderação[13], em função do método de expansão. Essa homogeneidade é importante em função da metodologia de imputação, que se baseia na distribuição dos dados dos questionários "bons" observados no lote, entendendo-se por questionários bons todos aqueles que não apresentaram nenhum erro segundo o conjunto de regras definido pelos analistas; e
- obtenção do menor número possível de lotes de modo a agilizar a análise da

[12] Ver Albieri e Oliveira (1994).
[13] Por área de ponderação entende-se a área geográfica a ser utilizada para o cálculo dos pesos associados às unidades da amostra.

imputação, não definindo tamanho mínimo, enquanto não fossem realizados testes com os planos de crítica definitivos. A idéia que está por trás de não estipular um tamanho mínimo está relacionada com a possibilidade de ocorrerem problemas na imputação devido ao tamanho do lote, e nesse caso seria necessário proceder a agregação de alguns lotes.

Dentre outras variáveis auxiliares incluídas nos arquivos de dados, durante a preparação dos lotes de apuração, encontravam-se a V100 (número do lote; códigos possíveis: 1 a 216) e a V1061 (situação do setor; códigos possíveis: 1 (área urbanizada), 2 (área não urbanizada), 3 (área urbanizada isolada), 4 (aglomerado rural com extensão urbana), 5 (aglomerado rural povoado), 6 (rural núcleo), 7 (rural outros) e 8 (rural exclusive aglomerados)).

Durante o processo de formação de lotes, os dados dos questionários da amostra, para cada Unidade da Federação, foram armazenados em arquivos contendo os controles específicos dos lotes de apuração.

Antes que cada lote de apuração fosse submetido à detecção e correção automática dos erros, já em ritmo de produção, os questionários referentes a um mesmo lote eram transferidos para um único arquivo, processo este denominado “seleção de lotes”, cujo objetivo era a obtenção de uma nova unidade de trabalho (conjunto de dados) segundo os critérios definidos anteriormente. Essa unidade de trabalho era composta dos dados transmitidos e validados. Desse modo, cada lote de apuração correspondeu a um arquivo de trabalho.

A tabela 2 apresenta a distribuição do número de lotes de trabalho e de setores censitários para as Unidades da Federação.

Tabela 2 - Número de lotes e setores segundo as Unidades da Federação

Unidades da Federação	Lotes	Setores
BRASIL	216	160.770
Rondônia	2	1.382
Acre	2	389
Amazonas	2	2.046
Roraima	2	254
Pará	2	4.338
Amapá	2	235
Tocantins	2	1.083
Maranhão	2	5.396
Piauí	2	3.169
Ceará	14	6.681
Rio Grande do Norte	2	2.702
Paraíba	2	3.582
Pernambuco	10	7.353
Alagoas	2	2.566
Sergipe	2	1.744
Bahia	14	12.410
Minas Gerais	26	17.340
Espírito Santo	2	2.671
Rio de Janeiro	20	14.267
São Paulo	40	32.725
Paraná	20	10.248
Santa Catarina	12	5.165
Rio Grande do Sul	16	12.878
Mato Grosso do Sul	2	2.040
Mato Grosso	2	2.145
Goiás	10	4.305
Distrito Federal	2	1.656

Fonte: Listagem de março de 1996 contendo informações sobre pastas, nº de setores, etc.

3.2 - Detecção e Correção Automática dos ErrosErro! Indicador não definido.

3.2.1 - Metodologia de Fellegi & HoltErro! Indicador não definido.

A apuração da fase centralizada do Censo Demográfico de 1991 foi realizada a partir da metodologia desenvolvida por Fellegi & Holt (1976). Esta metodologia[14] permite uma integração das etapas de detecção e correção de inconsistências em apenas um ciclo de processamento.

Isso significa dizer que quando um erro é detectado num registro, alguma correção ou alteração é introduzida sobre os dados desse registro, de modo a torná-los consistentes.

Os princípios básicos, que norteiam a metodologia, são os seguintes:

1) manter a maior quantidade possível de informação original, alterando assim o mínimo de variáveis;

2) não existe a necessidade de escrever os critérios de imputação, como ocorre no método que utiliza a matriz *Hot-Deck*[15] pois esses critérios são deduzidos automaticamente das regras de incompatibilidades. Isto significa que os dados imputados estarão de acordo com todas as regras, simplificando o trabalho de programação e, obviamente, eliminando a possibilidade de introdução de erros por ocasião da definição de regras de imputação; e

3) a imputação dos dados procura manter, na medida do possível, as distribuições marginais ou conjuntas de freqüências das variáveis referentes aos registros corretos.

A operacionalização dessa metodologia baseia-se na necessidade de especificação, por parte dos analistas, de um dicionário formado pelas variáveis a depurar e seus respectivos valores válidos e de um conjunto de regras de incompatibilidades, escritas em forma de erro (denominada forma normal), definindo as combinações inconsistentes dos subconjuntos de códigos das variáveis.

A partir dessas definições é então criado um conjunto de regras, formado por regras escritas pelos analistas (regras explícitas) e por outras deduzidas logicamente (regras implícitas), denominado conjunto completo de regras.

Exemplo de regra explícita:

V308(faixa de idade) = 1(0 a 4 anos) $\cap$ V3076(idade) = 5 a 9 $\Rightarrow$ erro

V308(faixa de idade) = 2(5 a 9 anos) $\cap$ V327(última série concluída) = br $\Rightarrow$ erro

[14] Uma descrição mais detalhada pode ser encontrada em Silva e outros (1990) e Rubio e Criado (1988).

[15] Procedimento que tinha como base a construção de matrizes formadas pelo cruzamento de variáveis pertinentes à variável a ser corrigida, e que eram continuamente atualizadas com as informações anteriormente consideradas corretas.

Exemplo de regra implícita**:**

V327(última série concluída) = br $\cap$ V3076(idade) = 5 a 9 $\Rightarrow$ erro

Segundo Fellegi & Holt, utilizando-se o conjunto completo, existirá sempre a possibilidade de se encontrar um conjunto mínimo de campos a imputar e algum código destes campos de tal forma que o registro imputado não falhe em nenhuma regra.

O processo de imputação é realizado com base no conjunto completo e no dicionário de valores válidos. Desse modo, quando um registro é detectado como errôneo, é então obtido um conjunto de campos a imputar que corrige o registro.

Uma vez selecionada(s) a(s) variável(is) que deve(m) ser imputada(s), é então eleito o código a imputar em cada variável.

Para isso, a metodologia oferece duas saídas para a imputação:

- imputação seqüencial (variável a variável) que procura manter a distribuição de freqüências marginal dos registros bons;

- imputação conjunta que ocorre quando uma variável é imputada em função do(s) valor(es) de outra(s) variável(is). Com isso, procura-se manter a distribuição de freqüências conjunta dos registros bons.

3.2.2 - O sistema DIA Erro! Indicador não definido.

A sigla DIA é uma abreviação de *Detección e Imputación Automática de errores para datos cualitativos*, nome dado ao pacote computacional desenvolvido pelo Instituto Nacional de Estadística - INE (Espanha) com base na metodologia de Fellegi & Holt.

O objetivo do desenvolvimento do sistema DIA era dispor de um pacote geral que facilitasse a apuração de censos e grandes pesquisas estatísticas.

Este sistema permite não só a utilização de correção probabilística sugerida na metodologia de Fellegi & Holt, como também a de correção determinística, usada no tratamento de erros sistemáticos, e ainda garante uma consistência entre as regras de incompatibilidades e as de imputação determinística.

A operacionalização do sistema é feita em ambiente centralizado, tipo mainframe IBM e trata somente de arquivos seqüenciais com um único tipo de registro de tamanho fixo com variáveis categóricas ou qualitativas[16].

O sistema DIA é constituído de dois subsistemas: de especificações e de tratamento dos dados.

As funções desenvolvidas pelos módulos que compõem o subsistema de

[16] Ver Hanono(1993).

especificações podem ser resumidas em:

- análise sintática das especificações;

- tradução das especificações para o formato interno do sistema;

- geração do conjunto completo de regras;

- geração e compilação dos programas a serem usados no subsistema de tratamento dos dados;

- criação dos arquivos de regras e auxiliares, no formato a ser usado no subsistema de tratamento dos dados; e

- criação da estratégia de depuração.

Tendo como entrada o plano de depuração fornecido pelos analistas, o sistema DIA procede a uma análise e passa então a gerar os programas e as estruturas de dados, denominado “Aplicação - DIA”, necessária para a depuração dos dados.

O subsistema de especificações é formado basicamente pelo analisador de regras, cuja função consiste em comprovar a consistência interna dos conjuntos de regras de incompatibilidades e de regras de imputação determinística, eliminando redundâncias existentes, como também verificar a consistência do conjunto de regras determinísticas em relação ao conjunto de regras explícitas.

O plano de depuração divide-se em dois tipos:

- especificação do usuário.
  - definição das variáveis a depurar e seus respectivos códigos possíveis;
  - definição das regras escritas em forma de erro (forma normal);
  - definição das regras de imputação determinística.
- a estratégia de depuração.

Algumas dessas estratégias são definidas a seguir:

- fixação de variáveis - trata-se de uma estratégia que foi bastante utilizada durante a apuração do Censo 1991, quando algumas variáveis já depuradas em fases anteriores não deveriam sofrer alterações durante o processo automático ou ainda, para as variáveis auxiliares que participavam do conjunto de regras com o intuito de auxiliar a depuração de outras variáveis;

- ponderação de variáveis - esta estratégia representa o grau de desconfiança da variável a imputar, cujos valores variam de 1 a 10. Desse modo, quanto menor esse valor maior o grau de confiabilidade da variável, permitindo aos analistas definirem que determinadas variáveis não fossem muito imputadas, em detrimento de outras; e

- critérios de imputação de códigos :

1) tipos de distribuição - nesse caso existe a possibilidade de utilizar distribuições de freqüências marginais ou conjuntas para a imputação de determinada variável;
2) composição das distribuições de freqüências (FNS) ou (FNS*) - tratam-se de alternativas para a obtenção das distribuições de freqüências a serem usadas durante a imputação das variáveis. Significa dizer que a imputação probabilística de códigos, seja através de uma distribuição de freqüências marginal ou de uma distribuição conjunta, é realizada utilizando-se a freqüência de registros não suspeitos (FNS) ou (FNS*), conforme o caso.

O uso da freqüência de registros não suspeitos (FNS) para a imputação de uma determinada variável ocorre quando, para o cálculo da distribuição de freqüências do sistema DIA, são contabilizados todos os registros bons (sem erros) e aqueles registros errôneos em que esta variável, se a distribuição é marginal, ou as variáveis, se a distribuição é conjunta, são não suspeitas.

Para a imputação segundo a FNS* valem os mesmos critérios adotados para a FNS, após a possível atuação das RIDs (regras de imputação determinística).

O subsistema de tratamento dos dados consiste em executar, verdadeiramente, a tarefa de depuração dos dados utilizando os programas e arquivos auxiliares gerados no subsistema de especificação.

Esse subsistema tem as seguintes funções:

- detecção de registros errados;
- estatísticas de erros;
- imputação determinística;
- imputação probabilística; e
- gravação de arquivos com os resultados da imputação.

3.2.3 - A execução da correção automática Erro! Indicador não definido.

O processo de detecção e correção automática do CD 1.02 foi planejado para ser executado sobre os dados armazenados nos lotes de apuração, os quais após serem selecionados foram submetidos às regras definidas pelos analistas, através de oito aplicações, sendo uma referente à imputação das características de domicílios e as outras sete referentes às características de pessoas.

Cabe registrar que o número de aplicações utilizado na depuração dos dados do CD 1.02 não foi simplesmente arbitrário, pois ao contrário, foi obtido através da

combinação de fatos relevantes como a quantidade de variáveis investigadas em cada bloco do questionário, a interdependência entre as regras descritas e a quantidade de regras, o que em várias ocasiões não permitiu a obtenção do conjunto completo de regras.

A princípio, foi considerada a hipótese de trabalhar, em conjunto, todas as regras correspondentes aos temas instrução, fecundidade, migração e mão-de-obra. Mediante a realização de alguns testes e de muita discussão, concluiu-se que seria impossível a execução do processo automático considerando todas as regras previstas em uma só aplicação, principalmente pela sobrecarga no sistema mas, também, pela complexidade na análise do conjunto de regras geradas.

É importante ressaltar que variáveis envolvidas em uma aplicação deveriam ser corrigidas e, posteriormente fixadas com vistas à correção de outras variáveis em aplicações subseqüentes, a fim de que se mantivesse coerência na estratégia pré-definida pelos analistas.

Julgou-se que não seria conveniente a estratégia de correção na qual cada aplicação abordaria um único tema, uma vez que a variável idade - V3076 (considerada, historicamente, uma variável de grande confiabilidade) estava envolvida nos temas de instrução, fecundidade, migração e mão-de-obra e que, por isso, sua imputação através de qualquer um desses temas, abordados separadamente, não teria embasamento suficiente para a sua seleção.

A partir daí, começou-se a avaliar uma determinada combinação de temas com o intuito de subsidiar a correção da variável idade. Um problema observado no tema fecundidade foi a existência de regras que envolviam, indiretamente, a variável idade com uma outra variável, V3351 (soma total de filhos tidos que moram no domicílio), criada e fixada durante a execução do processo automático, uma vez que suas componentes (V335 e V336 - filhos(as)tidos(as) que moram no domicílio) já tinham sido corrigidas na etapa descentralizada. Quando essa variável falhava com a idade da pessoa que estava sendo criticada não se tinha outra alternativa, a não ser corrigir a variável idade. Desse modo concluiu-se, então, que o tema fecundidade deveria constar da primeira aplicação.

Analisando a combinação dos temas fecundidade e migração verificou-se que poderiam ocorrer observações do tipo homens não migrantes[17] , de modo que não haveria embasamento suficiente para correção da variável idade. Posto isso, e considerando-se o volume de regras que envolviam a variável idade, pertencentes aos

[17] Nesse tipo de registro não são investigados os quesitos referentes à fecundidade e migração.

temas de instrução e fecundidade, optou-se por esta combinação.

No entanto, levando-se em conta a sobrecarga que significaria ao sistema a apuração das características de instrução e fecundidade se fossem consideradas todas as pessoas, o que impediria a obtenção do conjunto completo de regras, ficou acertado que as duas primeiras aplicações seriam destinadas à correção das características de instrução e fecundidade das pessoas de 10 anos ou mais, enquanto que a terceira trataria da apuração das características das pessoas de 0 a 9 anos. No caso da terceira aplicação foi discutida, ainda, a idéia de apurar as características de migração juntamente com as de instrução a fim de subsidiar a seleção, quando fosse o caso, da variável idade. Esta idéia não foi avante, devido ao mesmo problema relatado anteriormente.

No que diz respeito à apuração das características de migração e de mão-de-obra constatou-se que, após a realização de exaustivos testes, não seria possível gerar o conjunto completo referente à totalidade das regras previstas. Assim sendo, a estratégia adotada foi a de executar a correção dos dados de cada um desses temas em duas partes, onde as variáveis selecionadas para a primeira foram aquelas historicamente consideradas mais relevantes, sem que isso viesse a prejudicar a consistência dos dados.

É importante destacar que a viabilidade da execução das aplicações referentes à imputação das características de domicílios e de pessoas foi alcançada, mediante a implementação de alguns procedimentos definidos pela CTD, por parte de seus analistas e implementados pela Divisão de Sistemas para Censo e Pesquisas Demográficas (DIDEM), do Departamento de Atendimento (DEATE), da Diretoria de Informática.

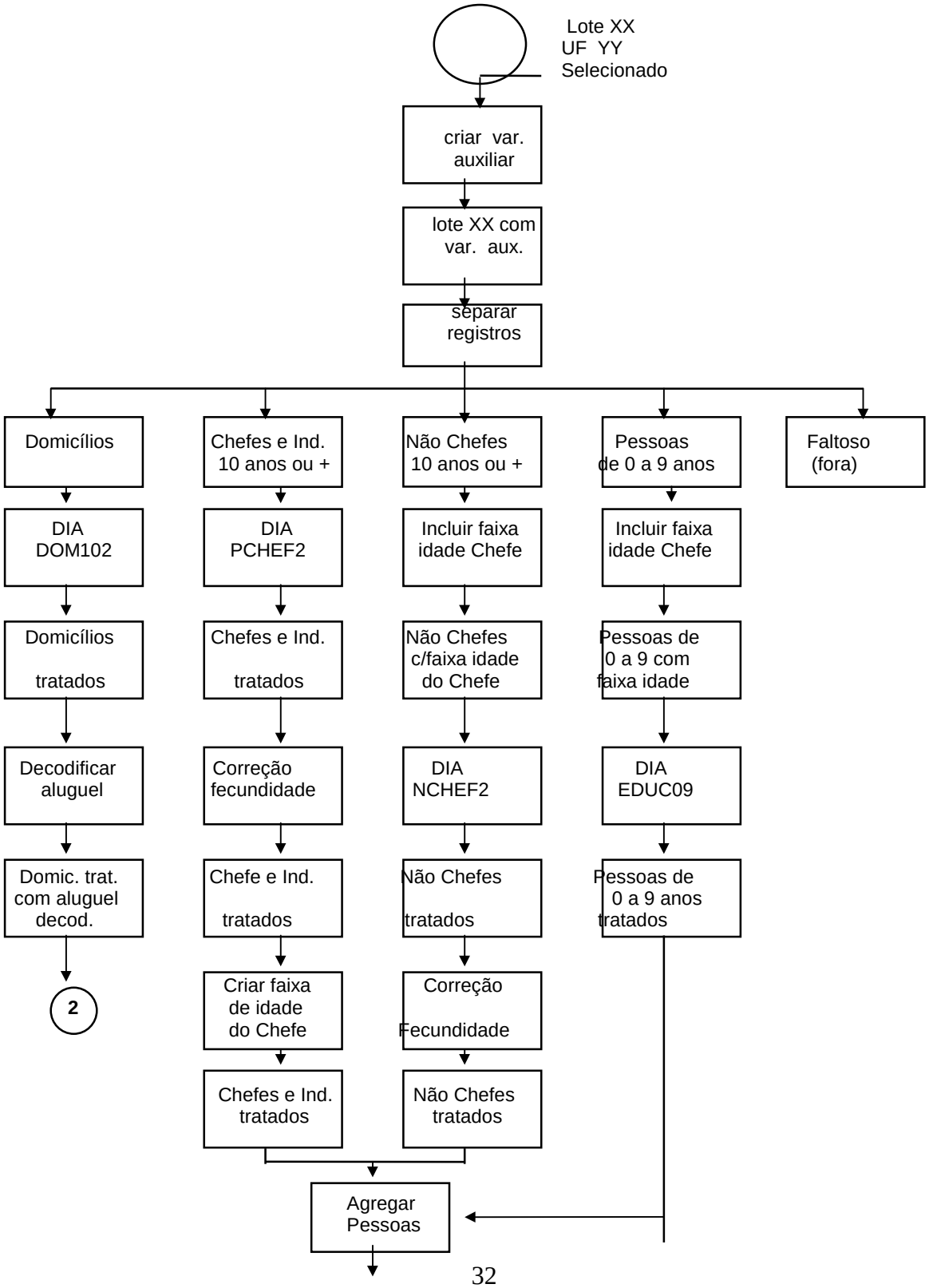
A seguir, são apresentados os quantitativos de regras segundo as aplicações definidas para a depuração dos dados e o fluxograma do processo automático.

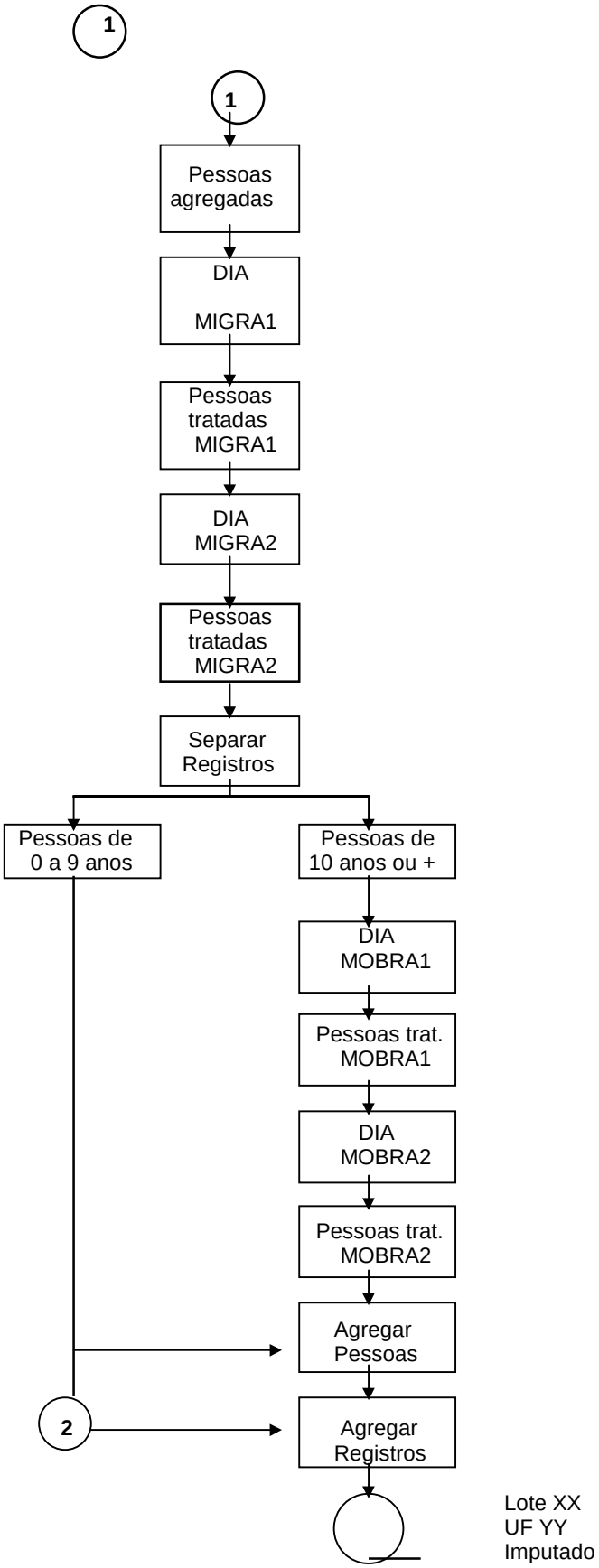
Quadro 1 - Número de regras de incompatibilidades e do conjunto completo, para cada aplicação

APLICAÇÃO	NÚMERO DE REGRAS	
	Incompati-bilidades	conjunto completo
Domicílios	**129**	**618**
Características do domicílio (DOM102)	129	618
Pessoas	**2010**	**4003**
Instrução e fecundidade dos chefes e individuais de 10 anos ou mais (PCHEF2)	345	1090
Instrução e fecundidade dos não chefes de 10 anos ou mais (NCHEF2)	347	1096

Instrução das pessoas de 0 a 9 anos (EDUC09)	27	38
Migração do total de pessoas 1º parte (MIGRA1)	412	628
Migração do total de pessoas 2º parte (MIGRA2)	385	568
Mão-de-obra das pessoas de 10 anos ou mais 1º parte (MOBRA1)	38	39
Mão-de-obra das pessoas de 10 anos ou mais 2º parte (MOBRA2)	456	544

Figura 2 - Fluxograma do processo automático de correção




1
1
Pessoas agregadas
DIA MIGRA1
Pessoas tratadas MIGRA1
DIA MIGRA2
Pessoas tratadas MIGRA2
Separar Registros
Pessoas de 0 a 9 anos
Pessoas de 10 anos ou +
DIA MOBRA1
Pessoas trat. MOBRA1
DIA MOBRA2
Pessoas trat. MOBRA2
Agregar Pessoas
2
Agregar Registros
Lote XX
UF YY
Imputado

As variáveis submetidas ao processo de detecção e correção automática, isto é, aquelas que foram tratadas na fase automática, são apresentadas, a seguir, para cada aplicação.

Aplicação 1 - Imputação das características de domicílios Erro! Indicador não definido.Erro! Indicador não definido.

Variáveis tratadas:

V202 - localização

V203 - paredes

V204 - cobertura

V205 - abastecimento de água

V206 - escoadouro

V207 - uso da instalação sanitária

V208 - condição de ocupação

V2090 - variável indicadora do preenchimento do aluguel mensal

V210 - combustível usado para cozinhar

V211 - total de cômodos

V212 - cômodos servindo de dormitório

V213 - banheiros

V214 - destino do lixo

V216 - filtro de água

V217 - telefone

V218 - automóvel particular

V219 - automóvel para trabalho

V220 - rádio

V221 - iluminação

V222 - geladeira

V223 - televisão preto e branco

V224 - televisão em cores

V225 - freezer

V226 - máquina de lavar roupa

V227 - aspirador de pó

Variáveis auxiliares à composição das regras:

Estas variáveis foram criadas na etapa preparatória da correção automática.

V7100 - total de pessoas registradas no questionário

Equivale à soma das V111 (total de homens) e V112 (total de mulheres).

V2205, V2206 e V2210 - variáveis indicadoras de: abastecimento de água, escoadouro e gás canalizado, respectivamente.

V1060 - tipo de setor

V2090 - (variável indicadora do preenchimento do aluguel mensal)

Para efeito de crítica no sistema DIA a V2093 (aluguel mensal) foi codificada na etapa preparatória da correção automática, dando origem à V2090, uma vez que o sistema só aceita variáveis categóricas de, no máximo, 4 dígitos.

Conforme já dito anteriormente, existia a possibilidade de se definir o tipo de depuração pretendido utilizando-se para isso as estratégias consideradas relevantes durante o processo automático de correção.

As estratégias de imputação utilizadas no plano de depuração foram:

Fixação de variáveis

Durante a execução do processo automático de correção, as variáveis descritas a seguir permaneceram fixas.

V7100 - total de pessoas

V1060 - tipo de setor

V201 - espécie do domicílio

V2205 - indicadora de abastecimento de água a nível de distrito ou subdistrito

V2206 - indicadora de escoadouro a nível de distrito ou subdistrito

V2210 - indicadora de gás canalizado a nível de distrito ou subdistrito

Modelo de imputação de códigos para cada variável

Todas as variáveis, com exceção da V202, pertencentes a essa aplicação foram imputadas segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não suspeitosos (FNS), utilizando o método proporcional.

Considerando que a imputação da V202 poderia ser, conforme o caso, probabilística ou determinística (dado que existiam regras determinísticas a fim de fazer com que nos setores especiais de favelas, o código registrado nessa variável ficasse coerente com tais setores), cabe ressaltar que a imputação probabilística foi realizada segundo a respectiva distribuição marginal, composta pela freqüência dos registros não

suspeitosos (FNS*), utilizando o método proporcional.

Ponderação de variáveis

Para todas as variáveis pertencentes a essa aplicação foi utilizado o peso médio (5).

A seguir são apresentadas as aplicações correspondentes à apuração das características de pessoas**Erro! Indicador não definido.**.

Aplicação 1 - Instrução e Fecundidade (chefes de famílias e individuais de 10 anos ou mais) Erro! Indicador não definido.

A execução dessa aplicação diz respeito à correção de eventuais inconsistências existentes em relação às características educacionais e à fecundidade das pessoas de 10 anos ou mais (chefes de famílias e individuais).

Cabe registrar que durante os testes de avaliação da correção automática foram utilizadas algumas estratégias de imputação (uso de distribuições condicionais), com vistas à correção das variáveis pertencentes ao bloco de fecundidade, considerando inválido o código 99 nas V337/V338, V339/V340 e V341/V342, a fim de que fossem imputados valores significativos, tentando assim aproveitar as informações de um número maior de mulheres para efeito de tabulação e análise.

As estratégias adotadas nos testes resumem-se na definição de três distribuições de dados, a saber:

- distribuição conjunta com a variável idade e uma variável indicadora de anos de estudo e estado conjugal;
- distribuição conjunta com variáveis formadas por grupos de idade qüinqüenais e decenais; e,
- distribuição marginal correspondente a cada variável de fecundidade.

No decorrer desses testes foi constatada, em alguns registros, a imputação de número de filhos bastante elevado, sem que tivesse algum registro de entrada com esse valor (as distribuições relativas entre a entrada e saída da imputação não se alteravam), resultado, talvez, de uma possível rarefação no interior da distribuição dos dados bons fazendo com que a probabilidade de ser selecionado 0, 1, 12, ou 20 filhos, por exemplo, fosse muito próxima uma das outras. Esse problema mostrou-se mais acentuado quando da utilização das duas primeiras estratégias.

Além disso, foram detectadas graves distorções em alguns indicadores de fecundidade e mortalidade (número de filhos por mulher, proporção de filhos mortos, proporção de filhos falecidos), calculados com os arquivos de entrada e saída, quando da utilização da primeira e terceira estratégias de imputação. Essa distorção foi mais acentuada no grupo de mulheres mais jovens que, provavelmente, não deviam ter tido filhos e não preencheram o bloco de fecundidade de modo adequado, fazendo com que durante a etapa descentralizada fosse aplicado código ignorado para todas as variáveis do bloco. Assim sendo, a substituição dos códigos 99 (ignorado) por outros códigos válidos gerou distorções que foram detectadas através do cálculo dos indicadores já citados.

Posto isso, os analistas da CTD e do DEMET reuniram-se para avaliar o problema e definir a solução mais adequada, de modo a não interferir nos indicadores. A decisão tomada foi a de considerar como válido o código 99 e utilizar a distribuição conjunta com as variáveis grupos de idades qüinqüenais e estado conjugal para a imputação das variáveis de fecundidade, tendo em vista a forte correlação existente entre elas.

Variáveis tratadas:

V3076 - idade calculada

V323 - sabe ler e escrever

V324 - série que freqüenta

V325 - grau da série que freqüenta

V326 - freqüência a curso não seriado

V327 - última série que concluiu com aprovação

V328 - grau da última série que concluiu com aprovação

V3291 - variável indicadora da espécie do curso mais elevado concluído com aprovação

V337/V338 - filhos(as) tidos(as) que moram em outro domicílio

V339/V340 - filhos(as)tidos(as) nascidos(as) vivos(as)que já morreram

V341/V342 - filhos(as) tidos(as) nascidos(as) mortos(as)

V343 - sexo do último(a) filho(a) nascido(a) vivo(a)

V3443 - idade do último(a) filho(a) nascido(a) vivo(a)

V3350 - soma total de filhos declarados

As variáveis auxiliares criadas na etapa preparatória da correção automática, foram:

V3303 **-** variável indicadora do estado conjugal

V3078 - variável indicadora dos grupos de idade qüinqüenais

V3350 e V3351 - soma total de filhos declarados e soma total de filhos tidos que moram no domicílio, respectivamente

A V3443, apesar de ter sido criada na etapa descentralizada, também foi utilizada durante a correção automática .

As estratégias de imputação utilizadas no plano de depuração foram:

Fixação de variáveis

V301 - sexo

V302 - parentesco com o chefe do domicílio

V303 - parentesco com o chefe da família

V308 - faixa de idade

V335 - filhos tidos que moram no domicílio

V336 - filhas tidas que moram no domicílio

V344 - mês e ano do(a) último(a) filho(a) nascido(a) vivo(a)

V3351 - soma total de filhos tidos que moram no domicílio

V3461 - variável indicadora da ocupação descrita no Q346 (ocupação habitual)

V3078 - variável indicadora dos grupos de idade qüinqüenais

V3303 - variável indicadora do estado conjugal

Modelo de imputação de códigos para cada variável

A imputação da V323, foi realizada segundo a distribuição conjunta com a V3076, enquanto que a imputação das V337 a V342 foi realizada segundo a distribuição conjunta com a V3303 e a V3078, cujas distribuições eram compostas pela freqüência dos registros não suspeitosos (FNS), utilizando o método proporcional.

Considerando que a imputação das V327 e V328 poderia ser, conforme o caso, probabilística ou determinística (dado que existiam regras determinísticas com o intuito de corrigir possíveis inconsistências provenientes de declarações equivocadas de série e grau equivalentes aos antigos primário, ginásio e colegial, ao invés de 1º grau ou 2º grau), cabe ressaltar que a imputação probabilística foi realizada segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não suspeitosos (FNS*), utilizando o método proporcional.

As demais variáveis pertencentes a essa aplicação foram imputadas segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não

suspeitosos (FNS), utilizando o método proporcional.

Ponderação de variáveis

Para todas as variáveis pertencentes a essa aplicação foi utilizado o peso médio (5), com exceção da V3076 cujo peso adotado foi (1).

É importante ressaltar que devido à limitação do sistema DIA, no que diz respeito ao tratamento de regras que envolviam a soma das variáveis componentes do bloco de fecundidade (V335 a V342), impossibilitando a imputação de algumas dessas variáveis quando a regra indicasse falha, foi necessária a utilização de algumas correções determinísticas,[18] realizadas através de programa "sob medida" após o processo de correção automática do sistema DIA, a fim de que os erros decorrentes dessas regras pudessem ser tratados adequadamente, isto é, alterando-se para "ignorado" os valores de algumas dessas variáveis, ao invés da soma (V3350).

Aplicação 2 - Instrução e Fecundidade (não chefes de famílias de 10 anos ou mais) Erro! Indicador não definido.

O processo automático de correção referente a essa aplicação foi executado segundo as mesmas orientações definidas para a aplicação 1 (pessoas), considerando também algumas alterações/inclusões.

Destinou-se à correção automática de inconsistências referentes às pessoas não chefes de famílias de 10 anos ou mais, a fim de que, as possíveis imputações da variável idade não fossem feitas independentemente, prejudicando a estrutura etária das famílias, de modo que tais imputações viessem a produzir algumas distorções entre as idades das pessoas de uma mesma família. Para levar adiante essa idéia, era necessário que a idade do chefe da família já estivesse limpa e conseqüentemente fixa nas aplicações posteriores, já que com base na idade do chefe seria criada uma outra variável V3077 (faixa de idade do chefe), após a execução da aplicação anterior, a qual serviria de controle para a imputação das idades dos não chefes.

As estratégias de imputação utilizadas nessa aplicação foram idênticas àquelas definidas na aplicação anterior, acrescida de um modelo tridimensional em que a imputação da V3076 levou em conta os valores de duas outras variáveis, V303 (fixa) e V3077 (fixa), com o intuito de garantir, na medida do possível, que as idades das pessoas

[18] Ver Indá, Oliveira, Mendonça, Lima e Guerra (1995).

de uma mesma família ficassem coerentes entre si.

Aplicação 3 - Instrução (Pessoas de 0 a 9 anos) Erro! Indicador não definido.

Nesse momento, dentre outras variáveis que foram analisadas e corrigidas, encontrava-se a variável idade que, ao final dessa aplicação, estava limpa e consistente com as demais, permanecendo fixa nas aplicações posteriores.

Variáveis tratadas:

V3076 - idade calculada

V323 - sabe ler e escrever

V324 - série que freqüenta

V325 - grau da série que freqüenta

V326 - freqüência a curso não seriado

V327 - última série que concluiu com aprovação

V328 - grau da última série que concluiu com aprovação

As estratégias de imputação utilizadas foram:

Fixação de variáveis

V303 - parentesco com o chefe da família

V308 - faixa de idade

V3077 - faixa de idade do chefe da família

Modelo de imputação de códigos para cada variável

A imputação da V323 foi realizada segundo a distribuição conjunta com a V3076, enquanto que a imputação da V3076 foi realizada segundo a distribuição conjunta com as V303 e V3077, cujas distribuições eram compostas pela freqüência dos registros não suspeitosos (FNS) utilizando o método proporcional.

As demais variáveis pertencentes a essa aplicação foram imputadas segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não suspeitosos (FNS), utilizando o método proporcional.

Ponderação de variáveis

Para todas as variáveis pertencentes a essa aplicação foi utilizado o peso médio (5), com exceção da V3076 cujo peso adotado foi (1).

Aplicação 4 - Migração (parte 1) Erro! Indicador não definido.

Essa aplicação abrangeu a maior parte das regras previstas para a correção do tema migração. Nela encontravam-se, basicamente, as regras que envolviam as V3076, V317 e V318 , cujos conjuntos de códigos possíveis são bastante extensos.

É válido ressaltar que o grande número de regras envolvendo as V3076, V313, V315, V317 e V318 impossibilitou a geração do conjunto completo de regras em uma única aplicação, devido ao enorme tempo de CPU necessário à análise das regras e dos inúmeros códigos possíveis das variáveis. Desse modo, a alternativa adequada foi a realização de uma outra aplicação envolvendo as características de migração (aplicação 5).

Variáveis tratadas:

V317 - há quantos anos mora nesta Unidade da Federação

V318 - há quantos anos mora neste município

V320 - na localidade indicada no Q.319, residia

V322 - na localidade indicada no Q.321, residia

As variáveis auxiliares criadas na etapa preparatória da correção automática foram:

V1010 - variável indicadora do município (Q.101)

V1011 - variável indicadora da Unidade da Federação (Q.101)

V3019 - variável indicadora de UF/País que residia antes de mudar para este município (Q.319)

V3021 - recodificação da V3214

V3191 - variável indicadora da relação entre as V3019 e V1011

V3192 - recodificação da V3193

V3194 - variável indicadora de município totalmente urbano em 1980 (município codificado na V3193)

V3215 - variável indicadora de município totalmente urbano em 1980 (município codificado na V3214)

V3245 - variável indicadora da diferença entre a V3076 e a V317

Variáveis criadas na etapa descentralizada que representavam os códigos aplicados à descrição dos respectivos quesitos, durante a codificação assistida:

V3193 - variável indicadora do Município/País que residia antes de mudar para

este município (Q.319)
V3214 - variável indicadora do Município/País que residia em set/1986 (Q.321)

No caso da V320 utilizou-se como referência a base de 1980 como um limite superior, tendo em vista que a resposta a esta variável correspondia ao Município/País que residia há menos de 10 anos; quanto à V322, segundo a DGC, não seria possível definir os municípios totalmente urbanos em setembro de 1986, pelo fato de não ser um ano de censo.

As estratégias de imputação utilizadas foram:

Fixação de variáveis

V1010 - variável indicadora de município (Q.101)
V314 - nasceu neste município
V3076 - idade calculada
V3021 - recodificação da V3214
V3191 - variável indicadora da relação entre a V3019 e a V1011
V3192 - recodificação da V3193
V3194 - variável indicadora de município totalmente urbano
V3215 - variável indicadora de município totalmente urbano
V3245 - variável indicadora da diferença entre a V3076 e a V317.

Modelo de imputação de códigos para cada variável

Considerando que a imputação das V320 e V322 poderia ser, conforme o caso, probabilística ou determinística (dado que existiam regras determinísticas para ignorar essas variáveis quando estivessem “em branco” e o código do município correspondente estivesse ignorado), cabe ressaltar que a imputação probabilística foi realizada segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não suspeitosos (FNS*), utilizando o método proporcional.

As demais variáveis pertencentes a essa aplicação foram imputadas segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não suspeitosos (FNS), utilizando o método proporcional.

Ponderação de variáveis

Foi adotado o peso médio (5) para todas as variáveis.

Aplicação 5 - Migração (parte 2) Erro! Indicador não definido.

A execução dessa aplicação teve a finalidade de corrigir, somente, as variáveis:

V312 - neste município morou

V313 - há quantos anos se deu a última mudança

V315 - em que ano fixou residência no País

As variáveis auxiliares criadas foram:

V3016 - variável indicadora da UF/País de nascimento

Esta variável foi criada na etapa descentralizada e representou o código aplicado à descrição do Q.316(Unidade da Federação ou País estrangeiro de nascimento), durante a codificação assistida.

V1061 - variável indicadora da situação do setor

Esta variável foi criada na etapa preparatória da correção automática.

As estratégias de imputação utilizadas foram:

Fixação de variáveis

V314 - nasceu neste Município

V3016 - variável indicadora de UF/País de nascimento

V1061 - situação do setor

V317 - há quantos anos mora nesta Unidade da Federação

V318 - há quantos anos mora neste município

V3076 - idade calculada

Modelo de imputação de códigos para cada variável

Todas as variáveis pertencentes a essa aplicação foram imputadas segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não suspeitosos (FNS), utilizando o método proporcional.

Ponderação de variáveis

Foi adotado o peso médio (5) para todas as variáveis.

Durante a execução das aplicações referentes ao tema migração, ocorreu um problema que fez com que o sistema, já em produção, fosse interrompido. A mensagem emitida pelo sistema salientava que naquele momento não estava sendo possível gerar o conjunto mínimo de campos a imputar.

Esse tipo de problema ocorre quando o conjunto completo de regras, utilizado no processo de imputação, gera alguma (s) regra (s) em que todas as variáveis envolvidas são fixas**,** isto é, não podem ser alteradas durante o processamento da aplicação. Devido a ausência de uma verificação instantânea sobre a existência dos códigos de Município/País, no momento da realização dos acertos na etapa descentralizada, especificamente em relação à V3193 e, algumas vezes, devido a erros no registro da V317 cuja diferença entre a V3076 (idade) e V317 era igual a -1, os analistas da CTD tiveram que alterar as informações dessas variáveis, *on line,* de modo que a aplicação pudesse seguir o curso normal.

Exemplo:

Uma pessoa com 9 anos de idade e código “branco” na variável auxiliar (V3191), indicando que a V3193 era igual ao código “branco”. Isso ocorreu pelo fato de que o código registrado na V3193 era inexistente, acarretando um código errado na variável auxiliar e, daí, gerando uma inconsistência com variáveis fixas em que uma pessoa com idade menor que 10 anos deveria ter preenchimento na V3193.

Aplicação 6 - **Mão-de-obra (parte 1)**Erro! Indicador não definido.

A execução da penúltima aplicação abrangeu a primeira parte da correção das características de mão-de-obra. Analogamente ao tema de migração, não foi possível realizar o trabalho de correção das características de mão-de-obra em uma única aplicação, por tratar-se de um tema que envolvia variáveis bastante ativas entre si e que possuíam um considerável conjunto de códigos possíveis, o que tornaria lento o processo e impediria a geração do conjunto completo de regras.

Diante da impossibilidade de executar a correção automática de uma só vez, foram feitos alguns testes com o intuito de dividir o processo em duas etapas sem perder de vista a consistência interna do tema. Essa forma de trabalho propiciou a execução da correção da primeira parte do processo, corrigindo apenas as variáveis V3461 e a V349 .

Em vista disso, foram separadas e tratadas as regras que envolviam essas variáveis e as V302, V323, V325, V3291 e V345, que naquele momento já encontravam-se fixas pois haviam sido corrigidas em aplicações anteriores.

Variáveis tratadas:

V3461 - variável indicadora da ocupação declarada no Q.346

V349 - relação de trabalho ou posição da ocupação

As estratégias de imputação utilizadas foram:

Fixação de variáveis

V302 - parentesco com o chefe do domicílio

V323 - sabe ler e escrever

V325 - grau de série que freqüenta

V3291 - variável indicadora da espécie do curso mais elevado concluído com aprovação

V345 - trabalhou em todos ou em parte dos últimos 12 meses

Modelo de imputação de códigos para cada variável

Considerando que a imputação da V3461 poderia ser, conforme o caso, probabilística ou determinística (dado que existia uma regra determinística para a imputação do código referente à ocupação "engenheiro agrícola", erroneamente classificada no banco de códigos), cabe ressaltar que a imputação probabilística foi realizada segundo a respectiva distribuição conjunta com a V349, composta pela freqüência dos registros não suspeitosos (FNS*), utilizando o método proporcional.

Vale ressaltar que a imputação da V349 ocorreu segundo a respectiva distribuição marginal, composta pela freqüência dos registros não suspeitosos (FNS), utilizando o método proporcional.

Ponderação de variáveis

Dentre as variáveis envolvidas nessa aplicação somente as V3461 e V349 foram passíveis de imputação. Para a V3461 foi utilizado o peso médio (5) e para a V349, por ser considerada uma variável de melhor qualidade de resposta, o peso utilizado foi (1).

Aplicação 7 - Mão-de-obra (parte 2) Erro! Indicador não definido.

A execução da segunda parte da correção das características de mão-de-obra

abrangeu todas as variáveis pertencentes ao tema, considerando fixas, porém, as variáveis que já foram corrigidas na primeira parte, de modo a tornar coerente o processo de correção.

Variáveis tratadas:

V3471 - variável indicadora do setor ou ramo de atividade em que exerceu a ocupação habitual

V350 - tem carteira de trabalho assinada

V351 - quantas pessoas trabalham no negócio, instituição, etc.

V352 - local de trabalho onde exerceu a ocupação habitual

V353 - é contribuinte de Instituto de Previdência Pública

V354 - horas trabalhadas por semana na ocupação habitual

V355 - horas trabalhadas por semana em outras ocupações

V3560 - variável indicadora do preenchimento do rendimento da ocupação habitual

V3570 - variável indicadora do preenchimento do rendimento de outras ocupações

V358 - se não trabalhou, indique a situação ou ocupação que tem

V359 - indique se é aposentado e/ou pensionista

V3600 - variável indicadora do preenchimento do rendimento de aposentadoria/pensão

V3610 - variável indicadora do preenchimento de outros rendimentos

Analogamente à V2093, pertencente à aplicação de domicílios, para efeito de crítica no sistema DIA, as variáveis V3563 (rendimento da ocupação habitual), V3573 (rendimento de outras ocupações), V3603 (rendimento de aposentadoria/pensão) e V3613 (outros rendimentos) foram codificadas durante a etapa preparatória da correção automática dando origem às variáveis V3560, V3570, V3600 e V3610.

As estratégias de imputação utilizadas foram:

Fixação de variáveis

V302 - parentesco com o chefe do domicílio

V325 - grau da série que freqüenta

V326 - freqüência a curso não seriado

V345 - trabalhou em todos ou em parte dos últimos 12 meses

V3461 - variável indicadora da ocupação descrita no Q.346

V349 - relação de trabalho ou posição da ocupação habitual

V3076 - idade calculada

Modelo de imputação de códigos para cada variável

A imputação da V3471 foi realizada segundo a distribuição conjunta com a V3461 e a V349, enquanto que a imputação da V352 foi realizada segundo a distribuição conjunta com a V3471, cujas distribuições eram compostas pela freqüência de registros não suspeitosos (FNS), utilizando o método proporcional.

Considerando que a imputação da V350 poderia ser, conforme o caso, probabilística ou determinística (dado que existia uma regra determinística para corrigir um erro sistemático de coleta referente aos “trabalhadores agrícola volantes”), cabe ressaltar que a imputação probabilística foi realizada segundo a respectiva distribuição conjunta com a V349, composta pela freqüência dos registros não suspeitosos (FNS*), utilizando o método proporcional.

As demais variáveis pertencentes a essa aplicação foram imputadas segundo as respectivas distribuições marginais, compostas pela freqüência dos registros não suspeitosos (FNS), utilizando o método proporcional.

Ponderação de variáveis

Foi utilizado o peso médio (5) para todas as variáveis.

3.3 - Análise da Correção Automática dos Dados Erro! Indicador não definido.

Seja qual for a metodologia empregada para detectar e corrigir automaticamente os dados de uma pesquisa, é de suma importância que, a posteriori, se proceda a uma análise dos dados imputados, como um todo.

Essa necessidade baseia-se no simples fato de que por motivos como a ocorrência de erros sistemáticos de coleta bem como de críticas imperfeitas, existe uma probabilidade de que os dados já corrigidos e dispostos em tabelas, previamente definidas, apresentem distorções prejudicando a divulgação dos resultados.

Não é por acaso que os analistas responsáveis pelo desenvolvimento do sistema DIA tiveram a sensibilidade de integrar o tratamento específico para erros sistemáticos, através das regras de imputação determinística, com a metodologia de Fellegi & Holt mediante uma especificação semelhante a das regras de incompatibilidades e operacionalização conjunta com a imputação probabilística, garantida pelo módulo “analisador de regras”.

Diferentemente dos erros denominados aleatórios, que podem ocorrer em qualquer momento, mas que são distribuídos de modo uniforme sem distorcer gravemente as distribuições, os erros sistemáticos podem distorcê-las gravemente, pois tendem a ocorrer em subconjuntos de dados de determinadas variáveis. Daí a necessidade de se corrigir deterministicamente esses tipos de erros.

Objetivando a liberação dos dados para serem carregados no banco de dados e, posteriormente, submetidos ao programa de cálculo de pesos, a análise da correção foi realizada a nível do lote de apuração e também a nível de município, uma vez que este último é considerado um importante segmento político-administrativo através do qual grande parte dos dados são divulgados.

O embasamento técnico para a realização desse trabalho encontra-se em documento elaborado pela CTD[19], no qual foram definidos, a priori, alguns indicadores e seus respectivos limites de tolerância, de tal forma que quando não fossem satisfeitos deveriam ser emitidas algumas tabelas para análise.

É bem verdade que através da análise da correção automática realizada para o CD 1.01 foi possível realizar algumas alterações, de maneira que o trabalho de análise da correção do CD 1.02 fosse mais ágil, isto é, restringindo-se um pouco mais a emissão de *Tablas*[20].

3.3.1 - Análise a nível de loteErro! Indicador não definido.

A execução da análise da imputação, a nível dos lotes de apuração, baseou-se nas informações constantes em uma tabela elaborada pela CTD, na qual era feita uma análise das regras segundo o percentual de erro. Para cada Unidade da Federação vinham descritos o percentual de registros com erro por lote e o percentual de registros falhados em determinada regra em relação aos registros maus, por aplicação.

Essa tabela permitiu que se tivesse uma visão geral do que ocorreu com os lotes de cada Unidade e através dela foi estabelecido um outro indicador, resultado do produto dos percentuais de registros com erros e de registros falhados na i-ésima regra. Para esse indicador ficou estabelecido um limite de tolerância de 10%, isto é, caso não fosse satisfeito esse limite aí sim seriam emitidas as *tablas* a nível de lote de apuração.

A tabela utilizada na análise do lote foi muito útil pois através das suas estatísticas, foi possível conhecer e compreender a crítica falhada, que algumas vezes repetia-se por

[19] Ver Plano de Análise da Correção Automática - Cd 1.02 (1995).
[20] Relatórios-padrão, emitidos pelo sistema DIA, contendo informações para subsidiar a análise do processo de imputação.

quase todos os lotes da Unidade da Federação.

Foram emitidos os primeiros relatórios do sistema DIA referentes aos lotes de Sergipe e Espírito Santo, com o intuito de avaliar os resultados do teste do processo automático de correção. Através desses relatórios, os analistas da CTD detectaram as seguintes situações de erros:

Domicílios:

Através da análise dos relatórios detectou-se a ocorrência de uma forte distorção na distribuição da V221 (iluminação), acarretando uma falsa eletrificação no setor rural, em virtude do preenchimento incorreto das V222 a V227 correspondentes à existência de: geladeira, Tv preto e branco, Tv em cores, freezer, máquina de lavar roupa e aspirador de pó, que vinham com preenchimento “não tem” ao invés de “branco”, já que de acordo com as instruções só deveriam ter preenchimento nas V222 a V227 quando houvesse iluminação elétrica nos domicílios, isto é, quando a V221 fosse igual ao código 1 ou 2 (iluminação elétrica com medidor ou sem medidor).

Partindo-se de um dos princípios do sistema DIA de que a alteração nas informações originais deve ser a menor possível, entende-se a ocorrência da forte distorção na V221, onde ao invés de imputar “branco” nas V222 a V227 foi imputada a V221 para 1 ou 2, conforme a respectiva distribuição.

Em vista da ocorrência desse erro sistemático foi incluída no conjunto de regras uma RID (regra de imputação determinística), de forma que os registros que estavam com esse tipo de erro preservassem a informação da V221 e as demais variáveis fossem alteradas para “branco”.

Migração:

Trata-se de um tipo de erro que ocasionou forte imputação na V317 (há quanto tempo mora nesta UF), ao contrário da V318 (há quanto tempo mora neste município), pois o erro detectado através da análise da regra envolvida e da tabulação segundo idade, mês de nascimento e a diferença entre as V3076 (idade) e V317, mostrou uma forte concentração em torno da diferença entre as V3076 e V317 (V3076 - V317 = -1), o que veio confirmar a suspeita de que este tipo de erro era decorrente do registro indevido na V317 para as pessoas que nasceram após a data de referência e sempre moraram na Unidade de Federação de nascimento.

Exemplo:

nascimento = out/1980 $\Rightarrow$ 10 anos de idade

Se a pessoa residia na Unidade da Federação desde o nascimento, o registro correto seria V317 = 10 anos, mas erroneamente vinha registrado V317 = 11, resultado da diferença entre 1991 e 1980, sem levar em consideração o mês de nascimento e o mês de referência.

Em vista do ocorrido foram incluídas algumas regras envolvendo os valores da

V317 e da V3076, de modo a forçar que o valor imputado fosse igual ao valor da V3076, já fixa nessa aplicação.

Cabe registrar que além das análises realizadas pelos analistas da CTD, com base nos relatórios emitidos pelo sistema DIA, houve também a participação de analistas do Departamento de População e Indicadores Sociais - DEPIS, no que tange à análise da correção automática dos dados através de estudos comparativos das distribuições conjuntas de variáveis, antes e depois da imputação, referentes aos temas de migração, escolaridade, fecundidade e domicílios e do Departamento de Emprego e Rendimento - DEREN responsável pela análise da imputação referente ao tema mão-de-obra.

Durante a execução das análises realizadas por esses Departamentos foram detectados alguns problemas, mencionados a seguir, para os quais a CTD investigou a fundo com o firme propósito de solucioná-los.

- Distorção detectada entre as distribuições dos dados “antes” e “depois” da imputação da V312 (neste município morou), especificamente em relação ao código 3 (nas zonas urbana e rural), referente aos registros das pessoas de 0 a 4 anos de idade, nos lotes de Sergipe e Espírito Santo.

Foram listados, para todos os registros das pessoas de 0 a 4 anos, além da situação do setor e das idades, as variáveis correspondentes ao bloco de migração, nos quais a V312 era igual a 1(só na zona urbana) ou 2(só na zona rural) antes da imputação e igual a 3 depois da imputação.

A solução encontrada, não só para os dados de Sergipe e do Espírito Santo, mas também para todas as Unidades da Federação, foi alterar a V313 (há quantos anos se deu a última mudança) de 00 para 98 de todos esses registros no arquivo “antes da imputação”, tendo em vista que na maior parte desses registros a V313 era igual a 00, em virtude de estar em branco e a digitação preencher com zeros, ou erroneamente ter sido registrado 00 anos, significando não ter mudado de zona.

- As distribuições “antes” e “depois” da imputação da V313 apresentaram uma distorção para os valores de 0 a 4 anos de mudança de zona da V313 quando a V312 era igual a 3. Em vista disso, foram listados todos os registros em que o código da V313, após a imputação, fosse de 0 a 4, considerando-se a hipótese de ter ocorrido um erro sistemático no preenchimento dessa variável , com valores de 01 a 11, significando tempo menor que 1 ano de mudança.

Após as análises desses registros, decidiu-se então, alterar a V313 para 00, somente naqueles registros em que a V312 fosse igual a 3 e a V313(01 a 11) maior que a V318 , visto que tanto em Sergipe como no Espírito Santo mais de 90% dos registros decorriam do erro: V313 (antes da imputação) maior que a V318 (antes da imputação).

- Detectou-se, também, uma distorção na V315 (ano em que fixou residência no país) referente às categorias “naturalizado brasileiro” e “estrangeiro”, para os registros, que antes da imputação, correspondiam a “brasileiros natos”. Na verdade, essa distorção foi provocada pelo preenchimento incorreto do quesito “Unidade da Federação ou País estrangeiro de nascimento”, uma vez que era registrado município ao invés da Unidade da Federação, que associado a uma certa deficiência no banco de códigos utilizado na codificação assistida, fez com que os códigos correspondentes fossem referentes a determinados países. Um outro tipo de erro, que possivelmente acarretou essa distorção, está relacionado a alterações equivocadas durante a etapa descentralizada. De posse da listagem contendo todos os registros, foi possível corrigi-los de acordo com o texto indicado.

Vale salientar que, na V3016, as declarações codificadas a partir do código 30, referem-se a países.

Exemplos:

- V315 = 100 e Q.316 = Salvador $\Rightarrow$ V3016 = 52(correspondente a El Salvador), ao invés de V3016 = 16, código referente a Bahia, nesta variável.

- V315 = 100 e Q.316 = Espírito Santo $\Rightarrow$ V3016 = 32(correspondente ao Canadá nesta variável, enquanto que em outras o código correspondia ao Espírito Santo) ao invés de V3016 = 18, código referente ao Espírito Santo, nesta variável.

- A análise da imputação do tema mão-de-obra detectou a presença de um erro sistemático no preenchimento da V350 (tem carteira de trabalho assinada), especialmente para os registros correspondentes aos trabalhadores agrícola volantes, código 01, na V349 (relação de trabalho ou posição da ocupação principal).

Ocorreu que para esses registros foi marcado erroneamente código 4 (não é empregado) na V350, ao invés de código 3 (não tem), uma vez que a grande maioria dos trabalhadores dessa categoria não tem carteira de trabalho assinada, enquanto que o

código 4 deveria ser marcado somente para os trabalhadores conta-própria, empregadores ou sem remuneração.

A estratégia adotada para a correção desse erro foi incluir uma regra determinística na aplicação de mão-de-obra (2ª parte), de modo que todo registro em que a V349 igual a 01 e V350 igual a 4 tivesse a V350 alterada para código 3.

- A análise da imputação do tema fecundidade foi realizada pelo DEPIS, cujo exame permitiu a identificação de um registro em que foi imputado 7 filhos em uma das V339/V340 (filhos(as) tidos(as) nascidos(as) vivos(as) que já morreram) para uma mulher de 19 anos de idade, o que segundo seus analistas poderia gerar distorção nos indicadores. Em vista disso, a CTD decidiu listar os registros de mulheres com idades de 10 a 14 e 15 a 19 anos, antes e depois da imputação, segundo o número de filhos, a fim de alterar aqueles registros nos quais tivesse havido erro de digitação ou mesmo imputação de um número de filhos considerado alto em relação à idade da mãe, apesar de estar de acordo com o conjunto de regras.

No entanto é válido salientar que antes mesmo de ter sido realizado este trabalho de listagem com possíveis alterações, a CTD, no intuito de subsidiar os trabalhos de análise da imputação desse tema, havia constatado, com base nos resultados obtidos para Espírito Santo e Sergipe, uma razoável coerência nos vários indicadores calculados sobre a fecundidade.

3.3.2 - Análise a nível de município Erro! Indicador não definido.

Para a análise da correção, a nível de município, utilizou-se o Quadro Resumo de Indicadores da Imputação contendo informações sobre o número de variáveis cujos valores de máximo de A_j (i) excederam o limite de tolerância estabelecido e a variável com o maior valor dessa estatística.

onde:

max 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 e A_j (i) = [fd_j (i) - fe_j (i)].

A primeira análise foi feita através desse quadro cujos informes descritos anteriormente eram, obviamente, a nível dos municípios que ultrapassavam os limites de tolerância impostos para a estatística max A_j (i). Analisando esses dados, algumas

vezes fazia-se necessária a emissão de outra tabela, “Registros imputados, de entrada, bons e depurados distribuídos segundo os códigos da variável (tabela A.5)”[21], para averiguar o que realmente ocorreu em determinada variável, fazendo com que a distância entre os dados de entrada e saída da imputação, ficasse acima do limite estabelecido.

Durante os trabalhos de análise da correção automática, observou-se, para vários municípios, a ocorrência do mesmo tipo de preenchimento para o bloco de escolaridade referente às pessoas que não estavam freqüentando escola, mas que já haviam freqüentado, isto é:

V327 - série concluída com aprovação - 4ª a 7ª séries

V328 - grau da série concluída com aprovação - 1º grau

V3291 - variável indicadora da espécie do curso mais elevado concluído com aprovação - primário

O processo de correção, preocupado em alterar o mínimo possível a informação original, alterava o conteúdo da V3291 de “primário” para “nenhum curso”. Esse tipo de ocorrência fazia com que esta variável fosse mais imputada que as demais e também apresentasse o maior valor de max A_j.

A imputação era considerada correta através da análise das regras e do conhecimento que se tinha sobre o conjunto de informações fornecido pelas pessoas no momento de informar a série e o grau quando estes equivaliam ao antigo primário. Nesses casos, as pessoas que freqüentaram da 4ª a 7ª séries do 1º grau tendiam a responder que já possuíam o curso primário uma vez que o 1º grau é equivalente às quatro séries do antigo primário mais as quatro séries do ginásio. Logo, tratava-se de uma correção adequada mas que poderia, algumas vezes, fazer com que fosse ultrapassado o indicador responsável pela emissão da tabela A.5.

3.3.3 - Considerações sobre os resultados consolidadosErro! Indicador não definido.

Posteriormente à execução e análise da correção automática, é possível trabalhar com esses dados, já consolidados, a nível das Unidades da Federação, Regiões e Brasil, elaborando algumas estatísticas e representações gráficas que propiciam conhecer o que efetivamente ocorreu a nível da detecção dos registros, bons e maus, e do número de variáveis imputadas nas aplicações utilizadas para a correção.

Nesse sentido, são apresentadas algumas estatísticas obtidas com base nos

[21] Ver Plano de Análise da Correção Automática - CD 1.02 (1995).

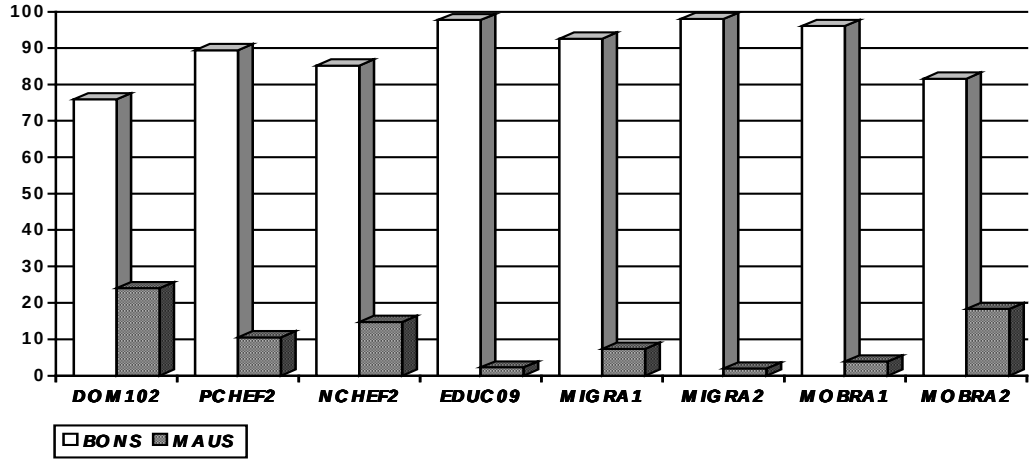
relatórios de análise dos lotes de apuração. A tabela 3 e o gráfico 1 apresentam, respectivamente, a distribuição dos registros, bons e maus, segundo as aplicações utilizadas na correção automática. A tabela 4 exibe a distribuição dos registros considerados errôneos, segundo o número de variáveis imputadas.

Tabela 3 - Distribuição dos registros bons e maus, para Brasil e Regiões, segundo as aplicações

APLICAÇÕES		BRASIL E REGIÕES											
		Brasil		Norte		Nordeste		Sudeste		Sul		Centro-Oeste	
		abs	rel	abs	rel	abs	rel	abs	rel	abs	rel	abs	rel
Dom102	**Bons**	3058766	75,99	120196	52,84	659511	61,73	1499151	84,71	574078	84,06	205830	74,39
Aplic. 1	**Maus**	966327	24,01	107280	47,16	408841	38,27	270516	15,29	108828	15,94	70862	25,61
Pchef2	**Bons**	3896910	89,48	214642	85,77	1050462	90,72	1696014	88,26	667406	91,38	268386	90,98
Aplic. 1	**Maus**	458154	10,52	35597	14,23	107462	9,28	225530	11,74	62958	8,62	26607	9,02
Nchef2	**Bons**	7387774	85,25	444573	81,39	2229502	86,48	3020571	83,72	1195189	87,41	497939	87,89
Aplic. 2	**Maus**	1278027	14,75	101644	18,61	348438	13,52	587290	16,28	172077	12,59	68578	12,11
Educ09	**Bons**	4041805	97,75	325619	97,11	1364583	97,70	1517456	97,74	569881	98,38	264266	97,54
Aplic. 3	**Maus**	92956	2,25	9681	2,89	32111	2,30	35083	2,26	9412	1,62	6669	2,46
Migra1	**Bons**	15716631	92,63	1043035	92,15	4842205	94,34	6405412	91,87	2473687	92,41	952292	90,42
Aplic. 4	**Maus**	1249811	7,37	88821	7,85	290348	5,66	566523	8,13	203236	7,59	100883	9,58
Migra2	**Bons**	16713682	98,05	1103334	97,48	5040690	98,21	6829633	97,96	2629117	98,21	1110908	98,10
Aplic. 5	**Maus**	332030	1,95	28522	2,52	91863	1,79	142302	2,04	47806	1,79	21537	1,90
Mobra1	**Bons**	12460331	96,13	767444	96,36	3608008	96,58	5259482	95,67	2000951	96,63	824446	95,70
Aplic. 6	**Maus**	501520	3,87	29012	3,64	127851	3,42	237914	4,33	69679	3,37	37064	4,30
Mobra2	**Bons**	10607819	81,67	619817	77,82	3073056	82,26	4480610	81,50	1744777	83,18	689559	80,04
Aplic 7	**Maus**	2381032	18,33	176639	22,18	662803	17,74	1016786	18,50	352853	16,82	171951	19,96

Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 1 - Percentual de registros, bons e maus, por aplicação (Brasil)



Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

De acordo com os resultados apresentados, pode-se verificar que as aplicações referentes à imputação das características de instrução das pessoas de 0 a 9 anos (EDUC09), e dos temas migração - 2ª parte (MIGRA2) e mão-de-obra - 1ªparte (MOBRA1) sinalizam os menores indicadores de registros detectados com erros. Por outro lado, as aplicações correspondentes à imputação dos chefes de famílias e individuais de 10 anos ou mais (PCHEF2) e ao tema migração - 1ªparte (MIGRA1) apresentam indicadores superiores às aplicações anteriores, não indicando, porém, nenhuma anormalidade. No entanto, cabe destacar que os maiores indicadores de registros errôneos referem-se à aplicação de domicílios (DOM102), à aplicação correspondente à imputação das características de instrução e fecundidade dos não chefes de famílias de 10 anos ou mais (NCHEF2) e à aplicação correspondente ao tema mão-de-obra - 2ªparte (MOBRA2). Em relação à DOM102, destacam-se as regiões Norte, Nordeste e Centro-Oeste com os mais altos indicadores, provavelmente devido à forte incidência do erro sistemático no preenchimento das V221 a V227, mencionado em 3.3.1. Quanto à NCHEF2, é bem provável que o grande número de variáveis investigadas juntamente com a quantidade de registros trabalhados tenham contribuído para uma maior incidência de registros errôneos; já MOBRA2, além do número de variáveis investigadas e da quantidade de registros (soma dos registros de PCHEF2 e NCHEF2) existe uma relativa complexidade inerente ao tema.

Tabela 4 - Distribuição percentual dos registros errôneos, segundo o número de variáveis imputadas, por aplicação, para as Unidades da Federação.

(continua)

UNIDADES DA FEDERAÇÃO	APLICAÇÕES E NÚMERO DE VARIÁVEIS IMPUTADAS															
	DOM102				PCHEF2				NCHEF2				EDUC09			
	1	2 a 5	6	7 ou mais	1	2	3 a 6	7 ou mais	1	2	3 a 6	7 ou mais	1	2	3 e 4	5 ou mais
RO	13,44	1,81	69,96	14,79	87,04	11,86	1,05	0,05	80,35	18,26	1,32	0,07	89,16	7,41	0,81	2,62
AC	14,05	2,01	70,01	13,93	90,05	9,28	0,67	-	89,23	10,14	0,63	-	95,52	4,48	-	-
AM	27,40	6,25	52,08	14,27	85,81	11,58	2,41	0,20	83,08	14,27	2,52	0,13	91,34	5,29	1,81	1,56
RR	21,46	5,27	40,21	33,06	87,51	11,66	0,83	-	83,87	14,42	1,71	-	91,91	6,25	0,74	1,10
PA	14,63	2,28	70,53	12,56	87,69	9,86	2,42	0,03	86,31	11,32	2,30	0,07	87,77	7,45	2,77	2,01
AP	29,34	4,58	52,07	14,01	90,83	9,17	-	-	87,00	12,41	0,59	-	94,87	5,13	-	-
TO	8,14	1,22	79,31	11,33	86,31	11,35	2,28	0,06	82,63	14,33	2,96	0,08	87,22	7,02	3,11	2,65
MA	8,97	1,22	83,69	6,12	86,45	9,94	3,58	0,03	85,01	11,82	3,10	0,07	84,77	9,60	3,24	0,39
PI	5,30	0,48	89,68	4,54	88,06	10,59	1,31	0,04	85,84	12,92	1,20	0,04	88,79	6,27	1,28	3,66
CE	12,86	1,40	80,44	5,30	90,56	7,81	1,58	0,05	89,49	8,37	2,06	0,08	91,16	5,53	1,13	2,18
RN	16,71	1,82	72,57	8,90	90,68	8,61	0,68	0,03	88,27	10,83	0,85	0,05	92,20	4,86	1,13	1,81
PB	11,56	1,29	80,78	6,37	87,33	11,75	0,92	-	85,00	14,09	0,90	0,01	93,14	5,44	0,88	0,54
PE	20,46	2,24	68,49	8,81	89,99	9,03	0,93	0,05	87,90	11,02	1,07	0,01	92,05	5,15	1,41	1,39
AL	19,69	2,34	68,97	9,00	89,30	9,48	1,18	0,04	86,68	12,36	0,94	0,02	92,64	5,64	0,70	1,02
SE	17,16	1,94	71,96	8,94	89,36	9,52	1,12	-	87,92	10,72	1,33	0,03	86,89	7,52	0,58	5,01
BA	14,32	1,93	74,02	9,73	87,83	9,87	2,20	0,10	84,84	13,12	1,97	0,07	86,54	7,03	3,35	3,08
MG	21,14	2,26	64,60	12,00	87,40	11,23	1,25	0,12	82,69	16,08	1,20	0,03	92,66	4,76	1,40	1,18
ES	39,59	4,42	44,72	11,27	87,98	11,21	0,73	0,08	85,60	13,70	0,63	0,07	93,99	2,78	0,45	2,78
RJ	58,08	11,08	13,44	17,40	85,49	13,08	1,38	0,05	82,72	15,65	1,59	0,04	91,74	6,17	1,02	1,07
SP	56,79	9,97	8,94	24,30	83,67	13,90	2,22	0,21	82,19	15,80	1,92	0,09	90,53	5,31	1,84	2,32
PR	34,72	3,98	47,67	13,63	79,85	19,43	0,68	0,04	76,71	22,56	0,73	-	93,39	5,55	0,81	0,25
SC	45,57	4,71	35,71	14,01	80,64	18,67	0,65	0,04	73,16	26,01	0,82	0,01	92,71	6,49	0,68	0,12
RS	27,61	2,82	55,29	14,28	89,70	8,68	1,49	0,13	86,91	11,54	1,49	0,06	91,89	4,06	1,77	2,28
MS	21,12	1,92	58,70	18,26	88,65	10,68	0,67	-	87,73	11,11	1,14	0,02	90,39	4,37	0,58	4,66
MT	19,93	2,20	59,98	17,89	87,93	11,59	0,48	-	83,19	15,86	0,94	0,01	94,49	5,05	0,38	0,08
GO	27,98	2,74	52,98	16,30	86,81	11,85	1,33	0,01	83,13	14,85	1,99	0,03	88,64	7,18	1,69	2,49
DF	55,10	8,31	13,96	22,63	87,02	10,69	2,26	0,03	86,05	11,63	2,31	0,01	86,26	5,69	2,84	5,21

Tabela 4 - Distribuição percentual dos registros errôneos, segundo o número de variáveis imputadas, por aplicação, para as Unidades da Federação.

(conclusão)

UNIDADES DA FEDERAÇÃO	APLICAÇÕES E NÚMERO DE VARIÁVEIS IMPUTADAS											
	MIGRA1				MIGRA2		MOBRA1		MOBRA2			
	1	2	3	4	1	2	1	2	1	2	3 e 4	5 ou mais
RO	89,33	9,30	1,30	0,07	97,02	2,98	99,30	0,70	69,08	22,18	8,27	0,47
AC	88,13	11,20	0,61	0,06	100,00	-	99,41	0,59	71,20	21,04	7,64	0,12
AM	78,60	18,35	2,81	0,24	97,12	2,88	98,74	1,26	61,76	25,27	12,01	0,96
RR	81,27	15,17	3,11	0,45	94,89	5,11	97,40	2,60	59,68	24,80	13,97	1,55
PA	88,29	10,66	0,93	0,12	95,31	4,69	99,08	0,92	69,79	22,01	7,73	0,47
AP	87,27	11,48	1,25	-	99,81	0,19	100,00	-	69,40	23,11	7,43	0,06
TO	87,56	11,06	1,31	0,07	99,67	0,33	98,86	1,14	70,40	22,78	6,36	0,46
MA	87,40	11,60	0,94	0,06	97,82	2,18	99,08	0,92	74,43	18,93	6,15	2,79
PI	89,39	9,55	0,94	0,12	98,11	1,89	99,50	0,50	69,58	24,74	5,44	0,24
CE	89,04	10,15	0,76	0,05	98,69	1,31	98,90	1,10	75,62	19,48	4,62	0,28
RN	89,64	9,64	0,68	0,04	98,44	1,56	99,26	0,74	72,68	21,37	5,72	0,23
PB	89,93	9,38	0,64	0,05	98,46	1,54	99,84	0,16	72,53	21,29	6,01	0,17
PE	87,96	11,26	0,74	0,04	99,22	0,78	99,48	0,52	76,42	17,95	5,43	0,20
AL	86,86	11,92	1,14	0,08	97,94	2,06	99,35	0,65	69,32	23,61	6,85	0,22
SE	87,69	11,24	1,03	0,04	99,36	0,64	99,64	0,36	60,09	33,51	6,23	0,17
BA	85,63	13,12	1,19	0,06	98,49	1,51	99,21	0,79	68,33	23,80	7,46	0,41
MG	88,25	10,28	0,81	0,06	98,84	1,16	99,38	0,62	72,48	21,03	6,29	0,20
ES	89,63	9,66	0,66	0,05	99,35	0,65	99,16	0,84	73,10	21,36	5,36	0,18
RJ	85,16	13,14	1,58	0,12	97,32	2,68	98,70	1,30	70,79	20,75	7,98	0,48
SP	86,36	12,20	1,31	0,13	97,90	2,10	98,84	1,16	70,38	20,88	8,17	0,57
PR	87,74	11,32	0,89	0,05	99,47	0,53	99,67	0,33	72,83	21,23	5,82	0,12
SC	87,76	11,33	0,87	0,04	99,27	0,73	99,43	0,57	70,60	22,51	6,66	0,23
RS	88,87	10,25	0,77	0,11	96,62	3,38	99,48	0,52	74,61	19,95	5,27	0,17
MS	89,41	9,70	0,86	0,03	97,50	2,50	99,41	0,59	75,32	18,69	5,81	0,18
MT	90,56	8,74	0,66	0,04	99,87	0,13	99,75	0,25	71,77	21,44	6,70	0,09
GO	88,36	10,58	0,98	0,08	99,11	0,89	99,46	0,54	71,46	21,44	6,87	0,23
DF	86,35	12,36	1,08	0,21	95,35	4,65	98,29	1,71	77,38	16,82	5,33	0,47

Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Como pode ser visto na tabela 4, a exposição dessas estatísticas vêm corroborar um dos princípios fundamentais da metodologia de Fellegi & Holt, que é a imputação do menor número possível de variáveis e, com isso, preservando, ao máximo, a informação original.

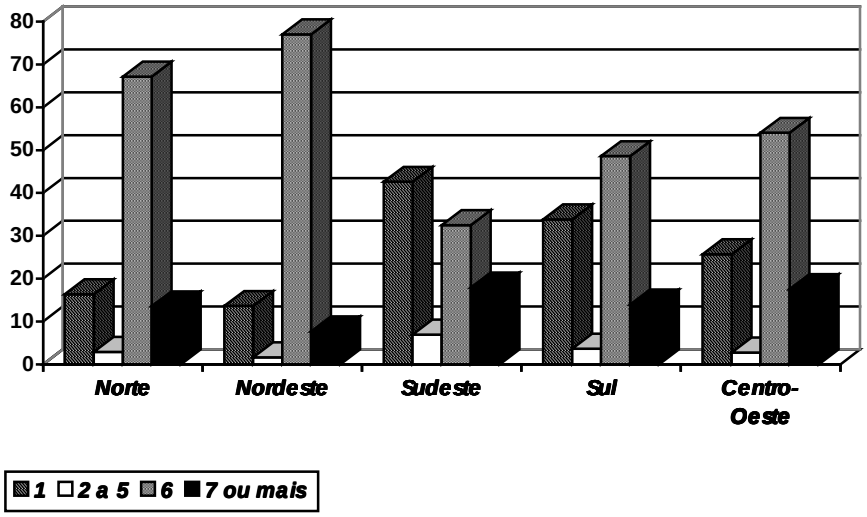
Esse fato poderia ser constatado em todas as oito aplicações, não fosse a ocorrência de um erro na aplicação referente à imputação das características de domicílios (DOM102), que acarretou uma "inversão" na distribuição dos registros errôneos segundo o número de variáveis imputadas, quando comparada com as demais distribuições onde o percentual de registros com apenas uma variável imputada abrange a grande maioria, ou seja, acima de 80%. Na realidade, trata-se do erro detectado no preenchimento das variáveis (V222 a V227) subordinadas à existência ou não de " iluminação elétrica"(V221) no domicílio, devidamente corrigido.

Em continuação à análise nota-se que essa "inversão" no número de variáveis imputadas ocorreu para todas as Unidades da Federação, sendo que Rio de Janeiro, São Paulo e Distrito Federal apresentam uma certa diferenciação em relação às demais Unidades, indicando os maiores percentuais na categoria " uma variável imputada" . Houve um certo destaque, também, na categoria correspondente à " 7 ou mais" na aplicação de domicílios, onde os percentuais encontram-se elevados se comparados com os correspondentes percentuais, nesta categoria, para as outras aplicações.

Um outro fato que deve ser salientado, diz respeito à imputação das variáveis nos registros errôneos referentes à aplicação do tema mão-de-obra - 2ª parte (MOBRA2), onde praticamente para todas as Unidades da Federação, a grande maioria dos registros foi corrigida imputando-se até duas variáveis. Acredita-se que isso deve ter ocorrido devido à fixação das V3461(variável indicadora da ocupação descrita no Q.346) e V349 (relação de trabalho ou posição da ocupação habitual), corrigidas na aplicação anterior, mas que nessa aplicação (MOBRA2) estavam envolvidas com outras variáveis que ainda seriam corrigidas e que, dependendo das regras falhadas, fazia com que determinados registros ficassem sem erros somente através da imputação de duas variáveis.

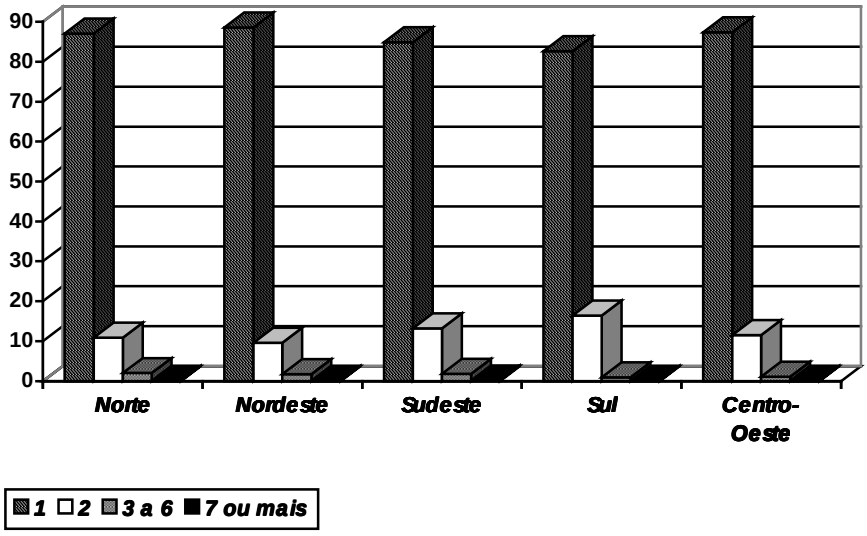
A fim de complementar a exposição dos resultados, encontram-se os gráficos, a seguir, elaborados através da consolidação dos dados para as regiões.

Gráfico 2 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(DOM102)



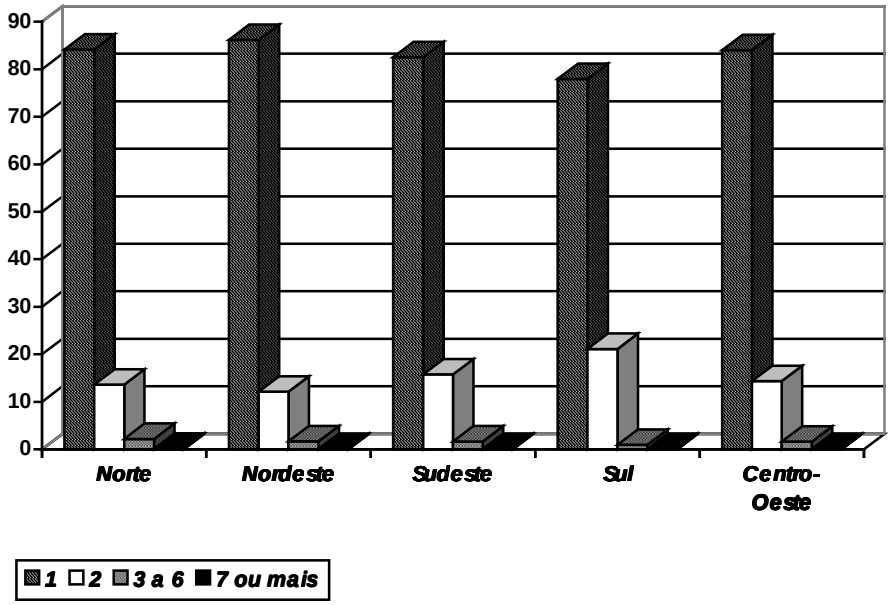
Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 3 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(PCHEF2)



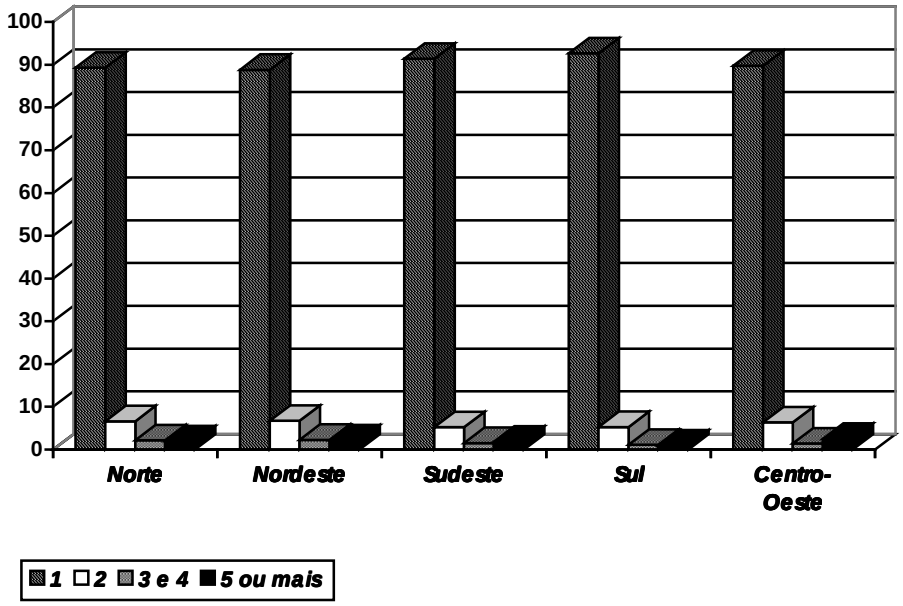
Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 4 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(NCHEF2)



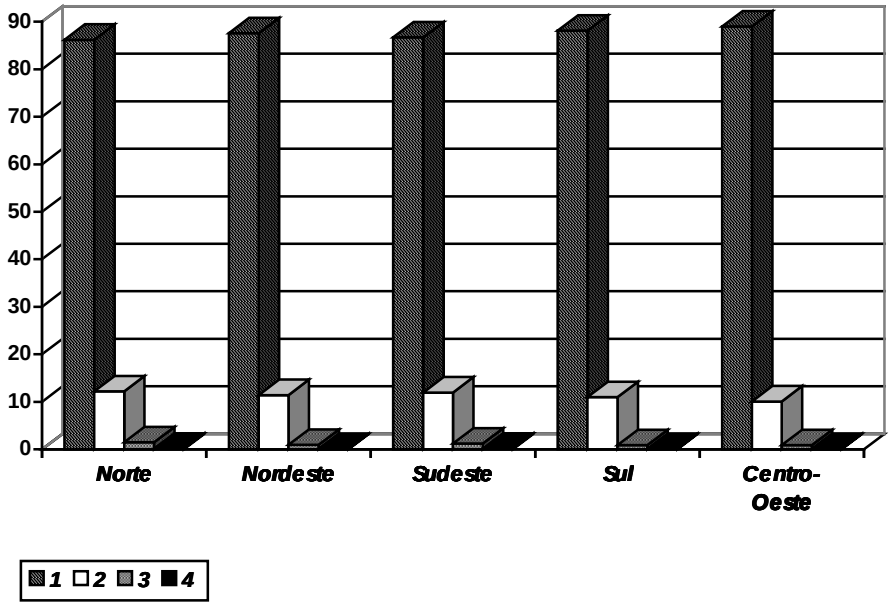
Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 5 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(EDUC09)



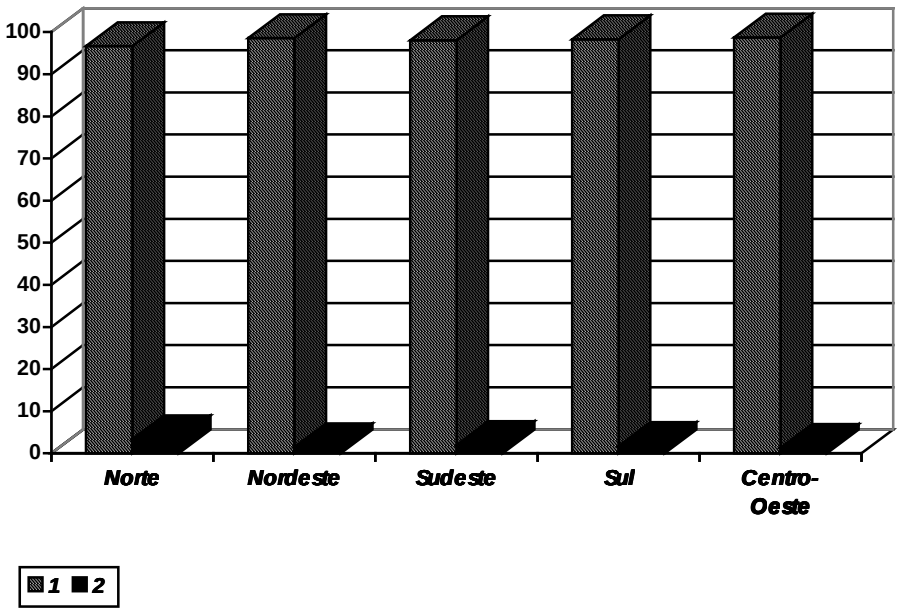
Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 6 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(MIGRA1)



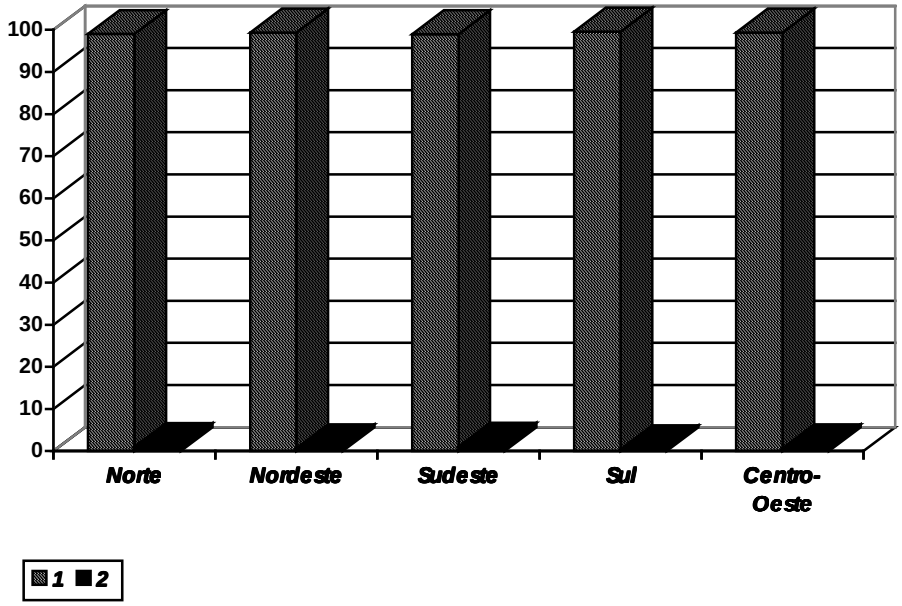
Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 7 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(MIGRA2)



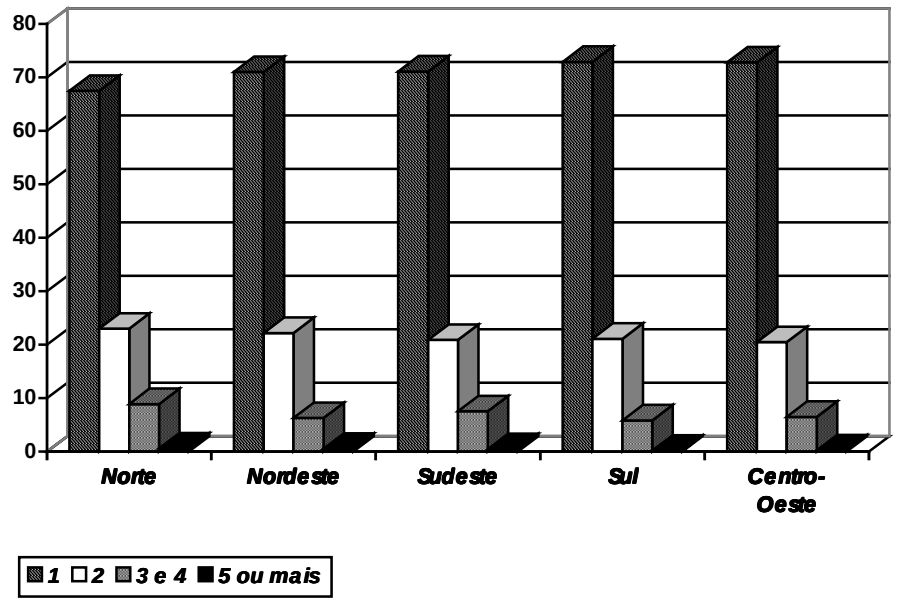
Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 8 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(MOBRA1)



Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

Gráfico 9 - Percentual dos registros errôneos, segundo o número de variáveis imputadas, para as Regiões.
(MOBRA2)



Fonte: Censo 91 - Resultados da Correção Automática do CD 1.02.

4 - EXPANSÃO DA AMOSTRAERRO! INDICADOR NÃO DEFINIDO.

Analogamente aos Censos Demográficos realizados em 1960, 1970 e 1980, o Censo de 1991 empregou a técnica de amostragem na coleta de informações sobre características de domicílios, migração, instrução, fecundidade, mão-de-obra e rendimento.

A intenção em se utilizar um determinado modelo de amostragem na apuração de um censo está relacionada, não só com a possibilidade de obtenção de relevantes informações de nível sócio-econômico e demográfico, para subsidiar a implementação de políticas governamentais e fornecer meios para outros usuários realizarem estudos comparativos, etc, mas também com a redução do custo de toda a operação censitária e a qualidade das informações obtidas.

A composição da amostra foi obtida pelos domicílios particulares e pessoas neles residentes e pelas famílias ou pessoas sós residentes em domicílios coletivos.

Pode-se considerar que a amostra do censo é uma amostra estratificada de domicílios, onde o estrato natural é o setor censitário.

Nesse censo foi introduzida uma inovação no processo de seleção da amostra, mediante a adoção de duas frações de amostragem segundo o tamanho dos municípios. Para os municípios com população superior a 15.000 habitantes a fração adotada foi de 10% e para os demais, 20%.

A obtenção das estimativas de totais das inúmeras características investigadas no questionário da amostra ocorreu, através da expansão das informações nele coletadas, baseada num determinado processo de estimação, a nível do conjunto de setores censitários pré-estabelecidos. A esses conjuntos dá-se o nome de “áreas de ponderação”, as quais representam as áreas geográficas construídas com o intuito de cálculo e divulgação dessas estimativas obtidas.

Com vistas à realização do trabalho de expansão da amostra do Censo Demográfico de 1991, o Departamento de Metodologia - DEMET, responsável pelo estudo de novas metodologias utilizáveis para os vários campos da pesquisa estatística, desenvolveu um estudo[22] através do qual analisou alguns métodos alternativos de estimação, dentre os quais o processo alternativo ao Processo Iterativo de Estimação por Totais Marginais - PIETOM, Grupo de Controle Iterativo - GCI sugerido por Costa (1987).

A razão da apresentação de um método alternativo para a expansão da amostra

[22] Ver Silva, Bianchini e Albieri (1992).

do censo derivou-se da tentativa de conseguir um método que se valesse das vantagens da utilização da pós-estratificação mas, também, que aumentasse a consistência das estimativas de totais, a nível das células da matriz com os totais conhecidos da população. Essa preocupação deve-se ao fato de que apesar do método utilizado na expansão do Censo de 1980 (PIETOM) utilizar também a matriz de pós-estratificação, o mesmo garante, apenas, a consistência das estimativas nas marginais da matriz.

Baseado nesse estudo, concluiu-se em pesquisar outras alternativas metodológicas para expandir os dados da amostra do Censo de 1991, que permitisse ajustar a pós-estratificação `as características de cada área de ponderação. Devido ao grande número de áreas de ponderação que se esperava utilizar na expansão da amostra, decidiu-se buscar métodos para formar os pós-estratos que fossem de fácil implementação computacional ou que já estivessem implementados em programas de computador acessíveis, para viabilizar a automação da tarefa de formação dos pós-estratos para a expansão em cada área de ponderação.

Dentre as alternativas avaliadas encontrou-se a metodologia desenvolvida por Bankier (1990), denominada *Generalized Least Squares Estimation Procedure* - GLSEP multivariada em dois estágios, também chamado pelos analistas do DEMET - Mínimos Quadrados Generalizados - MQG2, a qual foi adotada para expandir os dados da amostra do Censo de 1991.

Essa metodologia foi desenvolvida pelo Statistics Canada e utilizada na expansão da amostra do Censo Canadense de 1991. Sua aplicação ocorre através de programas em linguagem SAS, visando a obtenção de pesos e matrizes necessárias para o cálculo dos erros amostrais, o cálculo dos coeficientes de variação correspondentes às estimativas das características de domicílios e de pessoas.

Os dois estágios referem-se, na verdade, em ajustar os pesos duas vezes, com a intenção de garantir a consistência das estimativas de totais também a nível dos setores censitários, além das áreas de ponderação, cuja demanda por informação a nível de pequenas áreas vem crescendo ao longo dos anos.

A obtenção dos pesos dá-se através do ajuste de um modelo linear generalizado sujeito a restrições, formadas com os dados do universo. Os pesos são calculados de tal modo que as estimativas da amostra para os totais de determinadas características investigadas para o universo sejam bastante próximas dos totais conhecidos dessas mesmas características.

Ao contrário da metodologia empregada na expansão do Censo de 1980, apenas

um peso foi calculado para cada domicílio da amostra, fazendo com que as estimativas das características de pessoas fossem obtidas através do peso correspondente ao seu domicílio. Além disso, foram utilizados os pesos do modo como foram obtidos, isto é, inteiros ou fracionários.

Tal como nos censos anteriores, a metodologia MQG2 foi aplicada separadamente para cada uma das áreas de ponderação.

Cabe registrar que o presente trabalho procura dar, apenas, uma descrição sucinta sobre a metodologia adotada para a expansão da amostra, e que um detalhamento adequado pode ser visto em Albieri e Dias (1996).

5 - ARMAZENAMENTO DOS DADOSErro! Indicador não definido.

A fase final do trabalho de apuração do CD 1.02 está relacionada com o ambiente de armazenamento dos dados, previamente depurados, denominado *Relational Access for Integrated Databases* - Banco Rapid.

Nesse ambiente são encontrados os dados correspondentes às variáveis pertinentes à composição do Banco, devidamente armazenados, segundo a posição e o tamanho das mesmas. As variáveis incluídas no Banco são: de identificação, referentes às características de domicílios e pessoas e as auxiliares, tanto para domicílios (incluindo-se aí aquelas de identificação territorial e relativas à expansão), quanto para pessoas.

Posteriormente à execução da correção automática, os dados são armazenados no Banco Rapid, cuja fase inicial denomina-se carga do lote, segundo as unidades de trabalho que foram definidas com vistas à correção automática. Após o armazenamento dos dados é dada uma outra carga no Banco, denominada "carga na UF ", que tem a finalidade de rearrumar os dados armazenados até então, sob a forma de lotes de apuração, nas diversas áreas de ponderação.

Feito isso, o próximo passo é dar início ao processo de expansão dos dados, cujo resultado final é a geração dos fatores de expansão (pesos). Esses mesmos pesos, gerados durante a atividade de expansão, foram armazenados no Banco (denominado carga do peso) complementando as informações já existentes.

Todas as informações foram armazenadas, em caráter provisório, tendo em vista as possíveis análises estatísticas e demográficas a serem realizadas antes da carga final do Banco. Cabe registrar que a referência à expressão "dar carga final" tem o objetivo de armazenar os dados já corrigidos, de tal forma que venha a possibilitar maior segurança e facilidade na obtenção desses dados, seja para a tabulação ou para outras pesquisas, por ventura necessárias. Passada essa etapa, as informações relativas ao CD 1.02 encontram-se disponíveis para consultas e pesquisas de usuários.

6 - CONSIDERAÇÕES FINAIS

Erro! Indicador não definido.

A tarefa de elaborar este documento foi relevante, não só pela necessidade de registrar o desenvolvimento de etapas pertencentes à apuração do Censo Demográfico, que se não escritas perdem-se ao longo dos anos, mas, também, pela sua utilidade como subsídio para o planejamento de outros censos.

É importante ressaltar que, apesar das dificuldades inerentes à realização de um trabalho investigatório do porte do Censo Demográfico, foram introduzidas inovações relevantes em toda a etapa da apuração, como: descentralização de uma parte da fase de crítica; codificação assistida por computador dos quesitos em aberto; utilização do sistema DIA para a detecção e imputação automática e a obtenção dos fatores de expansão da amostra pelo método dos Mínimos Quadrados Generalizados.

Embora exista o risco de ser repetitivo, não se pode deixar de registar a sugestão de que a Instituição deve procurar investir, maciçamente, no treinamento das equipes de campo, na coleta e na supervisão, pois não basta ter a preocupação com métodos inovadores de crítica e correção, se a coleta dos dados não tiver sido feita dentro dos critérios pré-estabelecidos ou se a supervisão não tiver cumprido, razoavelmente, o seu papel identificando e solucionando os problemas inerentes à coleta como: incompreensão dos conceitos, má fé dos recenseadores, omissão de unidades e de seus componentes, etc.. Associado a isso, encontra-se o problema da remuneração dos recenseadores e supervisores que deveria ser a melhor possível, já que a qualidade dos dados coletados depende, em boa parte, da atuação dos mesmos.

REFERÊNCIAS: Erro! Indicador não definido.

ALBIERI, S. e DIAS, A.J.R. Metodologia de expansão da amostra do Censo Demográfico de 1991. Uma descrição resumida. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Divisão de Metodologia, 1994.

ALBIERI, S. e DIAS, A.J.R. A aplicação de mínimos quadrados generalizados na determinação dos fatores de expansão da amostra do Censo Demográfico de 1991. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Departamento de Metodologia, 1996.

ALBIERI, S. e OLIVEIRA, E.M. Censo Demográfico de 1991 - Critérios de formação de lotes de apuração dos questionários da amostra - CD 1.02. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Divisão de Metodologia, 1994.

COSTA, L.N. Processo de expansão da amostra do Censo Demográfico de 1980 - estudos experimentais. Rio de Janeiro: IBGE, 1987.

FELLEGI, I. P. e Holt, D. *A systematic approach to automatic edit and imputation*. Journal of the American Statistical Association, v.71, p. 17-35. 1976.

HANONO, R. M. DIA - Integração à arquitetura de informática do IBGE. Rio de Janeiro: IBGE, 1993 - (INFOTEC v.2 nº 9).

INDÁ, L.B. ; OLIVEIRA, L.C.S. ; MENDONÇA, M.S. ; LIMA, R.L.A. e GUERRA, V.S. Definições necessárias à implementação da apuração centralizada dos dados referentes ao questionário da amostra - CD 1.02 do Censo Demográfico de 1991, utilizando o sistema DIA para a detecção e correção automática dos erros. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Coordenação Técnica do Censo Demográfico, 1995.

MARTINEZ, A.F. ; RUBIO, E.G. e CRIADO, I.V. Sistema DIA - v.2. *descripción del sistema*. Madrid: Instituto Nacional de Estadística (INE), 1994.

METODOLOGIA do Censo Demográfico de 1980. Série Relatórios Metodológicos - Volume 4. Rio de Janeiro: IBGE, 1983.

PLANO de análise da correção automática do questionário da amostra - Censo Demográfico de 1991. Rio de Janeiro: IBGE, Coordenação Técnica do Censo Demográfico, 1995.

OLIVEIRA, L.C.S. ; INDÁ, L.B. ; LIMA, R.L.A. e BIANCHINI, Z.M. Apuração dos dados investigados no questionário básico - CD 1.01 do Censo Demográfico de 1991- Rio de Janeiro: IBGE, 1994 - 91p. (Textos para Discussão, n° 71).

RUBIO, E.G. e CRIADO, I. V. Sistema DIA - sistema de *Detección e Imputación Automática de errores para datos cualitativos* - Volumen 1. DIA: descripción del sistema. Madrid: Instituto Nacional de Estadística (INE), 1988.

SILVA, P.L.N. ; OLIVEIRA, E.M. ; OLIVEIRA, L.C.S. ; e LIMA, R.L.A. Uma nova metodologia para correção automática no Censo Demográfico brasileiro: experimentação e primeiros resultados. Rio de Janeiro: IBGE, 1990. 102p. (Textos para Discussão, n°28).

SILVA, P.L.N. ; BIANCHINI, Z.M. e ALBIERI, S. Uma proposta de metodologia para a expansão da amostra do Censo Demográfico de 1991 (versão preliminar). Rio de Janeiro: IBGE, 1993 (Textos para Discussão, nº 62) 106p.

SILVA, A.C.C.M. ; HANONO, R.M. e BARBOSA, D.M.R. Sistema gerador de aplicações de codificação assistida (SISDOC). Rio de Janeiro: IBGE, DI/DETEC/DISIA, 1993.

X RECENSEAMENTO Geral do Brasil de 1991. Manuais das atividades da apuração descentralizada do Censo Demográfico de 1991. Rio de Janeiro: IBGE, Coordenação Técnica do Censo Demográfico.